اصول و روش‌های ارزیابی فراگیران در علوم پزشکی

دکتر محمد جلیلی، دکتر محبوبه خباز مافی‌نژاد، دکتر رقیه گندم‌کار
دکتر سارا مرتاض هجری

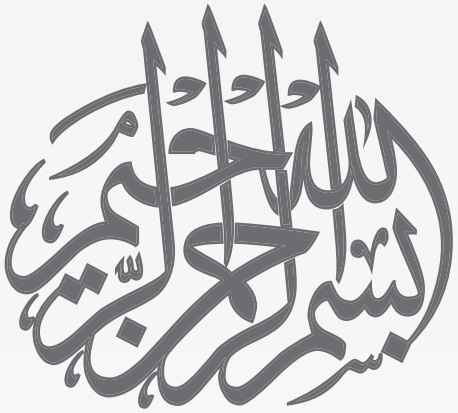
فرهنگستان علوم پزشکی جمهوری اسلامی ایران

اصول و روش‌های ارزیابی فراگیران در علوم پزشکی
محمد جلیلی، محبوبه خباز مافی نژاد، رقیه گندم کار، سارا مرتاض هجری
فرهنگستان علوم پزشکی جمهوری اسلامی ایران

Principles and Methods of Student Assessment in Health Professions

Mohammad Jalili (MD)
Mahboobeh Khabaz Mafinejad (PhD)
Roghayeh Gandomkar (MD,PhD)
Sara Mourtaz Hejri (MD,PhD)

The Academy of Medical Sciences
Islamic Republic of Iran


بسم الله الرحمن الرحيم

اصول و روش‌های ارزیابی فراگیران در علوم پزشکی

نویسندگان(به ترتیب حروف الفبا):

دکتر محمد جلیلی

دکتر محبوبه خباز مافی نژاد

دکتر رقیه گندم کار

دکتر سارا مرتاض هجری

اعضای هیات علمی دانشگاه علوم پزشکی تهران

عنوان کتاب: : **اصول و روش‌های ارزیابی فراگیران در علوم پزشکی**

چاپ اول : تهران ۱۳۹۶

تیراژ : ۱۰۰۰ جلد

ناشر : فرهنگستان علوم پزشکی جمهوری اسلامی ایران

تالیف : محمدجلیلی، محبوبه خباز مافی‌نژاد، رقیه گندم کار، سارا مرتاض هجری

صفحه آرا : مجید شمیمی

شابک : ۹۷۸-۶۰۰-۶۷۳۴-۰۷-۱

نشانی ناشر : تهران، بزرگراه شهید حقانی، مجموعه فرهنگستان‌ها، فرهنگستان علوم پزشکی، صندوق پستی ۱۹۳۹۵/۴۶۵۵

فهرست

پیشگفتار

آموزش در عرصه علوم پزشکی موضوعی است که از دیرباز مورد توجه بوده است. لکن پرداختن به آن به صورت علمی و روش‌مند در چند دهه اخیر پیشرفت‌های شگرفی داشته است. بهره‌گیری از مبانی نظری و تئوری‌های یادگیری و یاددهی و کاربست آنها در زمینه‌های گوناگون اعم از برنامه‌ریزی، فنون تدریس، روش‌های ارزیابی و ارزشیابی و موارد متعدد دیگر که می‌تواند باعث ارتقا کیفیت و افزایش کارایی آموزش شود، رو به گسترش است. ایران نیز دراین حوزه همزمان با تحولات و تغییرات جهانی گام‌های مهمی را در این مسیر برداشته است و دانشگاه‌های علم پزشکی سراسر کشور با تشکیل مراکز مطالعات و توسعه آموزش پزشکی با نگاهی تحولی به موضوع آموزش پرداخته‌اند.

اخیرا و با راه‌اندازی رشته آموزش پزشکی در مقطع کارشناسی ارشد و دکترا این مقوله وارد مقطعی جدیدی از رشد و بالندگی خود شده است. تدوین منابع و متونی که در این راه بتواند محل مراجعه مدرسان، محققان و سایر علاقمندان باشد یکی از اقداماتی است که می تواند در این مسیر کمک کننده باشد.

ارزیابی فراگیران با تاثیری که بر یادگیری آنها دارد یکی از نقاط حساس و اثرگذار در چرخه آموزش است و از این رو توجه به جنبه‌های مختلف آن در فرآیند یاددهی و یادگیری بسیار اثرگذار است. از سوی دیگر، یکی از تحولات اخیر حوزه آموزش پزشکی در کشور نیز که در قالب بسته‌های تحول آموزش مطرح شده است موضوع ارتقا روش‌های ارزیابی فراگیران را هدف قرار داده است و از این رو پرداختن به این مقوله می‌تواند با الزامات اسناد بالادستی نیز انطباق زیادی داشته باشد.

اصول و روش‌های ارزیابی فراگیران در علوم پزشکی مجموعه‌ای است که با حمایت فرهنگستان علوم پزشکی و با تلاش جمعی از اعضای هیأت علمی دانشگاه علوم پزشکی تهران تدوین شده است. این مجموعه غنی با مرور مستندات موجود در متون آموزش پزشکی به بیان مبانی اساسی ارزیابی فراگیران در حوزه آموزش علوم پزشکی می‌پردازد و در عین حال ابزارهای مورد استفاده در این زمینه را نیز یک به یک مرور می‌کند. مبتنی بر شواهد بودن مطالب و نیز استفاده از مثال‌ها و نمونه‌های کاربردی این مجموعه را برای طیف گسترده‌ای از مخاطبان به یک مرجع قابل استفاده تبدیل می‌کند.

ضمن تشکر از زحمات مولفین محترم این مجموعه که از اعضای هیات علمی و صاحب‌نظران ارزشمند حیطه آموزش پزشکی هستند امیدوارم انتشار این مجموعه گامی در جهت ارتقای نظام ارزیابی در آموزش علوم پزشکی باشد.

دکتر باقرلاریجانی
رییس گروه آموزش پزشکی
فرهنگستان علوم پزشکی

مقدمه نویسندگان

آموزش علوم پزشکی در چند دهه اخیر بهره مند از پیشرفتهای گوناگونی بوده که محصول نگاه تخصصی به مقوله آموزش در این حوزه است. از دیرباز معلمین گروه علوم پزشکی به سبقه برجستگی و تبحرشان در حوزه تخصصی خود به این مهم می‌پرداختند و کمتر از اصول و مبانی و روشهای روزآمد آموزش به معنای آکادمیک آن آگاهی داشتند. طبعا داشتن دانش و مهارت تخصصی در هر زمینه‌ای لازمه آموزش آن موضوع است اما برای این که یادگیری روی دهد به تنهایی کفایت نمی کند. آشنایی با مبانی برنامه ریزی آموزشی، اصول یاددهی و یادگیری و نظریه‌ها و روش‌های تدریس و ارزیابی و ارزشیابی امروزه جزء لاینفک توانمندی های معلمین است و اعضای هیأت علمی دانشگاههای علوم پزشکی کشور در طی سالهای گذشته با درک صحیح این موضوع به روشهای مختلف به کسب آشنایی با این مقولات روی آورده اند.

در کنار این پدیده میمون و مبارک که محصول آن استفاده از روشهای نوین در آموزش و ارزیابی فراگیران بوده است رخداد مهم دیگری در کشور به وقوع پیوسته است و آن ایجاد رشته تخصصی آموزش پزشکی در مقاطع کارشناسی ارشد و دکترای تخصصی بوده است. انتظار می رود با ایجاد گروههای آموزشی و مراکز تحقیقاتی خاص آموزش پزشکی تولید و کاربست دانش در این حوزه توسعه جدی یابد و دانش آموختگان این گروه‌ها بتوانند با نگاه عمیق و علمی به موضوعات مربوطه باعث ارتقا سطح کیفی آموزش در حوزه آموزش علوم پزشکی شوند.

طراحان برنامه‌های آموزشی عمدتاً بر تبین توانمندی‌های مورد انتظار، ایجاد فرصت‌های یادگیری همراه با ارزیابی صحیح فراگیران تاکید دارند. در این بین نقش و جایگاه ارزیابی بسیار بارز و برجسته است، به گونه‌ای که برخی آن را مهمترین جزء از این چرخه می‌دانند که به تنهایی می‌تواند تا حدود زیادی نقائص و کمبودهای سایر اجزا را پوشش دهد. به علاوه اگر چه در گذشته عمده نگاه به ارزیابی برای سنجش توانمندی فراگیران بود، در حال حاضر استفاده از ارزیابی برای یادگیری در کانون توجه قرار گرفته و آزمون‌ها خود ابزاری برای یادگیری محسوب می شوند.

ارزیابی چه از دیدگاه مدرسان و چه از دیدگاه فراگیران همواره مورد تاکید و مهم بوده است اما نکته دیگری که در مورد ارزیابی‌ها به طور ویژه جلب توجه می‌کند، حساسیت از سوی مسؤولان و نهادهای عمومی نسبت به آنها است. در حوزه آموزش به ندرت موضوعی مانند ارزیابی فراگیران محل کشمکش و اختلاف یا در معرض بازرسی و بازخواست قرار می‌گیرد. این حساسیت در کنار تاثیر و نقش ارزیابی در فرآیند یادگیری ضرورت توجه به این مقوله را بیش از پیش مشخص می‌کند.

گسترش دانش آموزش علوم پزشکی و توسعه حوزه‌های مختلف آن باعث شده که سنجش و ارزیابی نیز به صورت جدی‌تری دنبال شود به طوری که امروزه شاهد ایجاد و توسعه تئوری‌ها، اصول نظری و کاربردی، بحث و نقدهای مختلف، نوآوری‌ها و پژوهش‌های گوناگون در این عرصه هستیم. فعالیتهای زیادی برای ارتقا روشهای ارزیابی در طی سالیان گذشته در کشور صورت گرفته است و توفیقات زیادی در این زمینه حاصل شده است. در عین حال اطلاع صحیح، دقیق و عمیق از اصول و مبانی ارزیابی ضروری و لازمه استفاده درست از این روشهاست. از سوی دیگر مخاطبان این موضوعات طیف متنوعی دارند و محققان و متخصصان آموزش پزشکی از یک سو و مدرسان و مدیران برنامه‌های آموزشی از سوی دیگر به کسب اطلاعات و توانمندی در این زمینه نیازمند هستند.

نویسندگان کتاب طی سال‌های ارتباط خود با مبحث ارزیابی فراگیران در آموزش علوم پزشکی با این مشکل مواجه بودند که یک مرجع و منبع جامع و مستظهر به مستندات و متون علمی به زبان فارسی در این حوزه وجود نداشت. عمده کتب و منابع موجود یا در زمینه آموزش عالی به صورت کلی (و نه عرصه خاص آموزش پزشکی) بودند و یا به صورت مختصر و به صورت یک دستنامه یا به عنوان بخشی از یک کتاب جامع تهیه شده بودند. این موضوع در کنار ضرورت وجود مرجعی برای دانشجویان دوره دکترای تخصصی رشته های آموزش پزشکی و سنجش و ارزیابی در دروس مرتبط نویسندگان را برآن داشت تا با گردآوری آنچه که در طی سالیان گذشته در این حوزه به دست آمده استفاده از آن را برای علاقمندان تسهیل کنند و در چارچوب اصول و کلیات ارزیابی در آموزش پزشکی و نیز معرفی و تحلیل هر یک از ابزارها یک مجموعه جامع را فراهم کنند تا هم بتواند به عنوان درسنامه مورد استفاده متخصصین امر قرار گیرد و هم به عنوان مرجع سریعی برای کسانی باشد که به عنوان آموزش‌دهنده در ارزیابی فراگیران نقش دارند و می‌خواهند از طریق آشنایی بیشتر با ابزارها و روش‌ها بهره بیشتری از آنها ببرند یا تحلیل صحیح‌تری از نتایج ارزیابی‌هایشان داشته باشند.

مولفان این کتاب در دانشگاه علوم پزشکی تهران با درک اهمیت ارزیابی صحیح سعی کرده‌اند مجموعه‌ای جامع و به روز از نکات پایه و کاربردی در حوزه ارزیابی فراگیر فراهم کنند. با همه تلاشی که برای تامین و تضمین صحت مطالب کتاب شده است و سعی شده است تا مطالب کتاب بر مبنای مستندات روز در حوزه آموزش علوم پزشکی باشد، قطعاً کتاب خالی از اشکال نیست و چه از حیث محتوا و چه از نظر شکل ارائه نیازمند نقد صاحبنظران است. امیدواریم با استفاده از بازخوردهای خوانندگان محترم و کلیه صاحبنظران این عرصه در چاپ‌های آتی، نواقص برطرف و کتاب غنی‌تر شود.

در تدوین بخشی از این کتاب که به بحث ارزیابی استدلال بالینی می پردازد از راهنمایی های ارزشمند استاد گرامی جناب آقای دکتر منجمی بهره بسیار بردیم که نیازمند قدردانی ویژه از ایشان است. همچنین لازم است از استاد ارجمند جناب آقای دکتر لاریجانی که حامی و مشوق ما در انجام این کار بودند و پیگیری ها و حمایتهای ایشان در گروه آموزش فرهنگستان علوم پزشکی در تولید این اثر نقش مهمی داشت تشکر کنیم.

در پایان لازم می دانیم از همه همکاران دیگری که در تهیه این مجموعه ما را یاری کردند و به ویژه از حمایتهای فرهنگستان علوم پزشکی برای تهیه و نشر این مجموعه سپاسگزاری کنیم.

۱ کلیات ارزیابی فراگیر

فصل ۱

اهمیت و اهداف ارزیابی

اهمیت ارزیابی

«ارزیابی محرک یادگیری است» (میلر[1] ۱۹۹۰). این عبارت کوتاه و شناخته شده، نقش محوری ارزیابی را در هر شکلی از آموزش به صراحت بیان می‌کند. به ویژه در آموزش علوم پزشکی که عرصه‌ای پرخطر است، اغراق کردن در مورد اهمیت ارزیابی چندان دور از واقعیت نیست. علی‌رغم اهمیت این مقوله، دانشکده‌های پزشکی همچنان در انتخاب روش‌های ارزیابی فراگیر پیرو سنت‌های قدیمی بوده و محافظه کارانه عمل می‌کنند و از به کارگیری روش‌های نوین اجتناب می‌کنند. این امر از این مسأله نشأت می‌گیرد که عموماً ارزیابی به عنوان یک مقوله «پردردسر ولی ضروری»[2] در برنامه درسی پنداشته می‌شود. به عبارت دیگر، ارزیابی عملی است که انجام می‌دهیم چون مجبور به انجام آن هستیم. با این حال در بسیاری از مطالعات مطرح شده است که ارزیابی، به خصوص نوع برنامه‌ریزی شده و کاربردی آن، اثرات هدایت کننده مثبتی روی یادگیری و برنامه درسی دارد (امین[3] و همکاران ۲۰۰۶).

ارزیابی فراگیران یکی از قسمت‌های اصلی یک برنامه درسی است. ارزیابی صحیح می‌تواند با هدایت مدرسان و فراگیران تاثیرات قابل‌توجهی بر کل برنامه درسی اعمال کند و بیش از هر عامل دیگری در موفقیت یا شکست یک برنامه تاثیرگذار باشد. اینکه فراگیران چه مطالبی، در چه سطحی و به چه طریقی فرا می‌گیرند، به میزان زیادی به تفکر آنان درباره اینکه چگونه در انتها مورد ارزیابی قرار خواهند گرفت، بستگی دارد. ارزیابی‌های به‌عمل آمده باید به گونه‌ای باشند که به فراگیران دید درستی در خصوص اینکه چه مطالبی باید فرا گرفته شود و چگونه باید آن مطالب را فرا گیرند، ارائه نمایند.

در آموزش پزشکی سنجش فراگیران را می‌توان به عنوان یکی از روش‌های تضمین پاسخگویی به جامعه دانست. از طریق سنجش می‌توان اطمینان حاصل کرد که فراگیران، مهارت‌ها و شایستگی‌های لازم برای طبابت را کسب نموده‌اند و صلاحیت‌های لازم برای برآوردن نیازهای نظام سلامت در آن‌ها ایجاد شده است. توجه به این مهم ضروری است که ارائه ماهرانه مراقبت‌های سلامت نه تنها نیازمند داشتن مهارت‌های فنی و دانش مرتبط است، بلکه ارائه‌دهنده خدمت باید واجد دیگر مشخصه‌های کیفی از قبیل مهارت در برقراری ارتباط، ارائه مشاوره و مراقبت میان‌رشته‌ای، طب مبتنی بر شواهد و عملکرد مبتنی بر سیستم نیز باشد. بنابراین نظام سنجش نیز باید به حد کافی جامع و موثر باشد تا بتوان این ویژگی‌ها را هم‌راستا با آزمودن دانش و سایر مهارت‌های عملکردی مورد ارزیابی قرار دهد.

نگاهی به روند تغییرات صورت گرفته در علم سنجش و اندازه‌گیری نشان می‌دهد که امروزه شاهد افزایش روز افزون تعداد روش‌های ارزیابی و تخصصی‌تر شدن بیشتر آن‌ها هستیم. عواملی از قبیل رواج برنامه‌های درسی مبتنی بر پیامد،

1. Miller
2. Necessary Evil
3. Amin

ضرورت پاسخگویی و تضمین کیفیت، پیشرفت تکنولوژی و ارائه نظریه‌های اندازه‌گیری بر این امر موثر بوده است:

- **آموزش مبتنی بر پیامد**[1]: همسو با تحولات صورت گرفته در آموزش عمومی و آموزش عالی، در دو دهه گذشته شاهد بروز تغییراتی درباره اینکه پزشکان چگونه باید آموزش داده شوند نیز بوده‌ایم. برخلاف رویکردهای پیشین که تمرکز اصلی آنها بر موضوعات و عناوین درسی بود، در آموزش مبتنی بر پیامد، تاکید بر این است که پزشکان در طول و انتهای برنامه آموزشی باید از چه قابلیت‌هایی برخوردار باشند. آموزش مبتنی بر پیامد، در وهله اول شامل تعیین توانمندی‌های مورد انتظار از یک پزشک است. سپس برای نیل به آن توانمندی‌ها و متناسب با آنها، تغییرات در اجزای دیگر برنامه درسی شامل محتوا و ساختار برنامه، روش‌های یاددهی-یادگیری و جو آموزشی و روش‌های ارزیابی باید مورد توجه قرار گیرد. در این برنامه، ارزیابی نقش اساسی در تعیین اینکه آیا دانشجویان و دستیاران به توانمندی‌های مورد انتظار دست یافته‌اند و آیا برنامه آموزشی کارآمد بوده است، بازی می‌کند. این تغییر در نحوه تفکر و نیاز به ارزیابی توانمندی‌های متنوع پزشکان، فاکتور مهمی در تدوین و توسعه روش‌های جدید ارزیابی در طی سال‌های اخیر بوده و هست.
- **پاســـخگویی اجتماعی[2] و تضمین کیفیت:** حرکت به سمت آموزش مبتنی بر پیامد از طریق تلاش‌های چشمگیر در تقویت پاسخگویی پزشکان حاصل می‌گردد. استادان پزشکی به این موضوع واقف هستند که تعدادی از فراگیران از نظر دانش و مهارت‌های بالینی نواقص اساســی دارند. مســلماً تربیت فارغ‌التحصیلانی که از توانمندی‌های لازم برخوردار نیستند، در طولانی‌مدت منجر به تضعیف اعتماد عموم نسبت به جامعه پزشکی می‌شود. به این خاطر همواره تلاش می‌شــود تا گام‌های متعددی در جهت ارتقاء و تضمین کیفیت آموزش پزشــکی برداشته شود. قاعدتاً در چنین سیســتمی، عنصر مرکزی در پاســخگویی و تضمین کیفیت، ارزیابی خواهد بود. در واقع ارزیابی روشی برای بررسی عملکردها بر طبق استانداردهای موردنظر، شناسایی نقاط ضعف موجود و کمک در جهت ارتقاء کیفیت مداوم است که بالتبع این موضوع به ابداع روش‌های ارزیابی نوین و افزایش استفاده از روش‌های موجود پیشین کمک خواهد نمود.
- **تکنولوژی:** در طول ۵۰ سال گذشته، دسترسی به تکنولوژی پیشرفته دستخوش تحولات بسیاری شده است. با ظهور رایانه و امکان اسکن پاسخ‌های سؤالات چند گزینه‌ای و ارائه گزارش آن به صورت کارآمد و عینی، استفاده از این نوع سؤال به تعداد قابل ملاحظه و در مقیاس وسیع به راحتی میسر شده است. همچنین تکنولوژی هوش مصنوعی، از طریق ایجاد محیط‌های شبیه‌سازی شده، فراهم‌سازی استفاده از سؤالات تعاملی و امکان انتخاب سؤالاتی که به توانایی‌های اختصاصی هر فراگیر توجه دارد، فرصت بیشتری برای ارزیابی متوالی توانمندی‌ها، سنجش مهارت‌های تصمیم‌گیری و ارزیابی قضاوت‌های پزشکان منطبق با شرایط مختلف مهیا می‌کند.
- **نظریه‌های اندازه‌گیری**[3] به موازات تغییرات صورت گرفته در سایر علوم، پیشرفت‌های چشمگیری در زمینه نظریه‌های اندازه‌گیری صورت گرفته است. نظریه کلاسیک آزمون[4] با ظهور چشمگیر در قرن بیستم به تدریج راهی برای ارائه مدل‌های اندازه‌گیری گشود. پس از آن معرفی نظریه سؤال-پاسخ[5] و مدل‌های مختلف آن موجب تحولات فراوان در امر ارزیابی فراگیران شد که از جمله آن می‌توان به امکان تهیه نمرات هم‌تراز حتی در آزمون‌های متشکل از سؤالات مختلف، امکان اجرای آزمون‌های مبتنی بر رایانه در ارزیابی فردی فراگیران و امکان کوتاه نمودن آزمون‌ها اشاره کرد. نظریه تعمیم‌پذیری[6] نیز یکی دیگر از نظریه‌های اندازه‌گیری است که امکان شناسایی میزان خطا را در ابعاد مختلف اندازه‌گیری از طریق نرم‌افزارهای مخصوص فراهم می‌سازد. توضیحات بیشتر در این خصوص در فصل چهارم همین بخش خواهد آمد.

1. Outcome Based Education (OBE)
2. Social accountability
3. Measurement Theories
4. Classical Test Theory (CTT)
5. Item Response Theory (IRT)
6. Generalizability Theory (G-Theory)

به طور کلی می‌توان گفت که پیشرفت‌های صورت گرفته توانسته تا حدودی به بهبود کیفیت و کارآمدی ارزیابی‌های به‌عمل آمده کمک نماید. با توجه به موارد فوق، استادان درگیر در فرایند ارزیابی و تدریس فراگیران باید از نقش قابل توجهی که در این زمینه به عهده دارند، آگاه باشند و بدانند نمره‌ای که به فراگیران داده می‌شود به تنهایی اهمیت ندارد؛ بلکه انتخاب روش‌های ارزیابی، اجرا و پایش آن تعیین‌کننده پیامدهای نظام آموزشی است. این وظیفه اعضای هیأت علمی درگیر در ارزیابی است که با ابزارهای در دسترس و همین‌طور نقاط قوت و ضعف هر کدام از آن‌ها آشنا باشند.

اهداف ارزیابی

اینکه چرا باید ارزیابی انجام شود و هدف اصلی از انجام ارزیابی فراگیران چیست، از متداول‌ترین سؤالات مطرح در زمینه ارزیابی فراگیر است. نکته حائز اهمیت آن است که این سؤال نه تنها به اهداف ارزیابی اشاره دارد بلکه مشخص می‌کند چه کسانی قرار است از نتایج ارزیابی استفاده کنند.

در ارتباط با بخشی از سؤال که بر اهداف ارزیابی متمرکز است، شاید بتوان گفت که مهمترین کارکرد ارزیابی، تعیین میزان دستیابی به اهداف یادگیری است اما باید گفت که مسلماً دلایل دیگری نیز برای ارزیابی فراگیران وجود دارد. به عنوان مثال، با استفاده از نتایج حاصل از ارزیابی می‌توان نواقص و کمبودهایی را که در عملکرد فراگیران و برنامه آموزشی وجود دارد شناسایی نمود. به طور کلی، ارزیابی در صورتی که به درستی انجام شود، اهداف متعددی را محقق می‌سازد. تعدادی از این اهداف به شرح زیر است (امین و کو[1] ۲۰۰۳، نیوتن[2] ۲۰۰۷):

- تعیین میزان دستیابی به اهداف یادگیری و قضاوت در مورد توانمندی‌های فراگیران
- تعیین بهترین دانشجویان
- ترغیب فراگیران به یادگیری بیشتر
- کنترل میزان یادگیری
- درک فرایند یادگیری
- حمایت از یادگیری فراگیران از طریق ارائه بازخورد
- ارائه گواهی و اطمینان دادن به جامعه از اعتبار مدارک و گواهی‌ها
- تدوین و ارزشیابی برنامه‌های آموزشی
- پیش‌بینی عملکرد فراگیران در آینده

در پاسخ به بخشی از سؤال که به ذی‌نفعان ارزیابی می‌پردازد، باید گفت اهمیت اهداف فوق‌الذکر از دید ذی‌نفعان مختلف یکسان نیست. دایتل و همکاران[3] چهار گروه اصلی از ذی نفعان ارزیابی را مشتمل بر تصمیم‌گیرندگان، مدیران، استادان و فراگیران شناسایی کردند. هر یک از گروه‌های مذکور با اهداف متفاوتی از نتایج حاصل از ارزیابی استفاده می‌کنند (دایتل و همکاران ۱۹۹۱).

ارزیابی منجر به هدایت دانشجو به سمت ارزش‌های مورد نظر برنامه درسی می‌شود و نقش انگیزشی قوی در یادگیری فراگیران ایفا می‌کند. ارزیابی همچنین به دانشجویان کمک می‌کند تا شکاف بین برنامه درسی و آموزش را پر کنند. این مقوله به خصوص در مؤسسات بزرگ و در نظام‌های پیچیده آموزش بالینی صدق می‌کند. در این محیط‌ها که فراگیران در بیمارستان‌ها و بخش‌های مختلف حضور می‌یابند و با استادان بسیاری برخورد می‌کنند، وجود یک نظام ارزیابی ساختارمند، موجب حفظ وحدت و یکپارچگی عناصر برنامه درسی می‌شود و در حکم الگویی برای یکپارچه شدن آموزش در محیط‌های مختلف یادگیری است.

1. Amin & Khoo
2. Newton
3. Dietel et al.

به‌علاوه در بسیاری از مؤسسات آموزشی، ساختار برنامه درسی بر اساس امتحانات مهمی که دانشجویان باید بدهند، تدوین می‌شود. به عنوان مثال، در ایالات متحده دانشجویان پزشکی باید قبل از فارغ‌التحصیل شدن از دانشکده پزشکی مجموعه‌ای از آزمون‌ها را برای اخذ مدرک پشت سر بگذارند. این موضوع از یک سو به جهت‌دهی فعالیت‌های آموزشی برنامه درسی کمک می‌کند و از سوی دیگر باعث می‌شود دانشجویان پزشکی بتوانند از میزان پیشرفت خود در طول برنامه درسی اطلاع یابند. در چنین بستری، قاعدتاً ارزیابی به عنوان نیروی محرک فراگیران به سمت یادگیری تلقی می‌شود.

در نگاه وسیع‌تر، جامعه نیز حق دانستن این را دارد که آیا پزشکانی که از دانشکده پزشکی فارغ‌التحصیل شده‌اند از تبحر کافی برای ارائه خدمات برخوردار هستند یا خیر. این مسؤولیت در وهله اول بر عهده دانشکده پزشکی است که نشان دهد چنین توانمندهایی در فارغ‌التحصیلان ایجاد شده است و در سطح فراتر آن مسؤولیت سازمان‌های اعتباربخشی است که مشخص کنند برنامه‌های آموزشی توانسته‌اند به این موضوع پای‌بند باشند. بنابراین از این منظر، ارزیابی به عنوان قلب پاسخگویی اجتماعی دارای اهمیت است.

رانتری[1] درباره ذی‌نفعان مختلف و اهداف ارزیابی معتقد است که «محصول ماندگار ارزیابی، انگیزه یادگیری است» (رانتری ۱۹۸۷). از نظر وی اهداف انگیزشی مستقیم‌ترین هدف ارزیابی است که مرتبط با نیازهای فراگیران است. او معتقد است که هدف انگیزشی از دو جنبه ترغیب و اجبار قابل تحلیل است، زیرا گاهی اهداف انگیزشی ممکن است به عنوان ابزاری برای اجبار دیده شوند، یعنی راهی باشند برای سوق دادن فراگیران به یادگیری مواردی که در صورت نبود ارزیابی فرا نخواهند گرفت. بدین ترتیب، استادان ممکن است در مقایسه با فراگیران انگیزه بیشتری برای استفاده از ارزیابی داشته باشند.

همچنین رانتری استدلال می‌کند که گزینش و حفظ کیفیت از جمله اهدافی هستند که تمام ذی‌نفعان به جز فراگیران از آنها سود می‌برند، (هر چند فراگیران نیز نیاز به اطمینان از کیفیت آموزش دریافت شده دارند) و ذکر می‌کند که بازخورد می‌تواند به عنوان مهمترین هدف ارزیابی توسط فراگیران در نظر گرفته می‌شود.

از طرف دیگر، آمادگی برای ارتقاء حرفه‌ای نیز وابسته به نوع نگاهی است که هر کس می‌خواهد به سمت آن هدایت شود. در تفاسیر قدیمی‌تر، آمادگی به عنوان فرصت اشتغال مطرح می‌شد که خود جو رقابت را در این میان دامن می‌زد اما به نظر می‌رسد که امروزه این مفهوم در حال تغییر است و فردی که همکاری و تبادل اطلاعات را به عنوان یک مهارت ارزشمند زندگی امروزی در نظر می‌گیرد، قاعدتاً دیگر به دنبال رقابت حرفه‌ای نیست. در همین رابطه، بایرن‌بام[2] معتقد است که ارزیابی باید به تربیت فراگیران خود راهبری کمک کند که دارای مهارت برقراری ارتباط و همکاری با دیگران باشند (بایرن‌بام ۱۹۹۶).

اهداف ارزیابی از دیدگاه رانتری (۱۹۸۷)

رانتری در بررسی که انجام داد دریافت که حجم وسیعی از مستندات ارزیابی در ارتباط با این است که چگونه از نتایج حاصل با هدف نمره‌دهی و رتبه‌بندی دانشجویان استفاده شود و تنها نسبت کوچکی از مستندات موجود، ارزیابی را با هدف تقویت وضعیت آموزشی فراگیران مورد ملاحظه قرار داده بودند. در بررسی انجام شده، رانتری شش طبقه کلی برای اهداف ارزیابی ارائه کرده است:

- گزینش
- حفظ استانداردها یا کنترل کیفیت
- ایجاد انگیزه برای فراگیران
- ارائه بازخورد به فراگیران
- ارائه بازخورد به استادان
- آمادگی برای ارتقاء حرفه‌ای

در جدول ۱-۱ اهداف ارزیابی از منظر ذی‌نفعان مختلف به صورت خلاصه ارائه شده است.

1. Rowntree
2. Birenbaum

جدول ۱-۱: تقسیم بندی اهداف ارزیابی از منظر ذی‌نفعان مختلف (برگرفته از امین و همکاران ۲۰۰۶)

ذی‌نفعان	سؤالات	علایق
فراگیران	آیا شایستگی و دانش لازم را کسب کرده‌ام؟ چگونه می‌توانم بهتر عمل کنم؟	قضاوت در مورد شایستگی حمایت از یادگیری
استادان	آیا تدریس من موفق بوده است؟ چگونه می‌توانم بهتر عمل نمایم؟	پیشرفت برنامه اعتبار برنامه
جامعه حرفه‌ای (مشتری)	آیا پزشکان معتبری تربیت کرده‌ایم؟	ارائه گواهی و مجوز کار
دانشکده پزشکی	آیا منابع مالی به درستی خرج شده است؟ آیا مباحث درستی را تدریس کرده‌ایم؟ آیا به طریقه صحیح تدریس نموده‌ایم؟	توجیه برنامه تعدیل برنامه درسی پیشرفت برنامه درسی

ارزیابی برای یادگیری به جای ارزیابی یادگیری

در گذشته، کارکرد اصلی ارزیابی به سنجش و اندازه‌گیری میزان یادگیری‌های فراگیران خلاصه می‌شد اما به نظر می‌رسد که امروزه این نگاه در حال تغییر است و از ارزیابی به عنوان فرصتی جهت تقویت یادگیری فراگیران استفاده می‌شود. تفکر غالب در آن سال‌ها این بود که هدف اصلی ارزیابی، تعیین میزان موفقیت فراگیران در پشت سر گذاشتن یک دوره آموزشی است. در این شرایط، تصمیم در مورد اتمام موفقیت‌آمیز یک دوره آموزشی بر پایه آزمون‌های فردی صورت می‌گرفت. در واقع، تجسم اصلی این نگاه، در اجرای آزمون‌های پایان دوره ظهور می‌کند و عواقب و پیامدهای حاصل از آن نیز روشن است: در صورت کسب حدنصاب نمره در آزمون، فراگیر قبول در نظر گرفته می‌شود و قادر به پشت سر گذاشتن دوره است اما در شرایطی که فراگیر نتواند نمره قبولی را در آزمون پایانی کسب کند، ملزم به شرکت در آزمون مجدد یا گذراندن دوباره دوره آموزشی است. هر چند این نگاه به ارزیابی، همچنان خمیرمایه اصلی بسیاری از برنامه‌های ارزیابی است و چندان دور از عرف معمول به نظر نمی‌رسد، با این همه، مغایرت‌ها و اختلاف‌های آن با اصول حاکم بر محیط یادگیری به طور فزاینده‌ای احساس می‌شود (شوورث و ون‌درولوتن[1] ۲۰۱۱).

در عمل می‌توان کارکرد آزمون‌های پایانی را با آزمون‌های غربالگری در پزشکی (به عنوان مثال، سرطان پستان یا دهانه رحم[2]) مقایسه نمود که تاثیری در روند بهبود بیماری ندارند و تنها مشخص می‌کنند فرد احتمالاً سالم است یا بیمار. در حالی که مداخلات تشخیصی و درمانی به فردی که بیمار است برای بهبود کمک می‌کنند. به همین ترتیب آزمون‌های پایانی اطلاعات ارزشمندی را برای اطمینان از آنکه داوطلبان فاقد صلاحیت‌های لازم برای فارغ‌التحصیلی هستند فراهم می‌کنند اما اطلاعاتی درباره آنکه چطور یک داوطلب بی‌کفایت را می‌توان توانمند کرد، ارائه نمی‌دهند. بنابراین مشابه آزمون‌های غربالگری می‌توان گفت «ارزیابی یادگیری» بر خلاف «ارزیابی برای یادگیری» کمک زیادی به بهبود یادگیری فراگیران نمی‌کند (شوورث و ون‌درولوتن ۲۰۱۱).

از آنجا که هدف غایی برنامه‌های آموزشی تقویت یادگیری فراگیران است، ارزیابی نیز باید بهتر با این هدف منطبق شود. برنامه‌های ارزیابی که تمرکز اصلی خود را منحصراً بر غربالگری دانشجویان دارای صلاحیت از فاقد صلاحیت تنظیم می‌کنند، از پتانسیل موجود به منظور هدایت رفتارهای یادگیری فراگیران استفاده نمی‌برند. در همین رابطه گیپس (۱۹۹۹) نیز معتقد است که توجه صرف به ارزیابی، به عنوان ابزاری جهت اندازه‌گیری و سنجش عملکرد فراگیران کافی نیست (نورسینی و همکاران[3] ۲۰۱۱).

1. Schuwirth & Van der Vleuten
2. Cervical cancer
3. Norcini et al.

به علاوه در دیدگاه ارزیابی یادگیری، عمده تلاش‌های صورت گرفته در قالب تدوین یا کشف ابزارهای بهینه[1] برای سنجش توانمندی‌های مورد انتظار تجلی پیدا می‌کند که از نمونه‌های بارز آن می‌توان به طراحی آزمون OSCE برای سنجش مهارت‌های بالینی فراگیران در طول سالیان اخیر اشاره کرد. در این دیدگاه، سیستم ارزیابی[2] بهینه نیز به معرفی بهترین ابزار ارزیابی برای سنجش هر جنبه از توانمندی‌های مورد انتظار خواهد پرداخت (به عنوان مثال، استفاده از آزمون‌های چند گزینه‌ای برای ارزیابی دانش، آزمون OSCE برای سنجش مهارت‌ها، آزمون ویژگی‌های کلیدی برای توانایی استدلال و غیره). در حالی که در دیدگاه ارزیابی برای یادگیری، هدف اصلی جمع آوری اطلاعات از ابزارهای متعداد اندازه‌گیری و در زمان‌های مختلف برای پاسخگویی به سه پرسش زیر است (شوورث و ون‌درولوتن ۲۰۱۱):

- آیا اطلاعات کافی در خصوص وضعیت عملکردی یک داوطلب در دسترس است یا نیاز به کسب اطلاعات بیشتر وجود دارد؟ (سؤال تشخیصی[3])
- چه مداخله آموزشی برای این دانشجو در حال حاضر پیشنهاد می‌شود؟ (سؤال درمانی[4])
- آیا این فراگیر در حال حاضر در مسیر صحیح برای تبدیل شدن به یک فرد حرفه‌ای قرار دارد؟ (سؤال پیش‌آگهی[5])

همان‌طور که از مطالب فوق مشخص است پاسخگویی به این سؤالات از طریق استفاده از یک ابزار ارزیابی منفرد محقق نخواهد شد، بلکه اجرای مجموعه‌ای از روش‌های ارزیابی ضروری است.

1. Optimal
2. Assessment System, Assessment Program
3. Diagnostic question
4. Therapeutic question
5. Prognostic question

منابع

Amin Z, Chong YS, Khoo HE. Practical guide to medical student assessment: World Scientific; 2006.

Amin Z, Khoo HE. Basics in medical education: World Scientific; 2003.

Birenbaum M. Assessment 2000: Towards a pluralistic approach to assessment. Alternatives in assessment of achievements, learning processes and prior knowledge: Springer; 1996. p. 3-29.

Dietel R, Herman J, Knuth R. What does research say about assessment. NCREL, Oak Brook; 1991.

Miller GE. The assessment of clinical skills/competence/performance. Academic Medicine 1990; 87(7):63–67.

Newton PE. Clarifying the purposes of educational assessment. Assessment in Education. 2007;14(2):149-70.

Norcini J, Anderson B, Bollela V, Burch V, Costa MJ, Duvivier R, et al. Criteria for good assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. Medical teacher. 2011;33(3):206-14.

Rowntree D. Assessing students: How shall we know them? Taylor & Francis; 1987.

Schuwirth LW, van der Vleuten CP. General overview of the theories used in assessment: AMEE Guide No. 57. Medical Teacher 2011;33(10):783-97.

فصل ۲

تعریف اصطلاحات و عبارات رایج

مقدمه

عبارات و واژه‌های زیادی در بخش‌های مختلف این کتاب استفاده خواهند شد که عموماً در جای خود به تفصیل مورد بحث و بررسی قرار خواهند گرفت. با وجود این، کلمات و مفاهیم پایه‌ای در بحث ارزیابی فراگیر وجود دارد که سعی می‌کنیم در همین ابتدای کتاب تعریف مشخصی از آن‌ها ارائه کنیم.

از این رو در این فصل کتاب کوشش خواهد شد که مفاهیم بنیادی و کلیدی مرتبط با ارزیابی به روشنی توضیح داده شود. توجه به این نکته ضروری است که در این کتاب، مفاهیم ارزیابی و سنجش به جای یکدیگر به کار رفته‌اند و در متن حاضر تفاوتی از نظر معنایی با هم ندارند.

اندازه‌گیری، ارزیابی و ارزشیابی

اندازه‌گیری

عمل اندازه‌گیری[1] هنگامی صورت می‌گیرد که ما به مقادیری از یک صفت یا خصیصه که در یک فرد، یک شئ یا یک پدیده وجود دارد، براساس قواعد معینی عدد یا کمیت اختصاص می‌دهیم (سیف ۱۳۹۴). به عبارت دیگر عمل اندازه‌گیری، به معنای کمی ساختن صفات یا خصوصیات مورد اندازه‌گیری است. به عنوان مثال هنگامی که استاد در حال تصحیح برگه‌های امتحانی فراگیران است یا با مشاهده و بررسی عملکرد دانشجویان به آنان نمره می‌دهد، عمل اندازه‌گیری دانش فراگیران صورت می‌گیرد.

نظریه‌های اندازه‌گیری

نظریه‌های اندازه‌گیری به عنوان شاخه‌ای از علم سنجش و اندازه‌گیری، شامل روش‌ها و الگوهایی است که به تبیین فرایند اندازه‌گیری می‌پردازند و مدل‌های آماری را برای نمرات آزمون و سایر اندازه‌گیری‌ها توسعه و گسترش می‌دهد. نظریه‌های اندازه‌گیری که در این کتاب مورد بحث قرار می‌گیرد شامل نظریه کلاسیک آزمون، نظریه تعمیم‌پذیری و نظریه سؤال پاسخ است.

---

1. Measurement

ارزیابی

از ارزیابی[1] تعاریف متعددی ارائه شده است. در فرهنگ لغت وبستر[2] ارزیابی به عنوان عمل سنجش عملکرد تعریف شده است. ذوالفقاری و همکاران نیز ارزیابی را به معنای کشف میزان یادگیری دانشجویان تعریف می‌کنند که هدف از انجام آن، سنجش وسعت اندوخته‌های دانشجویان در سه حیطه دانش، مهارت و نگرش است (ذوالفقاری و همکاران ۱۳۷۹). در همین رابطه سیف معتقد است که ارزیابی یک اصطلاح کلی است که در طی آن از ابزارهای متعددی از جمله آزمون جهت اندازه‌گیری عملکرد فراگیران استفاده می‌شود (سیف۱۳۹۴)

ابزارهای ارزیابی

ابزارهای ارزیابی[3] روش‌هایی هستند که از طریق آنان معلمان به اندازه‌گیری و بررسی میزان پیشرفت تحصیلی فراگیران می‌پردازند. به طور کلی، روش‌های ارزیابی شامل آزمون یا هر روشی (به عنوان مثال، مشاهده، مصاحبه، پرسشنامه و ...) است که به منظور اندازه‌گیری علایق یا صلاحیت‌های فراگیران در طول دوره آموزشی مورد استفاده قرار می‌گیرند.

ارزشیابی

اگرچه لغت ارزشیابی و ارزیابی به کرّات در متون مترداف با یکدیگر مورد استفاده قرار می‌گیرد، با این حال از نظر مفهومی در پی پاسخ به سؤالاتی متفاوت هستند (شوورث و همکاران ۲۰۱۱). تعاریف متعددی از واژه ارزشیابی ارائه شده است. در فرهنگ دهخدا، ارزشیابی[4] به معنای «عمل یافتن ارزش هر چیز» و در فرهنگ معین به عنوان «عمل یافتن ارزش و بهای هر چیز، سنجش و بررسی حدود هر چیز و برآورد کردن ارزش آن» تعریف شده است. کرونباخ (۱۹۶۳) ارزشیابی را فرآیند نظامندِ گردآوری اطلاعات، به منظور اخذ تصمیم درباره برنامه آموزشی بیان می‌کند (سیف ۱۳۹۴). به طور کلی می‌توان گفت، ارزشیابی به معنای تجزیه و تحلیل نتایج حاصل از اندازه‌گیری است که در طی آن، درباره ارزش یا کیفیت برنامه آموزشی طبق ملاک‌ها و مقاصد معینی قضاوت و داوری می‌شود. از دیدگاه شوورث و همکاران (۲۰۱۱)، ارزیابی به معنای بررسی نظام‌مند عملکرد یا وضعیت تحصیلی فراگیران است، در حالی‌که ارزشیابی اشاره به بررسی وضعیت و کیفیت برنامه آموزشی و برنامه درسی دارد که قاعدتاً در دل آن مسائل مربوط به ارزیابی فراگیران، منابع، نیروی انسانی و پیامدهای کلی سازمان و برنامه نیز مطرح است.

در این کتاب از لغت ارزشیابی برای قضاوت در خصوص کیفیت برنامه آموزشی و از لغت ارزیابی برای سنجش عملکرد فراگیران استفاده شده است.

ارزشیابی پیشرفت تحصیلی

ارزشیابی پیشرفت تحصیلی[5] عبارت است از تعیین پیشرفت فراگیران در طول دوره آموزشی (ذوالفقاری و همکاران ۱۳۷۹). در ارزشیابی پیشرفت تحصیلی عملکرد یادگیرندگان مورد سنجش قرار می‌گیرد و نتایج حاصل با هدف‌های آموزشی از پیش تعیین شده مقایسه می‌گردد تا مشخص شود که آیا فعالیت‌های آموزشی به نتایج مطلوب انجامیده‌اند یا خیر و به چه میزانی. هدف از این آزمون‌ها، بهبود یادگیری فراگیران، ارائه بازخورد و ارزشیابی سطح پیشرفت فردی فراگیران در مقایسه با سایرین هم می‌تواند باشد.

1. Assessment
2. Webster
3. Assessment tools
4. Evaluation
5. Academic achievement evaluation

آزمون

سیف آزمون[1] را معمول‌ترین وسیله اندازه‌گیری ویژگی‌ها یا صفات روانی می‌داند که شامل مجموعه‌ای سؤال است که برای پاسخگویی در اختیار داوطلبان قرار می‌گیرد تا توانایی‌های آنها را اندازه‌گیری کند. گرانلاند و لین[2] ذکر کرده‌اند که آزمون وسیله یا روشی نظام‌مند برای اندازه‌گیری نمونه‌ای از رفتارها است. اناستازی و یوربین[3] از این تعریف فراتر رفته‌اند و آزمون را وسیله‌ای اندازه‌گیری عینی و استاندارد نمونه‌ای از رفتارها دانسته‌اند (اناستازی و یوربین ۱۹۹۷). برای درک بهتر این تعریف، آنها سه کلمه «عینی»[4]، «استاندارد»[5] و «نمونه‌ای از رفتارها» را به این صورت توضیح دادند:

- یک آزمون زمانی عینی تلقی می‌گردد که فرایند اجرا، تصحیح و تفسیر آن فارغ از قضاوت ذهنی ارزیابان بوده و بر اساس قواعد معین و مشخص باشد.
- یک آزمون استاندارد آزمونی است که طراحی سؤالات، نمره‌دهی، تفسیر و اجرای آن از یک فرد به فرد دیگر مشابه باشد (گرگوری[6] ۲۰۰۴). به بیان ساده‌تر، اگر ما قصد داریم تا نمرات داوطلبان را با یکدیگر مقایسه نماییم، این کار مستلزم آن است که تمامی آن‌ها را با سؤالات مشابه و تحت شرایط یکسان ارزیابی کنیم.
- یک آزمون باید نمونه‌ای از رفتارهای مشخص را بسنجد. به عنوان مثال اگر ارزیاب تمایل دارد که دانش لغات تخصصی فراگیران را در زمینه پزشکی بسنجد، نمونه سؤالات آزمون باید معرف و بیانگر دانش مربوط به اصطلاحات پزشکی باشد.

آزمون‌شونده و آزمون‌گر

آزمون‌شونده[7]، داوطلب یا آزمودنی فردی است که عملکرد او مورد آزمون قرار می‌گیرد. آزمون‌گر[8] یا ارزیاب به صورت سنتی فردی در نظر گرفته می‌شود که مسؤولیت طراحی یا اجرای آزمون نوشتاری یا عملکردی را بر عهده دارد. اما فلاچیکو[9] با ذکر موارد دیگری که شامل خودارزیابی[10]، ارزیابی توسط همتایان[11] و ارزیابی مبتنی بر رایانه[12] است، در این باره تصویر دقیق‌تری ارائه می‌دهد (فلاچیکو ۲۰۰۴):

- **خودارزیابی:** خودارزیابی اشاره به درگیر کردن فراگیران به قضاوت در خصوص عملکرد خودشان دارد. هرچند خودارزیابی معمولاً به شکل تکوینی انجام می‌شود، در مواردی ممکن است سهمی از نمره نهایی فراگیران را تشکیل دهد و به صورت تراکمی با طرح این پرسش صورت گیرد که آیا دانشجویان از نظر خودشان توانسته‌اند مطالب را همانگونه که انتظار داشتند، فرا گیرند یا خیر (بوود[13] و فلاچیکو ۱۹۸۹).
- **ارزیابی توسط همتایان:** در ارزیابی بوسیله همتایان نیز، فراگیران با استفاده از معیارها و استانداردهای مشخص به بررسی عملکرد یکدیگر می‌پردازند. این ارزیابی نیز به طور معمول به شکل تکوینی صورت می‌گیرد اما در مواردی ممکن است به شکل تراکمی نیز انجام شود.
- **ارزیابی مبتنی بر رایانه:** این ارزیابی همان‌طور که از نامش پیداست به معنای استفاده از رایانه برای تسهیل ارزیابی

1. Test
2. Grounland & Linn
3. Anastasi & Urbin
4. Objective
5. Standardized
6. Gregory
7. Testee, candidate, examinee
8. Tester, examiner, rater
9. Falchikov
10. Self-assessment
11. Peer-assessment
12. Computer Assisted Assessment (CAA)
13. Boud

جدول ۱-۲: مزایا و محدودیت‌های ارزیابی با کمک رایانه

مزایا	محدودیت‌ها
• عملکرد تعداد زیادی از دانشجویان در زمانی کوتاه ارزیابی می‌شود که منجر به صرفه‌جویی در وقت می‌گردد. • حجم وسیعی از محتوا در طول دوره آموزشی مورد ارزیابی قرار می‌گیرد. • از طریق انتخاب تصادفی سؤالات باعث کاهش بروز تقلب می‌شود. • نمره‌دهی تحت تاثیر خطای انسانی نیست و نیاز برای چک مجدد تصحیح را کاهش می‌دهد. • فراگیران می‌توانند از طریق خودارزیابی، میزان پیشرفت خود را به طور مستمر کنترل نمایند. • امکان ارائه بازخورد اختصاصی بلافاصله بعد از اجرای آزمون وجود دارد.	• به کارگیری آن توسط فراگیران نیازمند برخورداری از مهارت‌های رایانه است. • راه‌اندازی سیستم ارزیابی مبتنی بر رایانه می‌تواند هزینه بر باشد. • ارزیابان نیاز دارند که آموزش‌هایی را در زمینه کار با رایانه و مدیریت سیستم‌های ارزیابی فرا بگیرند. • اجرای آن نیازمند هماهنگی بسیار زیاد بین کارکنان، طراحان سیستم و مجرین است.

فراگیران است. با کمک گرفتن از رایانه می‌توان جدای از صرفه‌جویی در وقت و انرژی، به میزان قابل توجهی از بروز خطا در فرایند نمره‌دهی جلوگیری کرد. در جدول ۱-۲ شماری از مزایا و محدودیت‌های اجرایی ارزیابی به کمک رایانه آورده شده است.

تحلیل پس‌آزمون

حفظ کیفیت روش‌های ارزیابی به اندازه کیفیت فرایند یاددهی-یادگیری در هر فعالیت آموزشی اهمیت دارد. از این رو، امتحانات پس از اجرا باید با استفاده از مشخصه‌های روان‌سنجی به منظور شناسایی، کنترل و بهبود کیفیت فرایند و نتایج ارزیابی بررسی و تحلیل شوند. از نتایج تحلیل پس‌آزمون[1] که به صورت مختصر تحلیل آزمون نیز نامیده می‌شود، نه تنها می‌توان برای بهبود ارزیابی‌های آتی استفاده نمود بلکه همچنین می‌توان به بهبود کیفیت برنامه درسی و استراتژی‌های آموزشی کمک کرد (توکل و دنیک[2] ۲۰۱۱، توکل و دنیک ۲۰۱۲). تحلیل پس از آزمون، همچنین امکان شناسایی سؤالات نادرستی که ممکن است به کاهش کیفیت آزمون منتهی شوند را فراهم می‌سازد (رایت و استون[3] ۱۹۷۹). به طور کلی، منطق استفاده از روش‌های تحلیل پس آزمون، بهبود کیفیت و پایایی ارزیابی‌های به‌عمل آمده و انتخاب سؤالاتی است که به طور مناسب‌تر به سنجش عملکرد فراگیران می‌پردازند.

حساسیت و اهمیت آزمون

در طراحی و برنامه‌ریزی برای ارزیابی، مهم است که به درجه اهمیت نتایج حاصل از ارزیابی توجه شود. حساسیت آزمون به معنای با اهمیت‌تر بودن پیامدهای ارزیابی است. در صورتی که نتایج حاصل از آزمون تاثیر جدی بر سرنوشت آزمون‌شونده داشته باشند، می‌توان گفت که آزمون از درجه اهمیت بالایی برخوردار است و به آن اصلاحاً آزمون high stake گفته می‌شود. آزمون‌های ورودی دانشگاه‌ها و آزمون‌هایی که برای ارائه مدرک یا گواهی‌نامه حرفه‌ای برگزار می‌شود، نمونه‌های بسیار خوبی از آزمون‌های حساس و مهم هستند. این آزمون‌ها معمولاً توسط صاحب‌نظران یا نهادهای ملی طراحی می‌شوند و مسلماً در آنها توجه به حفظ کیفیت کلیه اجزاء آزمون و اطمینان از عادلانه بودن، روایی، پایایی و قابلیت تکرار ارزیابی، حیاتی است. اما این فاکتورها در ارزیابی‌های با خطرپذیری پایین از اهمیت کمتری برخوردار هستند.

1. Post examination analysis
2. Tavakol & Dennick
3. Wright & Stone

جدول ۲-۲: انواع ارزیابی از نظر اهمیت و سطح حساسیت (برگرفته از امین و همکاران ۲۰۰۶)

	اهمیت پایین	اهمیت متوسط	اهمیت بالا
نمونه ارزیابی	ارزیابی تکوینی	ارزیابی مداوم، ارزیابی درون بخشی	آزمون‌های ارائه مدرک حرفه‌ای و گواهی‌نامه
پیامدهای آموزشی	تصمیمات قابل‌برگشت یا عواقب ارزیابی خفیف	تصمیمات قابل‌لغو	تصمیمات غیرقابل‌لغو و پیامدهای ارزیابی بزرگ
فعالیت‌های ارزیابی	کم	متوسط	بالا
تضمین کیفیت	ندرتاً موردنیاز	توصیه‌شده	ضروری و لازم
سطح پایش و اجرا	استاد	گروه	دانشکده یا صاحب‌نظران
کنترل روایی و پایایی	معمولاً موردنیاز	توصیه‌شده	ضروری و لازم

در صورتی که پیامدهای حاصل از رد شدن در آزمون کم باشد و یا برگزاری آزمون‌های مجدد هزینه‌بر و دشوار نباشد، این آزمون‌ها به عنوان آزمون‌های با درجه اهمیت متوسط (medium stake) تا پایین (low stake) در نظر گرفته می‌شوند. تفاوت بین آزمون‌ها از این منظر به صورت خلاصه در جدول ۲-۲ آمده است.

ویژگی محتوا و تعمیم‌پذیری نتایج ارزیابی

مفهوم ویژگی محتوا[1] که در فارسی به آن اختصاصی بودن محتوا هم گفته‌اند، در متون به صورت ویژگی زمینه[2] (اختصاصی بودن زمینه) یا ویژگی مورد[3] (اختصاصی بودن مورد) نیز ذکر شده است. این عبارات به این موضوع اشاره دارند که توانمندی یا صلاحیت بالینی یا شایستگی بالینی[4] در پزشکی یک پدیده پیچیده است و جنبه‌ها و حیطه‌های مختلفی دارد که تعامل آنها با یکدیگر نهایتاً منجر به چیزی می‌شود که به عنوان صلاحیت بالینی می‌شناسیم. این امر موجب می‌گردد عملکرد پزشک در یک حیطه از صلاحیت بالینی (مانند درمان)، همبستگی مختصری با عملکرد وی در سایر حیطه‌ها (مانند تشخیص یا مهارت ارتباطی یا انجام پروسیجر) داشته باشد. همچنین عملکرد پزشک در ارتباط با یک بیمار، نتایج یکسانی با عملکرد وی در مواجهه با سایر بیماران نداشته باشد. یکی از مواردی که به کرّات در متون به آن اشاره شده است این است که در تمام صلاحیت‌های بالینی یک مهارت عمومی مشترک وجود ندارد. به عبارت دیگر مهارت‌هایی از قبیل شرح‌حال‌گیری، حل مسأله، استدلال تشخیصی، تصمیم‌گیری و برقراری ارتباط، به صورت مهارت‌های عمومی و ژنریک مطرح نیستند و در مورد تمامی ابعاد صلاحیت قابل تغییر می‌باشند (اپستین و هاندرت[5] ۲۰۰۲، نورمن[6] ۲۰۰۳).

پیامد منطقی این موضوع آن است که ارزیابی عملکرد دانشجو در یک بُعد مشخص از مشکل (به عنوان مثال درمان بیماری) نمی‌تواند ما را در مورد عملکرد وی در ارتباط با ابعاد دیگر آن مشکل (به عنوان مثال تشخیص بیماری) راهنمایی کند. به عبارت دیگر، ارزیابی عملکرد داوطلب در مواجهه با بیمار مبتلا به فشارخون ممکن است همبستگی ضعیفی با عملکرد همان داوطلب در شرایط دیگر (برای مثال در مواجهه با بیمار مبتلا به آرتریت روماتوئید) داشته باشد.

از همین رو، نمی‌توان در مورد شایستگی و صلاحیت یک داوطلب، تنها بر اساس عملکرد وی در یک مواجهه بالینی،

1. Content specificity
2. Context specificity
3. Case specificity
4. Clinical competency
5. Epstein & Hundert
6. Norman

با اطمینان قضاوت کرد. تنها روش کاربردی برای حذف ویژگی محتوا یا ویژگی زمینه، استفاده از راهبردهای متعدد نمونه‌گیری (مانند موارد بالینی متعدد، ارزیابان مختلف و سؤالات گوناگون) برای دستیابی به چشم انداز گسترده‌تری از عملکرد داوطلب است.

واژه‌ای که تقریباً با موضوع ویژگی محتوا یا زمینه در ارتباط می‌باشد، تعمیم‌پذیری[1] است که به چگونگی بسط نتایج یک ارزیابی به تعداد بیشتری از موقعیت‌ها یا شرایط اشاره دارد. به عبارت دیگر تعمیم‌پذیری به ما در مورد چگونگی پیش‌بینی عملکرد داوطلب در ورای آنچه که در آزمون اتفاق افتاده است، اطمینان می‌دهد. برای داشتن آزمونی پایا، معتبر و با نتایج تعمیم‌پذیر باید داوطلبان را در حیطه‌های متعدد مهارت‌های بالینی و در مواجهه با شرایط و بیماری‌های مختلف مورد ارزیابی قرار داد.

چرا نیازمند نمونه‌گیری متعدد در ارزیابی هستیم

نمونه گیری متعدد از بین مهارت‌های مختلف و حیطه‌های متفاوت مواجهه با بیماران (دانش، نگرش و مهارت) و سناریوهای بالینی متعدد و با ارزیابان گوناگون، برای دستیابی به یک آزمون معتبر و پایا لازم و ضروری است. بیشتر روش‌های قابل قبول در ارزیابی صلاحیت بالینی در حال حاضر استراتژی نمونه‌گیری متعدد را مورد استفاده قرار می‌دهند. روش‌هایی که از چنین استراتژی استفاده می‌کنند، شامل OSCE (نسبت به «مورد بالینی کامل») و سؤالات متعدد کوتاه‌پاسخ (نسبت به یک سؤال تشریحی واحد)، Mini-CEX ،DOPS و ارزیابی ۳۶۰ درجه است.
شواهد تجربی متعدد و چندین استدلال متقاعدکننده از این مسأله حمایت می‌کنند:

- صلاحیت بالینی در پزشکی وابسته به زمینه و نوع مشکل است و به عنوان یک مهارت عمومی محسوب نمی‌شود. به عبارت دیگر مهارت حل مسأله در یک موقعیت مشخص نمی‌تواند به یک صلاحیت مشابه در موقعیت دیگر تبدیل شود و یا میتوان گفت صلاحیت در یک موقعیت ویژه نمی‌تواند به دیگر موقعیت‌ها تعمیم یابد. به منظور بیان موضوع وابستگی به زمینه، نیازمند طراحی آزمون‌هایی هستیم که نمونه‌گیری متعدد داشته باشد، به نحوی که در مورد صلاحیت داوطلب در دامنه‌ای از موقعیت‌های بالینی اطمینان حاصل شود.
- احتمال بروز خطای سیستماتیک در آزمون‌های تک مواجهه (به طور مثال یک آزمون مورد بالینی کامل» با یا بدون آزمون شفاهی و سؤالات تشریحی) اثبات شده است. پایایی این چنین آزمون‌هایی به شدت پایین است. بیشتر اختلافات و تغییرپذیری در نمرات، اغلب در نتیجه عوامل مزاحم خارجی است تا تفاوت واقعی در سطح عملکرد دانشجویان. برآورد می‌شود که یک «مورد بالینی کامل» با آزمون شفاهی می‌تواند به میزان ۰/۳۹ پایا باشد که به معنای آن است که تقریباً دوسوم تغییرپذیری نمره‌دهی در نتیجه عواملی است که با صلاحیت داوطلب در آن موقعیت مشخص ارتباط ندارد.
- استفاده از استراتژی نمونه‌گیری متعدد (به طور مثال موقعیت‌های بالینی متعدد با آزمونگران مختلف) تغییرپذیری بین آزمونگران یا بین ارزیابان را کاهش می‌دهد. تغییرپذیری بین ارزیابان در بهترین موقعیت‌های آزمون نیز وجود دارد. اقداماتی مانند آموزش اعضای هیات علمی و استاندارد کردن الگوی تصحیح نیز احتمال بروز این خطا را کاهش می‌دهد. استفاده از نمونه‌گیری متعدد با ارزیابان متعدد و با تجربه، در صورتی که با استاندارد کردن الگوی نمره‌دهی و دیگر عملکردهای موثر در فرایند ارزیابی همراه شود، این خطا را به صورت چشم‌گیری کاهش خواهد داد.

بلوپرینت

در همین فصل در بحث روایی آزمون اشاره خواهد شد که برای یک دوره آموزشی به‌صورت بالقوه تعداد سؤالات زیادی می‌توان در نظر گرفت اما هنگام برگزاری آزمون از نظر اجرایی با محدودیت تعداد سؤال مواجه هستیم و باید از بین سؤالات، تعداد مشخصی را انتخاب کنیم. مسأله مهم، انتخاب سؤالات به گونه‌ای است که به خوبی معرف دوره باشند.

بلوپرینت[2] یا جدول مشخصات آزمون به ما کمک می‌کند تا قبل از اینکه وارد فرایند طرح سؤال شویم، استراتژی خود را برای انتخاب نمونه خوبی از سؤالات مشخص کنیم و در واقع مسیر ارزیابی را روشن کنیم. این کار مخصوصاً در دوره‌هایی که چندین مدرس وجود دارد، واقعاً ضروری است اما برای آزمون‌های دیگر نیز شدیداً توصیه می‌شود زیرا همان‌طور که بعداً مورد بحث قرار خواهد گرفت، روایی محتوایی[3] آزمون را تا حد زیادی تضمین می‌کند. بلوپرینت به اَشکال متفاوتی طراحی می‌شود.

اولین اقدامی که برای داشتن نمونه معرف و خوب به ذهن می‌رسد این است که سؤالات از «تمام» مباحث ارائه شده در طول

1. Generalizability
2. Blueprint
3. Content validity

دوره طرح شود. مثلاً اگر کتاب مرجع دارای ده فصل است، سؤالات از تمام فصول باشد و نه صرفاً از یکی دو فصل باشد. همین‌طور اگر درس شامل ۱۷ جلسه بوده، از تمام جلسات به تعداد مشخصی سؤال طرح شود. این روشِ معمولی است که استادان گروه به توافق برسند به ازای هر جلسه مثلاً ۲ یا ۳ سؤال در نظر گرفته شود. در جدول ۳-۲ بلوپرینت آزمون کورس قلب را برای دانشجویان پزشکی مشاهده می‌کنید که شامل ۱۲ جلسه بوده و کل آزمون حاوی ۶۰ سؤال است. این مثال ساده‌ترین شکل بلوپرینت است که تعداد سؤالات هر مبحث از تعداد جلسات آن مشخص می‌شود.

انتقادی که می‌توان به بلوپرینت فوق وارد کرد این است که ارزش و اهمیت همه فصول یا جلسات از منظر یادگیری دانشجو یکسان نیست. بنابراین دلیلی ندارد که تعداد سؤالات بر این مبنا مشخص شود. البته انتظار می‌رود که مسأله ارزش و اهمیت موضوعات قبلاً در تقسیم تعداد جلسات متجلی شده باشد اما می‌توان تصور کرد که برخی از موضوعات برای آموزش و درک مطلب به زمان بیشتری نیاز دارند، در حالی که ممکن است در هنگام ارزیابی با سایر مباحث یکسان باشند. در هرحال این نکته که صرف توجه به ساعات درس نمی‌تواند مبنای تقسیم‌بندی قرار گیرد، صحیح است. بنابراین تلاش می‌شود تا ابعاد دیگری نیز مدنظر قرار گیرند. یکی از ابعاد رایج مباحث در علوم بالینی، حیطه‌های توانمندی است یعنی در هر مبحث، کسب دانش مربوط به چه نوع وظیفه‌ای از دانشجو انتظار می‌رود. در این بلوپرینت جدول یک محور افقی دارد که در آن اهداف یادگیری، (یا رئوس محتوای دوره یا مشکلات بالینی) نوشته می‌شوند و یک محور عمودی دارد که حیطه‌های مورد نظر (اتیولوژی، پاتوفیزیولوژی، شرح‌حال، معاینه، تشخیص، درمان، پروگنوز یا ...) در آن فهرست شده‌اند. جدول ۴-۲ یک نمونه بلوپرینت را برای درس بیماری‌های محیطی مشخص کرده است. توجه کنید خانه‌ای که سیاه است، یعنی قرار است از آن سؤال طرح شود. در این حالت نه تنها سؤالات توزیع مناسبی از نظر موضوعات دارند، بلکه می‌توانیم مطمئن شویم که همه سؤالات در بحث تشخیص یا اتیولوژی متمرکز نشده‌اند و از این بابت هم واجد توزیع منطقی هستند. البته اینکه مثلاً چرا از اتیولوژی هایپوترمی سؤال طرح می‌شود و از تشخیص نیش بندپایان، مسأله‌ای است که به تخصص و اختیار طراح سؤال برمی‌گردد و در آزمون نوبت بعدی می‌تواند متغیر باشد. جدول ۵-۲ همان مثال آزمون کورس قلب را در این مدل جدید به نمایش می‌گذارد. در ابتدا تعداد سؤالات بر اساس تعداد ساعت جلسات مشخص شده (محور افقی) و سپس تعداد سؤالات هر حیطه هم مشخص شده (محور عمودی) و با توجه این دو بُعد، تعداد سؤالات هر مبحث به راحتی قابل تعیین است. توجه کنید که در اینجا نیز تعداد سؤالاتی که برای هر خانه تعیین شده است، ثابت نیست و بسته به نظر متخصصان برای سطوح مختلف فراگیران می‌تواند متفاوت باشد.

این مدل می‌تواند برای آزمون‌های عملی مانند OSCE نیز به کار رود. در بخش پنجم نمونه‌هایی از بلوپرینت OSCE را خواهید دید. همچنین این مدل می‌تواند برای دروس علوم پایه نیز به کار رود. کافی است در مورد هر آزمون به ابعاد

جدول ۳-۲: بلوپرینت بر اساس تعداد جلسه

عنوان جلسه	تعداد جلسه	تعداد سؤال
کل	۱۲	۶۰
فشارخون	۲	۱۰
آریتمی	۲	۱۰
نارسایی	۱	۵
ایسکمی	۳	۱۵
دریچه‌ای	۲	۱۰
سایر	۲	۱۰

جدول ٤-۲: بلوپرینت درس بیماری‌های محیطی بر اساس محتوای دوره و حیطه‌های توانمندی

عنوان‌جلسه	اتیولوژی	پاتوفیزیولوژی	شرح‌حال‌ومعاینه	تشخیص	درمان	disposition	پیش‌آگهی
Frostbite							❊
هیپوترمی	❊						
گرمازدگی							
نیش بندپایان				❊			
نیش مار					❊		
موجودات دریایی							
دیس‌باریم				❊			
سوختگی شیمیایی							
غرق‌شدگی							
سوختگی					❊		
برق‌گرفتگی		❊					
صاعقه‌زدگی			❊				
منوکسید کربن						❊	
مسمومیت با قارچ							
گیاهان سمی							

جدول۵-۲: بلوپرینت کورس قلب بر اساس محتوای دوره و حیطه‌های توانمندی

عنوان جلسه	تعداد جلسه	پاتوفیزیولوژی	اپیدمیولوژی	تشخیص	درمان	پیش آگهی	تعداد سؤال
کل	۱۲	۲۰	۵	۱۵	۱۵	۵	۶۰
فشارخون	۲						۱۰
آریتمی	۲						۱۰
نارسایی	۱						۵
ایسکمی	۳						۱۵
دریچه‌ای	۲						۱۰
سایر	۲						۱۰

بالقوه محتوای آن فکر کنیم و سپس محتوا را بر همان اساس تقسیم کنیم. یکی از ابعادی که می‌توان در نظر گرفت سطح شناختی سؤال است. برای اینکه تمام سؤالات محدود به سطح‌های پایین شناختی و حفظیات نشوند، می‌توان از ابتدا تعداد مشخصی از سؤالات را به سطوح بالاتر اختصاص داد. جدول۶-۲ یک مثال از بلوپرینت درس بیوشیمی را نشان می‌دهد.

اگر دوره آموزشی موردنظر ادغام‌یافته است، ردیف افقی جدول یعنی اجزای تشکیل‌دهنده دوره می‌تواند سیستم‌های مختلف بدن (قلب، تنفس، گوارش و غیره) باشد. همین ردیف افقی را می‌توان به اجزای خُردتری تقسیم کرد مثلاً با توجه

جدول ۶-۲: بلوپرینت درس بیوشیمی بر اساس محتوای دوره و سطوح شناختی

عنوان جلسه	اصول و حقایق	درک عمیق	استدلال	نوآوری
کربوهیدرات	❊		❊	
پروتئین		❊		❊
چربی		❊	❊	
مواد معدنی	❊			

به درجه اهمیت و میزان شیوع مشکلات و بیماری‌ها در هر یک از سیستم‌های بدن، مشکلات بالینی یا بیماری‌های خاصی تعیین شوند. سپس در محور عمودی، وظایف و توانمندی مورد انتظار در مقابل مشکل یا بیماری مربوطه مشخص گردند. توجه داشته باشید همانند جدول ۷-۲ برای هر بیماری می‌توان بیش از یک خانه را در نظر گرفت.

مسأله حائز اهمیت این است که هنگامی که صحبت از بلوپرینت در سطح یک برنامه می‌شود (و نه یک درس محدود)، معمولاً به ابزارهای ارزیابی متنوعی نیاز است. در مثال فوق مشاهده می‌کنید که مهارت ارتباطی و معاینه هم جزء اهداف دوره آمداند. از آنجا که انتظار نمی‌رود بتوان با سؤال چندگزینه‌ای به طرز مناسبی این توانمندی‌ها را ارزیابی کرد، باید به فکر انتخاب ابزار مناسب باشیم. پس نکته جالب این است که بلوپرینت به ما کمک می‌کند ابزار مناسب را انتخاب کنیم. به عنوان مثال، در صورتی که هدف آزمون، سنجش بُعد دانشی فراگیران است، از انواع آزمون‌های کتبی استفاده خواهد شد و چنانچه این انتظار می‌رود که دانشجو علاوه بر کسب دانش، بتواند اصول فراگرفته شده را در عمل نیز به کارگیرد، در این صورت استفاده از انواع آزمون‌های مبتنی بر محیط کار را می‌توان توصیه کرد. جدول ۸-۲ یک مثال از بلوپرینت دوره ادغام‌یافته و ابزارهای مناسب برای هر مبحث را نشان می‌دهد.

پس به صورت کلی بلوپرینت می‌تواند برای یک ابزار مشخص تدوین شود یا برای دوره‌ای که شامل ابزارهای متنوع است. مزیت استفاده از بلوپرینت این است که انتخاب ابزارهای مختلف را به صورت معنادار و در قالب یک چارچوب مفهومی برای ما مشخص می‌کند. پس از تهیه بلوپرینت آزمون، هر عضو هیأت علمی برای هر فعالیت مشخص شده، سؤالات آزمون را طراحی می‌کند. ابعاد دیگری نیز می‌توان به بلوپرینت داد که شامل جنس، سن، مشکل مزمن و حاد، شرایط اورژانس و پایدار و ... است. هر یک از این ابعاد برای طراحی آزمونی منطقی با توزیع متناسب سؤالات کمک‌کننده است.

جدول ۷-۲: بلوپرینت دوره ادغام‌یافته

سیستم	مشکل یا بیماری	اپیدمیولوژی	شرح‌حال	معاینه	تشخیص	پروسیجر	درمان	پروگنوز	ارتباطی
قلب	درد قفسه سینه نارسایی قلبی		❊		❊		❊		
تنفس	آمبولی آسم			❊			❊		❊
گوارش	خونریزی فوقانی التهابی روده	❊	❊			❊	❊	❊	
عصبی	سکته مغزی صرع	❊		❊	❊		❊	❊	

جدول ۸-۲: بلوپرینت دوره ادغام‌یافته و تعیین ابزار ارزیابی مناسب

سیستم	مشکل یا بیماری	اپیدمیولوژی	شرح‌حال	معاینه	تشخیص	پروسیجر	درمان	پروگنوز	ارتباطی
قلب	درد قفسه سینه نارسایی قلبی		OSCE		KF		MCQ		
تنفس	آمبولی آسم			OSCE			MCQ		OSCE
گوارش	خونریزی فوقانی التهابی روده	MCQ	OSCE			DOPS	KF	MCQ	
عصبی	سکته مغزی صرع	MCQ		OSCE	KF		MCQ	MCQ	

تقسیم‌بندی انواع ارزیابی

در ادامه تعریف اصلاحات و عبارات رایج باید گفت که اَشکال بسیار متنوعی از چگونگی انجام ارزیابی وجود دارد. به عبارت دیگر، ارزیابی‌ها را می‌توان از جنبه‌های گوناگون بررسی و تقسیم‌بندی کرد. یک نمونه از آن شامل ارزیابی تکوینی یا تراکمی است که در قسمت بعدی مفصلاً به آنها پرداخته خواهد شد. همچنین ارزیابی می‌تواند به صورت عینی یا ذهنی، رسمی یا غیررسمی، کمی یا کیفی، داخلی یا خارجی و ... صورت گیرد که در ادامه سعی می‌کنیم به صورت مختصر به ذکر هر یک از آنها بپردازیم:

ارزیابی عینی در برابر ذهنی[1]

ارزیابی ذهنی فرایندی است که در آن نمره فراگیر بر اساس قضاوت ذهنی آزمونگران تعیین می‌گردد. به همین دلیل این نوع ارزیابی می‌تواند تحت تاثیر عواملی که هدف اصلی سنجش نیستند، مانند وضعیت روحی و ترجیحات فردی مصححان، نحوه نگارش، دست‌خط فراگیر و ... قرار گیرد. در این رویکرد اگر افراد مختلف در یک زمان و یا یک نفر در زمان‌های متفاوت، برگه آزمون یک فراگیر را تصحیح نمایند، ممکن است نمرات متفاوتی به دست آید. برخلاف آن در ارزیابی عینی به دلیل مشخص بودن پاسخ از پیش تعیین‌شده، نتایج حاصل از اجرای آزمون همیشه ثابت است.

ارزیابی کمی در مقابل کیفی[2]

رویکرد کمی به ارزیابی در ارتباط با اندازه‌گیری‌های عددی است. به عنوان مثال، نمره فراگیر و یا رتبه او در رویکرد کمی از طریق شمارش تعداد پاسخ‌های صحیح، تعیین می‌شود.

در حالی که رویکرد کیفی با عدد و رقم سرو کار ندارد و ارزیابی عملکرد فراگیران در آن بر اساس میزان تطابق بین اهداف و پیشرفت فراگیران تعیین می‌شود. در این رویکرد نمره‌دهی بر اساس این پیش‌فرض است که آموزش چیزی فراتر از مذمت و توبیخ فراگیران است، بنابراین رویکردهای کیفی ارزیابی به پیچیدگی‌های یادگیری واقف بوده و معمولاً در ارتباط با فرایند یادگیری و از نوع معیار محور هستند (رانتری ۱۹۸۷). جدول ۹-۲ به مقایسه دو رویکرد کیفی و کمی در ارزیابی فراگیر پرداخته است.

1. Objective versus Subjective
2. Quantitative versus qualitative

جدول ۹-۲: رویکرد کمی و کیفی به ارزیابی (برگرفته از بیگز ۲۰۰۶)

ارزیابی کمی	ارزیابی کیفی
عملکردهای یادگیری را می‌توان به صورت یک جزء واحد در نظر گرفت.	یادگیری ساختار پیچیده‌ای دارد که در آن دانسته‌های جدید بر پایه دانش قبلی شکل می‌گیرد.
ارزیابی به صورت تحلیلی است و نمره فراگیر بر اساس ارزیابی اجزاء کوچک یادگیری بر طبق صحیح یا غلط بودن آن‌ها تعیین می‌شود (گاهی اوقات تعریف شیوه نمره‌دهی دلبخواهی است).	ارزیابی نیازمند تفکر در خصوص پیچیدگی‌های فرایند یادگیری است که این مستلزم داشتن دید کلی نگر نسبت به آن است.
تمامی اجزاء از ارزش یکسانی برخوردار هستند و نمره بر اساس تعداد پاسخ‌های صحیح به سؤالات محاسبه می‌شود.	ارزیابی برابر با جمع عددی نیست. نمره به معنای توصیف میزان خوب بودن پیشرفت فراگیر و اهداف مورد انتظار است.
عملکرد فردی داوطلبان ممکن است با یکدیگر مقایسه شود.	عملکرد هر داوطلب مستقل از سایرین در نظر گرفته می‌شود.

ارزیابی محصول در مقابل فرایند[1]

هرچند در عمل می‌توان هم محصول و هم فرایند یادگیری را مدنظر قرار داد، تصمیم‌گیری مشخص در این مورد به جهت‌دهی سایر فعالیت‌های ارزیابی کمک می‌کند.

- ارزیابی محصول شامل کلیه فعالیت‌های علمی است که به طور سنتی از قدیم در تمام موسسات آموزشی در حال اجرا است. به عنوان مثال، امتحانات پایان دوره و کارهای کلاسی که به شکل گزارش‌های آزمایشگاهی، پروژه و غیره ارائه می‌شوند، نمونه‌هایی از ارزیابی محصول یادگیری هستند.
- ارزیابی فرایند مشتمل بر تحلیل کلیه اتفاقاتی است که در بستر آموزش اتفاق می‌افتد. به عنوان مثال، بررسی میزان مشارکت فراگیران در بحث‌های کلاسی، تحلیل مهارت‌های کارگروهی و تیمی و برقراری ارتباط از جمله این موارد هستند.

ارزیابی داخلی در مقابل خارجی[2]

آیا ارزیابی در داخل مؤسسه و توسط استادان یا فراگیران انجام می‌شود یا نه افرادی خارج از موسسه مسؤولیت ارزیابی را بر عهده دارند؟ به عنوان مثال ارزیابی عملکرد کارآموزان در بخش‌ها توسط دستیاران نمونه‌ای از ارزیابی داخلی است. اما چنانچه سازمانی ملی به ارزیابی عملکرد فراگیران در انتهای دوره آموزشی بپردازد، ارزیابی از نوع خارجی خواهد بود.

تصمیم‌گیری در خصوص انجام ارزیابی داخلی یا خارجی به میزان زیادی به اهمیت کاربرد نتایج حاصل و امکانات موجود بستگی دارد. به طور معمول از ارزیابی خارجی زمانی استفاده می‌شود که قرار است قضاوتی سرنوشت‌ساز و بی‌طرفانه در مورد عملکرد داوطلبان صورت گیرد و اهمیت قضاوت‌های انجام شده به حدی است که ممکن است بر روند زندگی آنان تاثیر گذار باشد.

ارزیابی هنجارمحور در مقابل معیارمحور[3]

آزمون‌ها را می‌توان از منظر هدف آزمون و بر اساس نحوه تصمیم‌گیری در مورد رد و قبول دانشجویان به دو دسته هنجارمحور یا معیارمحور تقسیم کرد.

- در ارزیابی وابسته به هنجار، نتایج رد و قبول از مقایسه عملکرد فراگیران با یکدیگر حاصل می‌گردد و مشخص کننده این موضوع است که یک نفر در مقایسه با دیگران چگونه عمل کرده است. نتیجه عملکرد داوطلب در یک آزمون هنجارمحور،

1. Process versus product
2. Internal versus external
3. Norm-referenced versus Criterion-referenced

تحت تاثیر سطح و عملکرد کل فراگیران قرار دارد تا صرفاً میزان دستیابی خود او به اهداف به شکل مستقل و مجزا.

- در ارزیابی معیارمحور، نتایج رد و قبول حاصل از ارزیابی در قالب این‌که هر فراگیر تا چه حد به معیارهای مشخص و مورد انتظار یادگیری دست یافته است، ارائه می‌گردد. در این ارزیابی نتیجه عملکرد هر فراگیر در امتحان مستقل و جدا از سایر فراگیران خواهد بود.

انواع روش‌ها برای هر یک از دو نوع ارزیابی فوق‌الذکر در بخش هشتم کتاب به تفصیل مورد بحث قرار خواهد گرفت.

ارزیابی آغازین، تکوینی و تراکمی

۱) ارزیابی آغازین

نخستین ارزیابی که پیش از انجام فعالیت‌های آموزشی به اجرا در می‌آید، ارزیابی آغازین[1] نامیده می‌شود. این آزمون برای تعیین سطح فراگیران قبل از شروع دوره آموزشی اجرا می‌گردد و اطلاعات پایه‌ای که برای آموزش مورد نیاز است در اختیار مدرسان می‌گذارد. این نوع ارزیابی به منظور موارد زیر استفاده می‌شود:

- گاهی سنجش آغازین به منظور اندازه‌گیری مهارت‌ها و دانش‌های پیش‌نیاز یادگیری درس جدید به کار می‌رود. اطلاعات حاصل از اجرای این آزمون در واقع به این سؤال پاسخ می‌دهد که آیا یادگیرندگان بر دانش‌ها و مهارت‌های پیش نیاز درس جدید از قبل مسلط هستند. در این موارد از لفظ آزمون آمادگی[2] هم استفاده می‌شود.
- گاهی سنجش آغازین به بررسی سطح فعلی فراگیران در ارتباط با درس جدید می‌پردازد و مشخص می‌کند که یادگیرندگان چه مقدار از هدف‌ها و محتوای درس جدید را قبلاً یاد گرفته‌اند و آیا توانایی‌ها و مهارت‌های لازم را برای ورود به مطلب جدید دارند یا خیر. در این حالت از عبارت آزمون جایابی[3] یا پیش‌آزمون[4] استفاده می‌شود. آزمون جایابی در مراکز تعیین تناسب شغلی با ویژگی‌های داوطلبان کاربرد دارد و اجرای پیش آزمون در ابتدای دوره یا کارگاه آموزشی به کار می‌رود.
- گاهی نیز ارزیابی آغازین را به صورت ترکیبی از آزمون آمادگی و پیش‌آزمون اجرا می‌کنند که این حالت مطلوب‌تر به نظر می‌رسد.

۲) ارزیابی تکوینی

ارزیابی تکوینی[5] که به آن ارزیابی سازنده یا مستمر[6] نیز می‌گویند، زمانی به اجرا در می‌آید که فعالیت‌های آموزشی همچنان در جریان است. از نظر بولا[7] ارزیابی تکوینی بلافاصله بعد از آغاز به کار فعالیت آموزشی شروع می‌شود (بولا ۱۹۹۰). از آنجا که معمولاً تحقق هدف‌های آموزشی، به تدریج و به مرور زمان امکان‌پذیر می‌شود، به همین خاطر استاد تحقق هدف‌های آموزشی را در فاصله زمانی معینی، معمولاً در پایان هر بخش از مطالب تدریس شده و در طول سال تحصیلی متناسب با توانایی فراگیران و امکانات موجود مورد سنجش قرار می‌دهد.

زیربنای اصلی این نوع ارزیابی به این واقعیت بر می‌گردد که استاد در جریان تعامل خود با فراگیران می‌کوشد تا با توجه به هدف‌های آموزشی، به جهت‌دهی و سازمان‌دهی فعالیت‌ها و تجارب یادگیری فراگیران بپردازد. در این مسیر همواره این نگرانی وجود دارد که آیا یادگیری در حال اتفاق افتادن است؟ آیا حواس و ذهن فراگیران با موضوع یادگیری درگیر شده است؟ آیا فراگیران مطالب را یاد گرفته‌اند؟ و... استاد برای این که بتواند به این پرسش‌ها پاسخ دهد و براساس پاسخ آنها فرآیند یاددهی- یادگیری را درست پیش ببرد، مجبور است در خلال تدریس خود، مرتباً از فراگیران تأیید بگیرد که آیا درس را فرا گرفته‌اند؟ آیا به درس توجه دارند؟ آیا ابهامی در دریافت مطالب ندارند؟

---

1. Pre-assessment
2. Readiness Assessment
3. Placement test
4. Pre-test
5. Formative
6. Continuous
7. Bhola

سیر استفاده از «ارزیابی تکوینی»

منشا اولیه استفاده از «ارزیابی تکوینی» به مطالعات انجمن تحقیقات آموزشی امریکا بر می‌گردد که طی آن ها در سال ۱۹۶۷ اسکریون برای نخستین بار از این لغت در ارتباط با ارزشیابی برنامه‌های آموزشی استفاده نمود (اندرید و سیزک ۲۰۱۰). چندی پس از انتشار نتایج مطالعات صورت گرفته توسط اسکریون، به سرعت استفاده از این کلمه در بررسی اثربخشی برنامه‌های آموزش گسترش یافت.
با این حال، مفهوم ارزیابی تکوینی متعاقب انتشار «کتاب ارزشیابی تکوینی و تراکمی آموخته‌های فراگیران» توسط بنجامین بلوم شناخته شد و مورد استقبال همگانی قرار گرفت (بلوم و همکاران ۱۹۷۱). هرچند شاید شهرت اصلی این کتاب به دلیل ارائه تاکسونومی اهداف یادگیری بلوم باشد که پیش‌تر توسط او در سال ۱۹۵۶ ارائه شده بود، با این وجود شرح تفاوت بین ارزیابی تکوینی و تراکمی در ارتباط با سنجش دانسته‌های فراگیران در آن زمان بسیار مورد توجه موسسات آموزشی قرار گرفت. از زمان ارائه این مفهوم توسط اسکریون و بعدها توسط بلوم، مطالعات متعددی در این زمینه صورت گرفته است.

برعکس ارزیابی تراکمی که یک ارزیابی مبتنی بر پیامد است، این نوع ارزیابی مبتنی بر فرایند است. بدین ترتیب، ارزیابی تکوینی در پی پاسخ به این سؤال که فراگیران به چه سطح مشخصی از پیامد و یا اهداف کلی برنامه رسیده‌اند، نیست. ارزیابی تکوینی زمانی حاصل می‌شود که هدف ارزیابی تعیین ظرفیت بالقوه پیشرفت فراگیران است (جرج و کووان[1] ۱۹۹۱). از این رو، ارائه بازخورد به عنوان جزء کلیدی ارزیابی تکوینی در نظر گرفته می‌شود (بلک[2] ۱۹۹۶). ارزیابی تکوینی قلب آموزش اثربخش است و خودارزیابی به عنوان جزء حیاتی ارزیابی تکوینی به‌شمار می‌رود (بلک و ویلیام[3] ۱۹۹۶).

برای انجام ارزیابی تکوینی بسته به هدف‌های دوره تحصیلی، ویژگی‌های فراگیران، موضوع و محتوای دروس، روش تدریس، تعداد فراگیران و ... می‌توان از شیوه‌های اجرایی گوناگون استفاده نمود:

- انجام آزمون کتبی (چندگزینه‌ای، تشریحی، کوته‌پاسخ، جورکردنی و ...)
- استفاده از سؤال‌های شفاهی، تحلیلی و تبیینی
- یادداشت رویدادهای مهم عاطفی، روانی و حرکتی
- تهیه گزارش و خلاصه‌نویسی
- انجام آزمایش به صورت فردی و گروهی
- انجام پروژه‌های فردی و گروهی
- اجرای نمایش توسط فراگیران
- مشاهده عملکرد فراگیران در محیط‌های واقعی و بالینی و ارائه بازخورد
- جمع‌آوری و بررسی کار پوشه[4] یا مجموعه کارها
- برگزاری کنفرانس
- مشارکت فراگیران در ارزیابی از خود و یا سایر همکلاسی‌هایشان

به طور معمول مخاطبان اصلی نتایج حاصل از ارزیابی تکوینی استادان، برنامه‌ریزان و فراگیران هستند. این ارزیابی به فراگیران فرصت می‌دهد تا نسبت به آنچه یاد می‌گیرند، توجه کنند و با کسب آگاهی از نقاط قوت و ضعف خود، در جهت رسیدن به اهداف آموزشی، رفع نارسایی‌ها و تقویت جنبه‌های مثبت خود گام بردارند. همچنین برنامه‌ریزان آموزشی عمدتاً از اطلاعات حاصل از ارزیابی تکوینی برای بهبود و ارتقاء کیفیت بیشتر فرایندهای آموزشی استفاده می‌کنند. اهداف ارزیابی تکوینی به صورت خلاصه در جدول ۱۰-۲ آمده است.

1. George & Cowan
2. Black
3. Wiliam
4. Portfolio

جدول ۱۰-۲ : اهداف ارزیابی تکوینی

۱. فراهم آوردن شواهد معتبری در مورد یادگیری
۲. تعیین نقاط قوت و ضعف فراگیران در فرایند یاددهی و یادگیری
۳. تقویت اعتماد به نفس در فراگیران
۴. ایجاد انگیزه یادگیری در فراگیران
۵. پرورش روحیه تحقیق، تفکر، تلاش، ابتکار و خلاقیت فراگیران
۶. توجه به جنبه‌های مهم درس و اهداف آموزشی آن
۷. تشویق به استفاده از راهبردهای فعال یادگیری
۸. مشخص کردن نتایج و ارائه بازخورد تصحیح کننده به فراگیران
۹. کمک به فراگیران در پیگیری پیشرفت خود و توسعه مهارت‌های خود ارزیابی
۱۰. مطلع کردن فراگیران از سطح عملکرد مورد نیاز
۱۱. ایجاد انگیزه در فراگیران جهت تداوم یادگیری
۱۲. نمایان ساختن توانایی‌های بالقوه‌ی فراگیران که موجب خودشناسی و تبین تصویری ذهنی فراگیر از خود می‌شود
۱۳. ارزیابی برای شناخت توانایی و زمینه‌های علمی فراگیران و تصمیم‌گیری برای انجام دادن فعالیت‌های بعدی آموزشی
۱۴. ارزیابی به عنوان وسیله‌ای برای شناساندن هدف های آموزشی در فرایند یاددهی-یادگیری
۱۵. ارزیابی به عنوان وسیله‌ای برای بهبود و اصلاح فعالیت های آموزشی

۳) ارزیابی تشخیصی

از آنجا که گفته شد ارزیابی تکوینی در طول دوره انجام می‌شود، ذکر این نکته لازم است که در متون به نوع دیگری از ارزیابی هم اشاره شده است که آن هم در جریان آموزش صورت می‌گیرد و به آن ارزیابی تشخیصی[1] می‌گویند. علت این نام‌گذاری جداگانه آن است که این ارزیابی با هدف تشخیص مشکلات یادگیری فراگیران به کار می‌رود. این آزمون‌ها معمولاً به صورت انفرادی اجرا می‌شوند و در قیاس با آزمون‌های پیشرفت تحصیلی، سؤال‌های بیشتری را در برمی‌گیرند. این ارزیابی زمانی مورد استفاده قرار می‌گیرد که استاد با مشکلات مبرم و مکرر در یک یا چند فراگیر رو به رو می‌شود که با روش‌های اصلاحی معمول ارزیابی تکوینی قابل رفع نیست. در آزمون تشخیصی نمره کل فراگیران اهمیت چندانی ندارد، بلکه نحوه پاسخ‌دهی فراگیر به سؤالات مورد توجه است که نشان‌دهنده اشتباهات رایج فراگیران در کسب هدف‌های آموزشی است.

۴) ارزیابی تراکمی

در پایان هر دوره آموزشی، لازم است ارزیابی جامعی از میزان آموخته‌های فراگیران به‌عمل آید که به آن ارزیابی پایانی[2]، تراکمی[3] یا تجمعی می‌گویند که به عنوان متداول‌ترین شکل ارزیابی ذکر شده است (امین و همکاران ۲۰۰۶). معمولاً ارزیابی تراکمی در انتهای نیمسال تحصیلی و یا در زمان اتمام یک برنامه آموزشی صورت می‌گیرد اما امتحانات میان‌ترم که با هماهنگی قبلی در برنامه دانشجویان گنجانده شده است و قسمتی از نمره نهایی را شامل می‌شود نیز نوعی از ارزیابی تراکمی هستند.

هدف از انجام این نوع ارزیابی، بررسی این موضوع است که فراگیران به چه میزان مطالب ارائه شده را فرا گرفته‌اند و در چه حد به توانمندی‌های مورد انتظار رسیده‌اند. به صورت کلی از آنجا که ارزیابی تراکمی عموماً در قالب دریافت نمرات پایان دوره تجلی پیدا می‌کند، اکثر فراگیران از نتیجه امتحان می‌هراسند، چرا که آینده آن‌ها وابسته به این نمره است. بنابراین، برخلاف ارزیابی تکوینی که اشتباهات فراگیران باعث شناسایی نقاط ضعف فراگیران و کمک به آن‌ها در یادگیری بهتر

1. Diagnostic
2. Terminal
3. Summative

جدول ۱۱-۲ : اهداف ارزیابی تراکمی (برگرفته از امین و کو ۲۰۰۳)

۱. تعیین اینکه آیا فراگیران به سطح مشخصی از توانمندی رسیده‌اند یا خیر؟ (تصمیم‌گیری در خصوص پذیرش یا رد فراگیر)
۲. تعیین اینکه چه میزان از اهداف آموزشی محقق شده‌اند (به عنوان مثال تعیین ارزش یک برنامه آموزشی)
۳. مقایسه بین فعالیت‌های آموزشی متعدد و انتخاب بهترین آن‌ها

مطالب می‌شود، در این ارزیابی فراگیران تمایلی به پذیرش اشتباهات خود ندارند (بیگز ۲۰۰۶).

نتیجه حاصل از این ارزیابی به عنوان نمره نهایی فراگیران محسوب می‌شود. ارزیابی تراکمی به عنوان ابزاری برای اخذ تصمیم در خصوص عملکرد فراگیران در نظر گرفته می‌شود (جورج و کووان ۱۹۹۹) که از این قضاوت‌ها برای مقاصد تشخیصی و صدور مدرک نیز استفاده می‌شود (نوو[1] ۱۹۹۵). البته تنها هدف ارزیابی تراکمی اختصاص دادن نمره به دانشجو و تعیین وضعیت رد و قبول وی نیست. هدف از ارزیابی تراکمی، نمره دادن به دانش آموزان و همین‌طور قضاوت درباره اثربخشی آموزش است (بولا ۱۹۹۰). برنامه‌ریزان آموزشی از ارزیابی تراکمی استفاده می‌کنند تا در خصوص ارزش و جایگاه برنامه آموزشی تصمیم‌گیری کنند. از این رو، این نوع ارزیابی شکل رسمی‌تر و مشخص‌تری به خود می‌گیرد (امین و کو ۲۰۰۳). در این شرایط، کسب اطلاعات دقیق به منظور اطمینان از تضمین کیفیت تصمیم‌گیری، حیاتی است.

به طور خلاصه می‌توان گفت، هدف ارزیابی تراکمی تعیین میزان یادگیری فراگیران در طول یک دوره آموزشی به‌منظور نمره دادن و صدور گواهینامه، یا قضاوت درباره اثربخشی کار استاد و برنامه درسی، یا مقایسه برنامه‌های مختلف با یکدیگر است. هر چند از داده‌های حاصل از ارزیابی تراکمی می‌توان به منظور بهبود فعالیت‌های آموزشی آتی استفاده نمود، با این حال این موضوع، هدف اصلی ارزیابی تراکمی نیست. اهداف ارزیابی تراکمی به صورت خلاصه در جدول ۱۱-۲ آمده است.

ارزیابی تراکمی همواره در آموزش پزشکی بسیار مورد توجه بوده است (امین و کو ۲۰۰۳). این موضوع منجر شده است تا ما به غلط ارزیابی تراکمی را معادل واژه کلی ارزیابی در نظر بگیریم و به این ترتیب نقش مهم ارزیابی تکوینی را در فرایند آموزشی نادیده بنگاریم. این در حالی است که اگر این نوع ارزیابی صورت نگیرد و تنها به ارزیابی تراکمی اکتفا شود، دیگر فرصتی برای تغییر و اصلاحات برنامه آموزشی پیش نمی‌آید. انجام ارزیابی تراکمی خوب عمدتاً وابسته به انجام ارزیابی‌های تکوینی با کیفیت است. ارزیابی تکوینی در صورتی که به خوبی سازمان‌دهی شود، به بهبود عملکرد فراگیران در طول دوره آموزشی منتهی می‌شود و در نهایت دستیابی به پیامدهای مطلوب را در ارزیابی تراکمی مهیا می‌سازد. از این رو توصیه می‌شود که ارزیابی تکوینی و تراکمی به دلیل نقش مکملی که در فرایند ارزیابی بازی می‌کنند، به یک نسبت مورد توجه قرار گیرد. در جدول ۱۲-۲ مقایسه این دو نوع ارزیابی از چند منظر آمده است.

سیستم ارزیابی مبتنی بر توانمندی

سیستم یا برنامه ارزیابی

با نگاه اجمالی به فرایند ارزیابی، می‌توان آن را همانند سیستم و چرخه‌ای فرض کرد که کلیه اجزاء آن به نوعی با یکدیگر مرتبط می‌باشند. این فرایند که بالتبع با تدوین پیامدهای یادگیری در بدو فرایند آموزش شروع و تا تحلیل پس‌آزمون ادامه می‌یابد، بی‌شک فارغ از بروز هیچ گونه خطا نیست. آنچه در این میان حائز اهمیت است، توجه به نقش مهم انجام ارزشیابی و ارائه بازخورد به منظور به حداقل رساندن خطاهای احتمالی است. در بخش نهم کتاب این موضوع تحت عنوان سیستم یا برنامه ارزیابی به صورت تفصیلی مورد بحث قرار می‌گیرد. اما نکته مهم اینکه برخلاف باور نادرست رایج که آغاز فرایند ارزیابی را از فاز انتخاب روش

1. Nevo

جدول ۱۲-۲ : مقایسه بین ارزیابی تکوینی و تراکمی (برگرفته از امین و کو ۲۰۰۳)

ارزیابی تکوینی		ارزیابی تراکمی
زمان	اجرا در طول دوره آموزشی	اجرا عمدتاً در انتهای دوره آموزشی یا در زمان از پیش تعیین شده
اقدام	ارائه بازخورد به منظور بهبود	ثبت پیشرفت
هدف	راهنمایی و هدایت به سمت توسعه حرفه‌ای	تصمیم‌گیری در خصوص عملکرد حرفه‌ای
آزمون‌گر	استادان	استادان و هیأت‌های حرفه‌ای

ارزیابی یا طراحی سؤال در نظر می‌گیرد، در ابتدایی‌ترین مرحله طراحی برنامه ارزیابی باید به پیامدهای یادگیری[1] مناسب توجه شود.

توانمندی

یکی از رویکردهای آموزشی که از حدود ۱۹۶۰ در متون آموزش پزشکی پررنگ شده است، برنامه آموزشی مبتنی بر پیامد[2] است که در آن به محصول نهایی توجه می‌شود. بر خلاف رویکرد سنتی که در آن موضوعات درسی و محتوای برنامه، اساس برنامه‌ریزی را تشکیل می‌دهند، در برنامه مبتنی بر پیامد، محور و نقطه آغاز برنامه‌ریزی، توانمندی‌های[3] پایه و اصلی[4] هستند که لازم است دانش‌آموخته در پایان دوره کسب کرده باشد.

فرهنگ لغات آکسفورد در تعریف توانمندی، آن را «برخورداری از دانش و توانایی انجام موفقیت‌آمیز برخی کارها» تعریف می‌کند. پژوهشگران زیادی برای تعریف این واژه وارد عمل شدند. یکی از تعاریف رایج توانمندی عبارت است از عملکرد تعریف شده در برنامه که دانش‌آموخته باید بتواند در حد تسلط[5] آن را به انجام برساند (اسمیت[6] ۱۹۹۹). به بیان ساده‌تر می‌توان گفت توانمندی، توانایی انجام مجموعه‌ای از وظایف و یا نقش‌ها به شکل موثر یا کافی است که فراگیران بعد از طی دوره آموزشی باید قادر به انجام آن باشند.

در آموزش پزشکی، توانمندی معنایی فراتر از کسب دانش و درک دارد و یک فراگیر زمانی توانمند محسوب می‌شود که بتواند مجموعه وظایفی را که توسط بورد حرفه‌ای به عنوان شرط لازم برای عمل به عنوان یک پزشک مستقل در نظر گرفته شده است، به شکل مناسب انجام دهد. مفهوم توانمندی در حوزه علوم پزشکی به معنای برخورداری از قضاوت صحیح و عادت به استفاده از دانش، مهارت‌های فنی و ارتباطی، استدلال بالینی، احساسات، ارزش‌ها و بازاندیشی در فعالیت‌های روزانه با هدف ارائه خدمت به جامعه و افراد است (اپستین و هاندرت ۲۰۰۲). پزشک توانمند قادر است خدمات پزشکی و یا سایر خدمات حرفه‌ای را در ارتباط با استانداردهای وضع شده به طریقی که انتظارات جامعه را محقق سازد، ارائه نماید (ویتکمب[7] ۲۰۰۲). توانمندی‌ها به طور معمول در برگیرنده هر سه جزء دانشی، مهارتی و نگرشی می‌باشند و دانسته‌های فرد صرفاً به عنوان توانمندی محسوب نمی‌شوند بلکه آنچه دانشجو در مواجهه با مسائل کاری واقعی باید انجام دهد، در این تعریف می‌گنجد. با توجه به نقش کلیدی توانمندی در برنامه آموزشی، اهمیت آن در بحث ارزیابی فراگیران نیز تجلی پیدا می‌کند. در واقع لازم است ارزیابی فراگیر به نوعی باشد که به بهترین شکل توانمندی او را در مواجهه با وظایفش نشان دهد.

1. Learning outcomes
2. Outcome Based Curriculum, Outcome Based Education
3. Competency
4. Core
5. Mastery
6. Smith
7. Whitcomb

چارچوب توانمندی‌ها

سازمان‌های مختلف تلاش کرده‌اند تا با تدوین فهرست توانمندی‌هایی که از دانش‌آموختگان رشته‌ها و مقاطع گوناگون انتظار می‌رود، به صورت شفاف منظور خود را از این واژه بیان کنند. در حال حاضر چارچوب‌های گوناگون توانمندی در مقطع پزشکی عمومی و دوره‌های تخصصی وجود دارد که توسط دانشگاه‌ها و سازمان‌های مختلف تهیه شده است:

- در مقطع پزشکی عمومی، می‌توان پزشکان فردا[1] در انگلیس، حداقل الزامات عمومی[2] تهیه‌شده توسط انستیتوی بین‌المللی آموزش پزشکی[3]، توانمندی‌های دانش‌آموختگان دانشگاه‌های براون[4]، ایندیانا[5]، دانذی و بارسلونا را نام برد. دانشگاه علوم پزشکی تهران نیز توانمندی‌های مورد انتظار خود از دانش‌آموخته پزشکی عمومی را در قالب هشت حیطه تدوین نموده است.
- در مقطع تخصصی می‌توان به چارچوب CanMEDs[6] که توسط کالج سلطنتی پزشکان و جراحان کانادا[7] طراحی شده است و توانمندی‌های پزشکی را از منظر نقش‌های یک پزشک توصیف می‌کند، اشاره کرد. عملکرد پزشک توانمند نیز که توسط شورای پزشکی عمومی[8] انگلیس طراحی شده است، عناصر عملکرد خوب پزشکان را شرح می‌دهد. شورای اعتباربخشی آموزش پزشکی تخصصی[9] و بورد تخصصی پزشکی در آمریکا[10] شش توانمندی کلی را برای فارغ التحصیلان پزشکی برگزیده‌اند که چارچوبی از پیامدها را برای طراحی برنامه‌های آموزشی دوره‌های دستیاری و فلوشیب فراهم می‌سازد. موسسه پزشکی[11] در امریکا نیز پنج صلاحیت پایه‌ای توصیه کرده است که در حکم چارچوبی برای ارزیابی عملکرد و ایجاد انگیزه برای حرکت به سمت اصلاح در آموزش پزشکی است.

همان‌طور که در جدول ۱۳-۲ نمایش داده شده است، علی‌رغم آنکه تفاوت‌هایی در شماری از اصول این چارچوب‌ها دیده می‌شود، اما با اندکی تأمل می‌توان دریافت که همپوشانی قابل ملاحظه‌ای در توصیف عملکردهای یک پزشک از منظر آن‌ها وجود دارد. از موارد فوق سه چارچوب با جزئیات بیشتری شرح داده می‌شوند:

چارچوب توانمندی ACGME

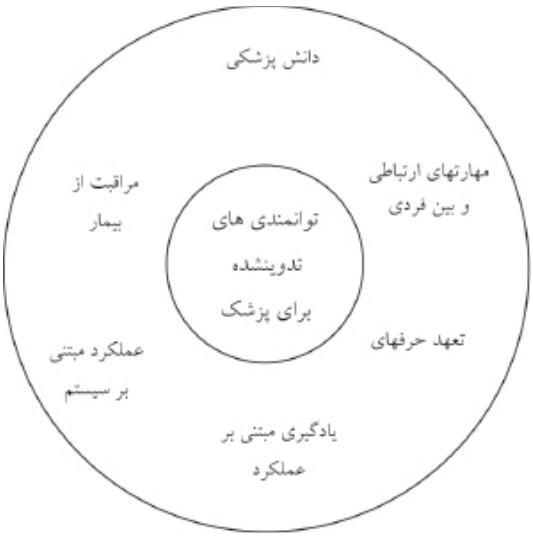
شش حیطه توانمندی تدوین‌شده توسط شورای اعتباربخشی آموزش پزشکی تخصصی که در آمریکا فعالیت می‌کند، به شرح زیر است (شکل ۱-۲):

- **دانش پزشکی[12]:** دانشجویان، دستیاران و پزشکان باید از دانش پزشکی (علوم پایه و بالینی) برخوردار باشند و قادر باشند آن‌ها را در فرایند مراقبت از بیمار نیز به کار گیرند. به‌علاوه از آنها انتظار می‌رود که بتوانند مشکلات بالینی را بر طبق یک رویکرد مناسب تحلیل، تفسیر و استدلال نمایند.
- **مهارت‌های ارتباطی و بین فردی[13]:** دانشجویان، دستیاران و پزشکان به منظور تبادل اطلاعات به شکل موثر با سایر اعضای تیم سلامت و بیماران نیازمند برخورداری از مهارت‌های ارتباط بین‌فردی هستند. کسب این توانمندی قطعاً در شکل‌گیری و تداوم ارتباط درمانی و اخلاقی بین بیمار و پزشک بسیار تاثیرگذار است.

1. Tomorrow Doctors
2. Global Minimal Essential Requirements
3. International Institute of Medical Education (IIME)
4. Brown University
5. Indiana University
6. Canadian Medical Education Directions for Specialists (CanMeds)
7. Royal College of Physicians and Surgeons of Canada
8. General Medical Council (GMC)
9. Accreditation Council for Graduate Medical Education (ACGME)
10. American Board of Medical Specialist (ABMS)
11. Institute of Medicine (IOM)
12. Medical knowledge
13. Interpersonal & communication skills

جدول ۱۳-۲: توانمندی‌های مورد انتظار پزشکان از چند دیدگاه (برگرفته از هولمبو و هاوکینز ۲۰۰۸)

CanMEDS	GMC	ACGME	IOM	RIME	دانشگاه علوم پزشکی تهران
متخصص	مراقبت بالینی	دانش پزشکی	عملکرد مبتنی بر شواهد	گزارش‌گر	مهارت‌های بالینی
ارتباط‌دهنده	ارتقاء عملکرد حرفه‌ای	مهارت های ارتباطی و بین‌فردی	کار در تیم بین رشته‌ای	تفسیرگر	مهارت‌های ارتباطی
هماهنگ‌کننده	تدریس، نقد و ارزیابی آموزشی	مراقبت از بیمار	ارائه مراقبت بیمارمحور	مدیر	مراقبت بیمار
مدیر	ارتباط با بیمار	تعهد حرفه‌ای	کاربرد اصول ارتقاء کیفیت	معلم	تعهدحرفه‌ای، اخلاق و حقوق پزشکی
مدافع سلامت	کار با همکاران	یادگیری‌مبتنی‌برعملکرد	استفاده از انفورماتیک		استدلال، تصمیم‌گیری و حل مسأله
پژوهشگر	صداقت	عملکرد مبتنی بر سیستم			رشد فردی
حرفه‌ای	سلامتی				پیشگیری و ارتقای سلامت
					نظام سلامت و نقش پزشک در آن



شکل ۱-۲: توانمندی‌های مورد انتظار تدوین‌شده توسط شورای اعتباربخشی آموزش پزشکی تخصصی در آمریکا

- **مراقبت از بیمار[1]:** مراقبت مناسب از بیماران نیازمند وجود پزشکانی حاذق و دلسوز است. آن‌ها باید قادر باشند که از یک سو به شکل موثر با بیماران ارتباط برقرار نمایند و از سوی دیگر بتوانند بر اساس شواهد و اطلاعات موجود به ارائه خدمات بپردازند. در این فرایند انتظار می‌رود فارغ‌التحصیل پزشکی قادر باشد با به کارگیری دانش پزشکی و در نظر گرفتن ترجیحات بیمار، در خصوص برنامه درمانی وی بهترین تصمیم‌گیری را داشته باشد.

1. Patient care

- **تعهد حرفه‌ای[1]:** دانشجویان، دستیاران و پزشکان باید در عین وفاداری به اصول اخلاقی، متعهد به انجام مسؤولیت‌های حرفه‌ای خود در قبال بیماران باشند. تعهد حرفه‌ای شامل داشتن میل به تعالی شغلی، نوع دوستی، عدالت، صداقت و پاسخگویی است.
- **یادگیری مبتنی بر عملکرد[2]:** از دانشجویان و پزشکان انتظار می‌رود تا با به کارگیری شواهد علمی و استفاده از روش‌های مختلف برای پژوهش و ارزیابی، به ارائه خدمات با کیفیت برای بیماران بپردازند. مسلماً کسب این توانمندی نیازمند آن است که دانشجو بتواند علاوه بر کشف فرصت‌های بهبود، شناسایی و اصلاح خطاهای پزشکی، از فناوری اطلاعات برای ارائه خدمات موثر به بیماران استفاده کند.
- **عملکرد مبتنی بر سیستم[3]:** دانشجویان و پزشکان نیازمند کسب یک درک عمیق از سیستم‌های ارائه خدمات هم در سطح خُرد و هم در سطح کلان هستند. دانشجویان باید یاد بگیرند که از دانسته‌هایشان برای بهره‌گیری کارآمد از منابع موجود برای ارائه خدمات موثر نه تنها در سطح فردی بلکه جامعه تحت پوشش استفاده کنند. برای این توانمندی کسب دانش، مهارت و نگرش انجام کار تیمی حیاتی است.

چارچوب توانمندی‌های دوره پزشکی عمومی دانشگاه علوم پزشکی تهران

در این قسمت به معرفی اجمالی اجزای سند توانمندی دانش‌آموختگان پزشکی دانشگاه علوم پزشکی تهران می‌پردازیم:

- **مهارت‌های بالینی:** دانش‌آموخته دوره پزشکی عمومی دانشگاه علوم پزشکی تهران باید توان‌مندی لازم را در طیف گسترده مهارت‌های بالینی، شامل گرفتن شرح‌حال و معاینه بالینی، ثبت و ارایه اطلاعات پزشکی حاصل از آن‌ها و انجام اقدامات عملی (پروسیجرها) و تست‌های آزمایشگاهی طبق استانداردهای تعیین شده داشته باشد.
- **مهارت‌های برقراری ارتباط:** دانش‌آموخته دانشگاه باید توانایی لازم را برای برقراری ارتباط مؤثر با بیماران، همراهان بیمار و همکاران خود داشته باشد. علاوه بر این وی باید بتواند صلاحیت خود را در برقراری ارتباط در تمام عرصه‌ها به صورت شفاهی، نوشتاری، الکترونیکی یا تلفنی نشان دهد.
- **مراقبت بیمار (تشخیص، درمان، بازتوانی):** دانش‌آموخته دانشگاه باید با داشتن دید کل‌نگر به بیمار توانایی تهیه فهرستی از مشکلات بیمار و تشخیص‌های افتراقی، انتخاب روش تشخیصی مناسب و تعیین برنامه مراقبتی به منظور دستیابی به اهداف مورد نظر در مواجهه با مشکل بیمار را داشته باشد. در ضمن وی باید بتواند شرایط خاصی را که نیاز به مشاوره یا ارجاع به متخصص مربوطه است، تشخیص دهد. از دانش‌آموخته دانشگاه انتظار می‌رود بتواند در جنبه‌های مهم مراقبت از بیمار از جمله اقدامات طبی و جراحی، تجویز دارو، تغذیه، مراقبت در موارد حاد و مزمن و اورژانس، کنترل درد و بازتوانی، توانایی‌های خود را نشان دهد.
- **ارتقای سلامت و پیشگیری:** دانش‌آموخته دانشگاه به منظور همکاری یا راهبری گروه ارائه‌دهندگان خدمات در جهت ارتقای سطح سلامت در فرد و جمعیت در تماس، باید توانایی ارزیابی وضعیت سلامت، تعیین عوامل خطرساز، شناسایی علل بیماری‌ها و عوامل تعیین‌کننده پیش‌آگهی آنها را داشته باشد. او باید بتواند به عنوان عضوی از تیم سلامت، راهبردهای متناسب ارتقای سلامت در سطوح پیشگیری ابتدایی، اولیه و ثانویه را به عنوان مداخلات مورد انتظار انتخاب کرده و به کار برد.
- **رشد فردی:** دانش‌آموخته دانشگاه باید اهمیت رشد فردی از جمله ارتقای مراقبت از خود، توانایی‌های ذهنی، روانی، اجتماعی، اقتصادی و شغلی را بپذیرد و دانش‌های غیرپزشکی مؤثر در زندگی فردی و حرفه‌ای مانند خودشناسی، روان‌شناسی تغییر، اصول رهبری و مدیریت، و دانش انفورماتیک را بداند و به کار بندد.

1. Professionalism
2. Practice based learning & improvement
3. System based practices

- **تعهد حرفه‌ای، اخلاق و حقوق پزشکی:** دانش‌آموخته دانشگاه باید با باور به این که شفای بیماران به دست خداوند است و وی از سوی او این توفیق را پیدا کرده است تا وسیله آن را فراهم کند، مجموعه ارزش‌ها، خصوصیات و رفتارهایی را که متضمن اعتماد جامعه به حرفه پزشکی هستند، به عنوان تعهدات حرفه‌ای پزشکی بپذیرد و در طبابت خود به کار بندد. او همین‌طور باید پایبند به رعایت سوگندنامه و راهنماهای اخلاق پزشکی منبعث از ارزش‌های انسانی و معارف اسلامی باشد و بداند تقوای الهی مبنای رعایت تعهد حرفه‌ای پزشک است. همچنین باید توانایی شناسایی مسائل اخلاقی را در طبابت خود داشته باشد و بتواند ضمن توجه به الزامات قانونی و اخلاقی و با احترام به فرهنگ و باورهای افراد ذی‌نفع، در مورد این مسائل تحلیل و تصمیم‌گیری مناسب انجام دهد.
- **مهارت‌های تصمیم‌گیری، استدلال و حل مسأله:** دانش‌آموخته دانشگاه باید در رویارویی با یک مسأله، قادر به شناسایی مشکل و ابعاد آن باشد، توانایی جمع‌آوری و ارزیابی اطلاعات مرتبط را از بهترین منابع در دسترس داشته باشد، راه‌حل‌های مختلف را شناسایی و ارزیابی نماید، قادر به برآورد احتمال پیامدهای هریک باشد و سرانجام مناسب‌ترین گزینه را با توجه به شرایط عدم قطعیت در هنگام تصمیم‌گیری انتخاب کند. او باید بتواند جهت اخذ تصمیم نهایی، این توان‌مندی را با اطلاعات خود در حوزه‌های دیگر مانند اولویت‌ها و ارزش‌های مورد قبول خدمت‌گیرندگان و جامعه و همچنین هزینه–اثربخشیِ راه‌حل‌های ممکن، ادغام کند.

- **نظام سلامت و نقش پزشک در آن :** دانش‌آموخته دانشگاه باید در نظام و شبکه سلامت به عنوان پزشک، آموزش‌دهنده، پژوهش‌گر، مدیر واحد ارایه خدمات سلامت و راهبر سلامت ایفای نقش کند.

مدل RIME

این مدل به توصیف حداقل سطوح مورد انتظار عملکرد دانشجویان پزشکی در محیط‌های بالینی می‌پردازد. بر طبق مدل RIME[1] چهار سطح مورد انتظار عملکرد کارآموز در بخش‌های بالینی مشتمل بر گزارش‌گر[2]، تفسیرگر[3]، مدیر[4] و معلم[5] است. به عنوان مثال، چنانچه یک دانشجوی پزشکی، عملکرد قابل‌قبولی به عنوان «گزارش‌گر» در جمع‌آوری اطلاعات صحیح در خصوص علایم بیماران، یافته‌های حاصل از معاینات فیزیکی و آزمایشگاهی نشان ندهد، اجازه رفتن به سال‌های بالاتر به او داده نخواهد شد. به دنبال معرفی این مدل، به دلیل سهولت پیاده‌سازی آن در بخش‌های مختلف بالینی (بتیستون و همکاران[6] ۲۰۰۲)، به سرعت در آمریکای شمالی گسترش یافت (همر و همکاران[7] ۲۰۰۸). این مدل در واقع بسط ساده‌ای از نحوه مراقبت از بیمار شامل جمع‌آوری یافته‌های بالینی، تفسیر آنها و تدوین برنامه تشخیص، درمان، مشاوره و آموزش به بیمار فراهم می‌سازد.

در جدول ۱۴–۲، مثالی از سطوح مورد انتظار عملکرد دانشجویان بر طبق مدل RIME ارائه شده است.

ارزیابی مبتنی بر توانمندی

درباره ارزیابی توانمندی باید گفت که یک فرایند چندوجهی است، شامل اندازه‌گیری رفتارها و ویژگی‌های پیچیده و

1. Reporter-Interpreter- Manager-Educator
2. Reporter
3. Interpreter
4. Manager
5. Educator
6. Battistone et al.
7. Hemmer et al.

جدول ۱۴-۲: سطوح مورد انتظار عملکرد دانشجویان بر طبق مدل RIME (برگرفته از پنگارو و تن کیت ۲۰۱۳)

۶ ماه	۱۲ ماه	۱۲ ماه تا ۱۸ ماه	۲۴ ماه
گزارش‌گر	تفسیرگر	مدیر	معلم
اخذ شرح حال مرتبط و صحیح از بیمار	ترکیب اطلاعات موجود، تشخیص مشکل اصلی بیمار	تحت نظارت استاد، بیماران مبتلا به تظاهرات بالینی متداول	تدوین سیستم پیگیری و بازاندیشی مشکلات بالینی

متعدد که اجزاء دانشی، مهارتی و نگرشی را در بر می‌گیرد (کراچیو و همکاران[1] ۲۰۰۲). به بیان ساده‌تر، ارزیابی توانمندی به عنوان پیامد نهایی دستیابی به تعداد زیادی رفتار و عملکرد پیچیده در نظر گرفته می‌شود (امین و کو ۲۰۰۳). به همین دلیل ارزیابی آن باید شامل فعال کردن دانسته‌های قبلی، کسب و تحلیل داده‌ها، حل مسائل بالینی، اتخاذ تصمیمات مناسب برای مدیریت بیمار و در نهایت انجام یک وظیفه مشخص باشد. با توجه به تعریف فوق مسلم است که در عمل، ارزیابی توانمندی از طریق یک آزمون میسر نخواهد بود بلکه مجموعه‌ای از آزمون‌ها برای سنجش این پارامترها در یک زمان مورد نیاز است (امین و کو ۲۰۰۳).

از دیدگاه کراچیو و همکاران عناصر ارزیابی مبتنی بر توانمندی شامل موارد زیر است:

- **ابزارهای ارزیابی مبتنی بر عملکرد:** در ارزیابی مبتنی بر توانمندی، ابزارهای ارزیابی باید معتبر و مبتنی بر عملکرد باشند تا بتوانند تصویری مشابه با عملکرد واقعی آزمودنی‌ها فراهم سازند.
- **اندازه‌گیری‌های متعدد:** ارزیابی توانمندی باید آمیخته‌ای از تعداد زیادی ویژگی‌ها و توانایی‌ها مرتبط را در برگیرد. در این میان قاعدتاً استفاده از آزمون‌های متعدد، ابزارهای معتبر و پایا برای سنجش دقیق توانمندی ضرورت دارد.
- **مشاهده مستقیم:** از آنجا که ارزیابی مبتنی بر توانمندی هر سه جزء دانش، نگرش و عملکرد را در برمی‌گیرد، از طریق مشاهده مستقیم عملکرد فراگیران در محیط‌های واقعی است که می‌توان درک کاملی از آن پیدا کرد.
- **ارزیابی معیار محور:** واضح است که ارزیابی معیارمحور، برای تفسیر داده‌های حاصل از ارزیابی توانمندی مقدم برشمرده می‌شود. زیرا بدین ترتیب می‌توان اطمینان حاصل کرد که آزمودنی‌ها توانسته‌اند حداقل توانایی‌های مورد انتظار را کسب کنند. در حالی که در ارزیابی هنجار محور عملکرد آزمودنی‌ها با یکدیگر مقایسه می‌شود، فارغ از آن که بتوان استنباط کرد آن‌ها در چه سطحی از توانمندی مورد انتظار قرار دارند. به عبارت دیگر، در ارزیابی هنجار محور این شانس برای تعدادی از آزمودنی‌های غیرتوانمند وجود دارد که عملکردشان قابل قبول در نظر گرفته شود (در صورتی که سایر آزمودنی‌ها در سطح بسیار ضعیف قرار داشته باشند). بر عکس این موضوع نیز صادق است. ممکن است عملکرد یک آزمودنی نامطلوب در نظر گرفته شود، حتی اگر در سطح قابل قبولی از توانمندی مورد انتظار قرار داشته باشد (در شرایطی که سایر آزمودنی‌ها در سطح بسیار مطلوبی قرار داشته باشند). بنابراین توصیه می‌شود که ارزیابی توانمندی از نوع معیار محور باشد.
- **ارزیابی تکوینی:** در ارزیابی مبتنی بر توانمندی، سیستم ارزیابی باید به منظور بهبود توانمندی فراگیر، شرایط لازم را برای انجام ارزیابی تکوینی و ارائه بازخورد به قدر کافی و به صورت مناسب فراهم سازد.

1. Carraccio et al.

توصیه‌های شورای پزشکی عمومی انگلیس در ارتباط با ارزیابی توانمندی‌های فراگیران (برگرفته از شان‌وی و هاردن ۲۰۰۳)

۱) اصول ارزیابی
- برنامه ارزیابی باید به نحوی تدوین شود که در عین آنکه از تمامی اجزاء برنامه درسی حمایت می کند، بتواند این فرصت را در اختیار فراگیران قرار دهد که میزان دستیابی به پیامدهای برنامه درسی را به نمایش بگذارند. به عبارت دیگر، فراگیران باید قادر باشند از طریق برنامه ارزیابی، وسعت و عمق دانسته‌های خود را ارائه کنند و آنچه که می توانند انجام دهند، به نمایش بگذارند. مسلماً در این میان علاوه بر ارزیابی دانش و مهارت، نگرش‌ها و رفتارهای حرفه‌ای نیز باید سنجیده شود.
- دانشکده‌های پزشکی به منظور سنجش پیامدهای مورد انتظار برنامه‌های درسی باید از طیف وسیع ابزارهای ارزیابی مناسب استفاده کنند. دانشکده‌های پزشکی باید مناسب‌ترین برنامه ارزیابی را مطابق با برنامه‌های درسی تعیین کنند. برنامه تدوین شده باید در عمل قابلیت اجرا داشته باشد. به علاوه دانشکده‌های پزشکی باید قادر باشند شواهدی در خصوص روایی و پایایی ارزیابی‌ها و فرایند تعیین حدنصاب قبولی و نحوه تصمیم گیری در خصوص عملکرد فراگیران ارائه نمایند.
- دانش، نگرش، مهارت و رفتار فراگیران باید در زمان فارغ‌التحصیلی بررسی و ارزیابی شود تا از کسب توانمندی‌های لازم برای ورود به عرصه ارائه خدمات اطمینان حاصل شود.

۲) روش های ارزیابی
برنامه ارزیابی باید منصفانه و متناسب با استانداردهای موجود باشد. دانشکده‌های پزشکی باید اطمینان حاصل کنند که:
- شواهدی از چگونگی تطبیق طرح ارزیابی با پیامدهای مورد انتظار برنامه درسی وجود داشته باشد.
- شواهدی از مشارکت ارزیابی‌ها و امتحانات فردی در ارزیابی کلی از پیامدهای مورد انتظار برنامه درسی وجود داشته باشد.
- در زمان طراحی و تدوین ارزیابی‌ها، شواهدی از چگونگی دستیابی به پیامدهای مورد انتظار برنامه درسی وجود داشته باشد.
- فراگیران راهنمایی‌های واضح و مشخصی را در مورد آنچه از آنها در هر امتحان انتظار می‌رود کسب کرده باشند.
- ارزیابان برای انجام نقش‌های و مسؤولیت‌هایی که در جریان ارزیابی برعهده دارند آموزش‌های لازم را دریافت کرده باشند.
- ارزیابان دستورالعمل‌های واضح و مشخصی در مورد شیوه نمره دهی امتحانات داشته باشند.
- سیستم‌های ارزیابی، برای تعیین نمره قبولی وجود داشته باشد.
- از ارزیابان خارجی به منظور اطمینان از تحقق استانداردهای مورد انتظار دعوت به عمل آمده باشد.

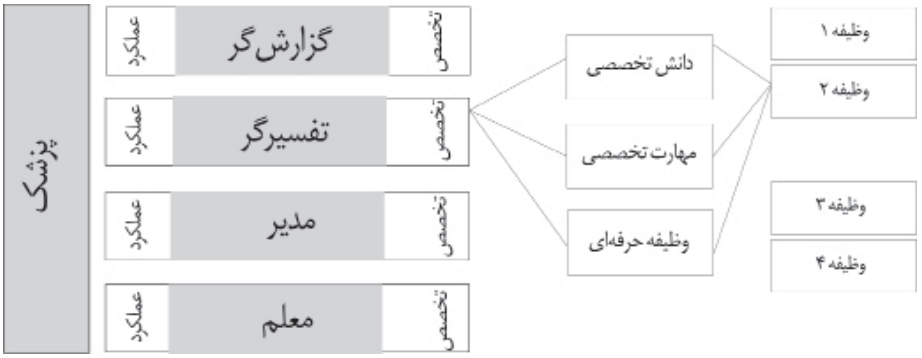
تقسیم‌بندی چارچوب‌های ارزیابی مبتنی بر توانمندی

پنگارو و تن‌کیت[1] (۲۰۱۳) به طور کلی سه دسته کلی برای تقسیم‌بندی چارچوب‌های ارزیابی ارائه می‌دهند که در ادامه به شرح مختصر هر یک از آنها خواهیم پرداخت.

- **چارچوب‌های تحلیلی[2]:** در این رویکرد تمرکز اصلی بر ارزیابی اجزاء تشکیل‌دهنده توانمندی‌های موردانتظار (به عنوان مثال، دانش، مهارت و نگرش) به منظور تسهیل فرایند ارزیابی است. توانمندی‌های اصلی معمولاً در قالب نقش‌ها و یا حیطه‌هایی بیان می‌شوند که خود مشتمل بر زیرتوانمندی‌هایی است که همراه با جزئیات توصیف شده‌اند (شکل ۲-۲). به طور کلی می‌توان گفت، اکثر چارچوب‌های پیامد محور ماهیت تحلیلی دارند که معمولاً به صورت مجموعه صلاحیت‌های موردانتظار توصیف می‌شوند. از این دسته می‌توان ACGME و CanMEDs را نام برد. به عنوان مثال، در مدل CanMEDs، یک پزشک باید متخصص، ارتباط‌دهنده، هماهنگ‌کننده، مدیر، پژوهشگر، حرفه‌ای و مدافع سلامت باشد (فرانک[3] ۲۰۰۵).
- **چارچوب‌های ترکیبی[4]:** چارچوب‌های ترکیبی برخلاف انواع تحلیلی بیشتر از نوع ادغام‌یافته و کمتر سنجش محور هستند (پنگارو ۱۹۹۹). یکی از انواع این چارچوب‌ها مدل RIME است. چارچوب‌های ترکیبی در پی پاسخ به این سؤال اساسی هستند که چه وظایف یا فعالیت‌هایی می‌توانند نشان‌دهنده کسب توانمندی مورد انتظار در فراگیران باشند. در این رویکرد، عناصر هر حیطه ممکن است با سایر حیطه‌ها در قالب یک وظیفه مشخص تعریف شوند. به عنوان مثال دانش مربوط به حیطه شناختی با مهارت‌های ارتباطی، مدیریتی و همکاری ادغام و ارائه می‌گردد (شکل ۳-۲).

1. Pangaro & Ten Cate
2. Analytic frameworks
3. Frank
4. Synthetic frameworks

شکل ۲-۲: شیوه ارائه توانمندی‌ها در مدل CanMEDs به عنوان یک چارچوب تحلیلی



شکل ۳-۲: ادغام توانمندی‌ها در قالب وظایف مشخص حرفه‌ای در چارچوب ترکیبی

- **چارچوب‌های توسعه‌ای[1]:** در چارچوب‌های توسعه‌ای فرایند پیشرفت فراگیران از مبتدی به خبرگی به تصویر کشیده می‌شود. برخلاف چارچوب‌های تحلیلی مبتنی بر توانمندی که ایستا هستند، در این رویکرد مراحل یا گام‌های رشد و پیشرفت فراگیران در هر فاز تعریف می‌شود. از انواع این چارچوب‌ها می‌توان مدل دریفوس و دریفوس[2] را نام برد. توجه به این نکته ضروری است که اگرچه هر کدام از مدل‌های موجود را می‌توان در یکی از دسته‌های فوق نامگذاری نمود اما در عمل اکثر آنان مخلوطی[3] از ویژگی‌های رویکردهای فوق را دارا هستند. در جدول ۱۵-۲ ویژگی‌ها، مزایا و محدودیت‌های هر یک از چارچوب‌ها نشان داده شده است.

1. Developmental frameworks
2. Dreyfus & Dreyfus
3. Hybrid

جدول ۱۵-۲: خلاصه چارچوب‌های ارزیابی توانمندی: مفهوم و مثال‌ها، مزایا و محدودیت‌ها (برگرفته از پنگارو و تن‌کیت ۲۰۱۳)

مفهوم و مثال‌ها	تقسیم توانمندی‌ها به حیطه‌ها (دانش، مهارت و نگرش) ACGME, CanMEDs	ترکیب حیطه‌ها به وظایف مشخص (گزارش‌کننده، تفسیرکننده، مدیر، معلم) RIME	توصیف فرایند پیشرفت فراگیران (مبتدی، کارآموز، متخصص، ماهر) دریفوس و دریفوس
مزایا	پوشش تمامی حیطه‌ها، امکان ارزیابی دقیق و ارائه بازخورد در هر حیطه به طور مشخص	منطبق با محیط واقعی کار، قابل کاربرد در واقعیت	امکان ارزیابی میزان پیشرفت فراگیران به صورت فردی و در طول دوره چند ساله آموزش
محدودیت‌ها	امکان کلی‌گویی، پیچیدگی درک آن، ارتباط ضعیف با فعالیت‌های بالینی	امکان ناتوانی در شرح دلایل مشخص عدم پیشرفت فراگیران به دلیل ارزیابی جامع عملکرد آنها	تعاریف متفاوت حیطه‌های مختلف در سطوح مختلف خبرگی، ارزیابی مبتنی بر هنجار پیشرفت تحصیلی فراگیران در عین استفاده از استانداردهای ثابت

مراحل تدوین برنامه ارزیابی مبتنی بر توانمندی

اولین مرحله تدوین برنامه ارزیابی مبتنی بر توانمندی تعیین توانمندی‌هایی است که باید ارزیابی شوند. در مرحله دوم سطح ارزیابی درخواستی باید تعیین شود و در نهایت، ارزیابی میزان پیشرفت فراگیران بر اساس ویژگی‌های آنان باید مشخص گردد.

- **تدوین چارچوب توانمندی‌ها:** فهرست صلاحیت‌ها به عنوان اولین گام در شناسایی اهداف یادگیری برنامه‌های آموزشی در نظر گرفته می‌شود و انتظار می‌رود که تدوین محتوا، آموزش و تمرین تخصص و ریزتخصص‌های مورد نظر بر اساس آن‌ها صورت گیرد. اطلاعاتی که از طریق ارزیابی این توانمندی‌های حاصل می‌شود، پایه‌ای برای قضاوت در خصوص کیفیت عملکردهای فراگیران و آموزش آنها فراهم می‌سازد. به‌علاوه به توسعه مستمر حرفه‌ای فراگیران و برنامه‌های آموزشی نیز کمک خواهد نمود. در صفحات پیشین مثال‌هایی از چندین چارچوب توانمندی ذکر شد.
- **تعیین سطح ارزیابی:** همان‌طور که پیش‌تر گفته شد، ماهیت چندبعدی توانمندی مبین این حقیقت است که هیچ روش منحصر به فرد ارزیابی نمی‌تواند اطلاعات جامعی برای قضاوت در خصوص عملکرد دانشجویان و دستیاران فراهم سازد. بر اساس اینکه یک توانمندی قرار است از منظر دانش زیربنایی لازم سنجیده شود یا قرار است داشتن توانایی انجام کار ارزیابی شود یا قرار است در محیط واقعی کار اتفاق بیفتد، از ابزارهای متفاوتی می‌توان بهره گرفت.
- **ارزیابی میزان پیشرفت:** واضح است که کسب توانمندی فرایندی نیست که یک‌شبه حاصل شود بلکه از ابتدای ورود به دانشگاه آغاز می‌گردد و در سرتاسر زندگی حرفه‌ای یک پزشک ادامه می‌یابد. در این صورت می‌توان انتظار داشت که عملکرد فراگیران با گذشت از مراحل مختلف آموزشی، به تدریج در حیطه‌های مختلف بهبود یابد. بنابراین استادان باید قادر باشند زمان‌هایی را که فراگیران از دانش، مهارت و نگرش کافی برخوردار هستند برای ورودشان به مقاطع بعدی شناسایی نمایند. رسیدن به این موضوع مستلزم تعریف استانداردهای مناسب برای هر فاز مشخص است. به عنوان مثال برای دانشجوی پزشکی می‌توان در پایان هر یک از فازهای علوم پایه، فیزیوپاتولوژی، کارآموزی و کارورزی به ارزیابی توانمندی‌های مورد انتظار پرداخت. قطعاً تصمیم‌گیری در خصوص حداقل توانمندی‌های مورد انتظار در هر فاز به ویژگی‌های فراگیران، طول دوره تحصیلی و اهمیت کسب آن بستگی دارد.

با توجه به موارد فوق، به صورت کلی می‌توان گفت که متناظر با هر یک از توانمندی‌های مورد بحث، روش‌های ارزیابی مناسبی وجود دارند که توصیه می‌شود مورد استفاده قرار گیرند. نمونه‌ای از روش‌های ارزیابی در جدول ۱۶-۲ آمده است.

جدول ۱۶-۲: روش‌های توصیه شده برای ارزیابی پیامدهای یادگیری در یک پزشک توانمند (برگرفته از شان‌وی و هاردن ۲۰۰۳)

حیطه	توانمندی	روش ارزیابی
آنچه پزشک قادر به انجام آن است	مهارت‌های بالینی مهارت‌های عملی بررسی بیمار مدیریت بیمار ارتقاء سلامت و پیشگیری برقراری ارتباط مدیریت اطلاعات	OSCE، مشاهده، لاگ‌بوک، آزمون کتبی OSCE، مشاهده، کارپوشه، لاگ‌بوک، کتبی کتبی، OSCE، مشاهده، کارپوشه کتبی، OSCE، مشاهده، کارپوشه OSCE، کارپوشه، مشاهده، کتبی OSCE، مشاهده، ارزیابی همتایان، خودارزیابی، کارپوشه کارپوشه، OSCE، مشاهده، کتبی
چگونگی رویکرد پزشک به طبابت	اصول علوم اجتماعی، پایه و بالینی نگرش، اخلاق و مسؤولیت قانونی تصمیم‌گیری، قضاوت و استدلال بالینی	کتبی، کارپوشه، مشاهده، OSCE مشاهده، کارپوشه، ارزیابی همتایان، خودارزیابی، OSCE، کتبی کارپوشه، مشاهده، کتبی، OSCE، ارزیابی همتایان، خودارزیابی
پزشک به عنوان یک فرد حرفه‌ای	ایفای نقش حرفه‌ای توسعه و رشد فردی	مشاهده، ارزیابی همتایان، خودارزیابی، OSCE، کتبی کارپوشه، مشاهده، ارزیابی همتایان، خودارزیابی، OSCE، کتبی

چالش‌های ارزیابیِ مبتنی بر توانمندی

با وجود استقبال نسبتاً خوب از ارزیابی مبتنی بر توانمندی، چالش‌هایی نیز برای آن برشمرده شده است:

- **لزوم داشتن دید سیســـتماتیک در طراحی برنامه ارزیابی به منظور پوشش محورهای مختلف توانمندی:** پیچیده بودن مفهوم توانمندی، دانشکده‌های پزشکی را به این سمت سوق داده است تا از تکنیک‌های مختلف ارزیابی برای سنجش آن استفاده کنند که قاعدتاً فراهم‌سازی بستر لازم برای اجرای آن با دشواری‌هایی مواجه است. هرچند مدت طولانی تحصیل، امکان ارزیابی عملکرد دانشجویان علوم پزشکی به خصوص دستیاران را فراهم می‌کند اما به نوبه خود بر پیچیدگی ارزیابی در فازهای مختلف تحصیل می‌افزاید.
- **خُرد کردن بیش از اندازه توانمندی به ریزمهارت‌ها:** برخی معتقد هســـتند توانمندی، توانایی اداره کردن حرفه در شـــرایط پیچیده‌ای است که از طریق تلفیق دانش، مهارت‌های عملکردی و نگرش فراگیر حاصل می‌شود. هرچند کوریکولوم‌ها بر پایه چنین توانمندی‌ها و اهدافی تدوین شده‌اند و در حیطه یاددهی-یادگیری، کم و بیش رویکردهای تلفیقی به کار می‌روند، به نظر می‌رسد در حوزه ارزیابی فراگیر تغییر مشخصی از این حیث ایجاد نشده باشد و همچون قبل، توانمندی‌ها به واحدهای کوچک و جدا از هم شکســـته می‌شـــوند. در واقع فرض بر این بوده است که قبولی در آزمون‌های جداگانه، به معنای رسیدن به حد تسلط در انجام وظایف پیچیده حرفه‌ای است که تا حد زیادی ساده‌انگارانه و تقلیل‌گرایانه می‌باشد. از طرف دیگر، تصور اینکه ارزیابی توانمندی به صورت کلی و ادغام‌یافته راه‌حل آسانی دارد، نیز درست نیست. مخصوصاً اگر قرار باشد مواردی مانند کار تیمی، مهارت‌های فراشناختی[1]، رفتار حرفه‌ای[2]، بازخورد دادن، نقد خود[3] و ... ارزیابی شـــوند. زیرا برای سنجش این موارد بر استفاده از رویکردهایی مانند جمع‌آوری داده‌های کیفی، توصیفی و روایی[4] تاکید می‌شود که با رویکردهای پیشین که بیشتر کمی بودند، متفاوت است.

1. Meta-cognitive skills
2. Professionalism
3. Self-appraisal
4. Narrative

- **تعیین خصوصیات روش‌های مختلـــف ارزیابی فراگیران از جمله روایی و پایایی:** دشـــواری‌های مربوط به تعیین روایی و پایایی ابزارهای ارزیابی در کنار مشخص کردن سایر ویژگی‌ها از قبیل پذیرش یا تاثیر آموزشی آن بر فراگیران، موسســـات آموزشی را با چالش‌هایی از حیث انتخاب ابزارهای کارآمد برای سنجش توانمندی‌های مختلف روبه‌رو ساخته است.

بحث بیشتر در خصوص سیستم یا برنامه ارزیابی مبتنی بر توانمندی در فصل نهم کتاب ارائه خواهد شد.

سودمندی

ابزارهای مختلف ویژگی‌ها و کاربردهای متفاوتی دارند که در موقعیت‌های مختلف و حسب شرایط می‌توان از آنها بهره گرفت. برای انتخاب ابزار ارزیابی مناسب باید به تعدادی از عوامل توجه کرد که در این فصل به آنها اشاره می‌شود. در واقع، در سال ۱۹۹۶ ون‌درولوتن و همکاران مدلی برای تعریف سودمندی[1] ارائه دادند که بر اساس آن می‌توان در مورد ابزار ارزیابی قضاوت کرد. در این مدل علاوه بر پایایی[2] و روایی[3] که از قبل به کرات در متون مورد توجه قرار می‌گرفتند، به تاثیر آموزشی[4]، هزینه[5] و مقبولیت[6] نیز اهمیت داده شد. فرمول سودمندی ابزار به این شکل است:

$$u=R_w x V_w x E_w x A_w x C_w$$

چند نکته در خصوص استفاده از این مدل باید مدنظر قرار گیرند:

- این فرمول از ۵ فاکتور تشکیل شده است اما حرف w در واقع نشان‌دهنده این موضوع است که وزن فاکتورها با یکدیگر برابر نیست. به عبارت دیگر، در عین حال که توجه همزمان به هر ۵ فاکتور باید مدنظر قرار گیرد، میزان توجه به هر کدام به موقعیت مکانی و زمانی استفاده‌کننده از ابزار بستگی دارد. به عنوان مثال، اهمیت پایایی در یک امتحان high stake مانند اخذ مدرک نسبت به یک امتحان درون‌بخشی که برای ارزیابی تکوینی و دادن بازخورد استفاده می‌شود، بیشتر است.
- علامت ضرب بین فاکتورها به معنای ضرب ریاضی نیست. به عبارت دیگر منظور این نیست که برای تعیین سودمندی یک ابزار باید مقادیر پایایی و روایی و ... را به صورت کمی محاسبه کنیم و حاصل‌ضرب آنها را در فرمول تعیین کنیم. این مسأله صرفاً نشان می‌دهد که اگر یکی از فاکتورها وجود نداشته باشد (ارزش آن صفر باشد)، امکان استفاده از ابزار به صورت کلی وجود ندارد. به عنوان مثال، اگر قرار است از آزمون OSCE استفاده شود اما بودجه مالی مربوطه به هیچ عنوان قابل تأمین نیست، باید کلاً از اجرای آن منصرف شد.
- مسأله دیگری که با ارائه این فرمول مطرح می‌شود، نگاه کردن به مقوله ارزیابی فراگیر به عنوان مفهومی فراتر از اندازه‌گیری و ارزیابی صرف است. در این رویکرد، با توجه ویژه به جنبه‌های آموزشی، اقتصادی و اجرایی، ابزار به عنوان عاملی تاثیرگذار در دستیابی به حداکثر سود در برنامه‌ریزی آموزشی مطرح می‌شود. به عنوان مثال، حتی در صورت وجود بهترین روایی و پایایی ابزار اندازه‌گیری، در برخی مواقع مجبور هستیم که به دلیل محدودیت‌های اجرایی از انتخاب آن صرف نظر کنیم. در این شرایط جنبه‌های اقتصادی و اجرایی نقش بسیار کلیدی در انتخاب ابزار دارند.

1. Utility
2. Reliability
3. Validity
4. Educational impact
5. Cost
6. Acceptability

- نکته دیگر در ارتباط با این فرمول، نگاه ویژه به نقشی است که ارزیابان و به ویژه دانشجویان در قبول و قابلیت پیاده‌سازی آن بازی می‌کنند. بی‌شک عدم توجه به این عامل در بسیاری از مواقع منجر به شکست بسیاری از ابزارهای آموزشی نوین در عمل می‌شود.
- در آخر آنکه کاربرد این فرمول تنها به انتخاب بهترین ابزارهای ارزیابی ختم نمی‌شود و در سطح کلان‌تر در طراحی و تحلیل «سیستم‌های ارزیابی مبتنی بر توانمندی» بسیار کارساز است. هرچند به این منظور مدل‌های پیچیده‌تری وجود دارد که در بخش نهم کتاب به آنها خواهیم پرداخت.

در این قسمت سعی شد که در مورد مفهوم سودمندی ابزارهای ارزیابی به صورت کلی مرور اجمالی صورت گیرد. در هر یک از بخش‌های بعد با معرفی تفصیلی تک‌تک ابزارها، سودمندی هر یک از آنها با جزئیات بیشتر و بر پایه مطالعات موجود تحلیل و بررسی خواهد شد.

در ادامه این فصل، به شرح پنج جزء فرمول ون‌درولوتن شامل پایایی، روایی، تاثیر آموزشی، هزینه و مقبولیت خواهیم پرداخت.

پایایی

مفهوم پایایی

شاخصی که تعیین می‌کند یک ابزار در شرایط مختلف تا چه حد نتایج یکسانی به دست می‌دهد، پایایی ابزار است. به بیان دیگر، پایایی در یک آزمون به این معنا است که نتیجه آزمون برای یک دانشجو تا چه اندازه با تکرار اندازه‌گیری‌ها در شرایط متفاوت یکسان باقی می‌ماند و ثبات دارد. همچنین پایایی آزمون بیان می‌کند که نتیجه‌ای که دانشجو از یک امتحان گرفته است، تا چه حد قابل تعمیم به سایر شرایط است.

در واقع در بحث سنجش ویژگی‌های روانشناختی و آموزشی، فاکتورهای متعددی در نمره آزمون‌شوندگان دخیل است که لزوماً منعکس‌کننده سطح واقعی آنها نیست و می‌تواند به عنوان منابع خطای اندازه‌گیری بر نتایج حاصل از آزمون تاثیرگذار باشد. به عنوان مثال، انتظار می‌رود نمره دانشجو در یک آزمون شفاهی در دفعات برگزاری با چند استاد یکسان به دست آید. در غیر این صورت، تغییر مشاهده شده در نمره به خطای آزمون منتسب می‌شود که می‌تواند از منابع مختلف مانند توجیه نبودن استادان در خصوص محتوای آزمون، سختگیر بودن یک استاد نسبت به سایرین، تعداد متفاوت سؤالات و ... ناشی شود.

البته هنگامی که از ثبات نتیجه آزمون صحبت می‌کنیم، صرفاً نمره خام دانشجو منظور نیست. نتیجه آزمون در یک آزمون معیارمحور می‌تواند تصمیم رد یا قبول باشد. بنابراین اگر دانشجویی در آزمون نوبت اول رد شود و در آزمون نوبت دوم قبول شود، می‌توان گفت که نتیجه آزمون پایا نبوده است. همچنین در یک آزمون هنجارمحور که نمره خام فراگیران چندان اهمیت ندارد و دانشجویان نسبت به یکدیگر مورد سنجش قرار می‌گیرند، پایایی مشخص می‌کند که تا چه اندازه رتبه‌بندی دانشجویان در اندازه‌گیری‌های مکرر یکسان باقی می‌ماند. نمونه‌هایی از انواع مختلف ثبات نتایج آزمون می‌تواند شامل موارد زیر باشد:

- ثبات نتایج رد و قبول آزمون در طول زمان
- ثبات نمره دانشجو بین ارزیابان مختلف
- ثبات رتبه دانشجو در آزمون در محیط‌های مختلف (درمانگاه، اورژانس، بخش بستری و ...)

روش‌های مختلفی برای اندازه‌گیری پایایی آن وجود دارد که در بخش هشتم کتاب به آنها اشاره خواهد شد. در اینجا سعی می‌کنیم با ذکر چند نکته و مثال مفهوم پایایی را روشن‌تر کنیم.

- اولین نکته اینکه پایایی، ویژگی ذاتی یک ابزار نیست. به عنوان مثال، ترازویی را در نظر بگیرید که با دقت ۱± کیلوگرم وزن را نشان می‌دهد. شاید بتوانیم بگوییم که این ترازو برای اندازه‌گیری وزن بزرگسالان پایایی خوبی دارد.

اما بدیهی است که ترازوی مخصوص نوزادان باید دقتی بیشتر از این میزان داشته باشد. چون اگر بر فرض، یک بار وزن یک کودک را سه کیلوگرم نشان دهد و بار دیگر چهار کیلوگرم، خطای زیادی است و اندازه‌گیری را فاقد ارزش می‌سازد. بنابراین، نمی‌توانیم بگوییم پایایی این ترازو همیشه و در همه شرایط خوب است. همین مسأله در خصوص ارزیابی فراگیر وجود دارد. فرض کنید یک بار امتحان آسیب‌شناسی را از دانشجویان سال سوم پزشکی می‌گیریم و بار دیگر، کل دانشجویان پزشکی در آزمون شرکت می‌کنند. به خاطر مسأله‌ای که در سطور بعدی به شرح آن می‌پردازیم، در حالت دوم پایایی می‌تواند بیشتر از حالت اول باشد. بنابراین، پایایی، ویژگی ذاتی یک ابزار نیست و به تعامل ابزار با جامعه‌ای که برای آن استفاده می‌شود، بستگی دارد.

- دومین مسأله اینکه پایایی به صورت نسبت واریانس نمرات حقیقی[1] به واریانس نمرات مشاهده‌شده[2] تعریف می‌شود و با مقدار خطا[3] رابطه عکس دارد. به عنوان مثال، وزن سه نفر را با ترازوی فوق سنجیده‌ایم و مقادیر ۴۵، ۷۲ و ۶۶ کیلوگرم به دست آمده‌اند در حالی که نمی‌دانیم که وزن واقعی هر یک از این افراد چیست. احتمالاً وزن آنها واقعاً با هم متفاوت بوده و اختلاف اعداد ناشی از تفاوت واقعی بین‌فردی است. اما قسمتی از تفاوت بین نمرات، تصادفی و ناشی از خطای اندازه‌گیری است. همین مسأله در ارزیابی دانشجویان صدق می‌کند. هنگامی که از یک آزمونگر می‌خواهیم که ده دانشجو را در حین گرفتن شرح حال از بیمار مشاهده کند و به مهارت ارتباطی آنها نمره دهد، آزمونگر به دانشجویان نمرات متفاوتی می‌دهند. قبول داریم که قسمتی از تفاوت‌ها، یک تفاوت واقعی و غیرقابل حذف است و ذاتاً بین دانشجویان وجود دارد اما قسمتی از تفاوت بین نمرات، تصادفی و ناشی از خطای اندازه‌گیری است و ما نمی‌دانیم که دقیقاً به چه چیزی مربوط می‌شود. پس نمره واقعی دانشجو را نمی‌دانیم. نمره واقعی هر دانشجو مجموع نمره مشاهده شده و خطای تصادفی است. اگر در شرایط فرضی خطای اندازه‌گیری صفر باشد، نمره مشاهده شده برابر نمره واقعی می‌شود. ضریب پایایی نسبت این دو مقدار را مشخص می‌کند و به صورت عددی از صفر (بدون پایایی) تا یک (پایایی ایده‌آل) بیان می‌شود:

$$\text{پایانی} = \frac{\text{واریانس نمرات حقیقی}}{\text{واریانس نمرات مشاهده شده}}$$

- مسأله سوم ارتباط مفهوم پایایی با پدیده تکرار اندازه‌گیری است. از آنجا که نمره واقعی هرگز بر هیچ‌کس معلوم نیست، تنها راه برای نزدیک شدن به مقدار آن تکرار اندازه‌گیری است. یعنی می‌توان امیدوار بود که با انجام ارزیابی به دفعات و محاسبه میانگین نمرات حاصله، خطای اندازه‌گیری را کاهش دهیم و به مقدار نمره حقیقی نزدیک شویم. از همین روست که تعریف رایج پایایی به تکرار اندازه‌گیری دلالت دارد.
- مسأله چهارم منابع مختلف بروز خطا در یک اندازه‌گیری است. در مثال ارزیابی مهارت ارتباطی فرض کنید که هر دانشجو توسط سه آزمونگر مورد سنجش قرار می‌گیرند و نمره نهایی هرکس از میانگین سه ارزیابی به دست می‌آید. در اینجا تفاوت مشاهده‌شده بین نمرات، فراتر از تفاوت بین‌فردی و خطای تصادفی است و ممکن است ناشی از اختلاف‌نظر بین آزمونگران باشد. پس سه منبع برای تفاوت نمرات این آزمون می‌توان در نظر گرفت. اگر اختلاف کل، عمدتاً ناشی از تفاوت بین‌فردی واقعی باشد، و سهم دو مورد دیگر جزئی باشد، می‌توان گفت که ابزار پایایی خوبی دارد. زیرا همان‌طور که ذکر شد، پایایی نسبت تفاوت بین‌فردی به تفاوت کل را نشان می‌دهد. یکی از مشکلاتی که در مورد آزمون‌های ذهنی و مبتنی بر عملکرد مانند کارپوشه وجود دارد، این است که منابع خطا در آنها زیاد است که این امر باعث کاهش پایایی می‌شود. به عبارت دیگر، متغیرهای تأثیرگذار در ارزیابی بسیارند و

1. True score
2. Observed score
3. Error

معمولاً غیرقابل‌شناسایی و غیرقابل‌کنترل هستند. به عنوان مثال در مورد کارپوشه، توانمندی‌هایی که مورد ارزیابی قرار می‌گیرند متنوع هستند و استادان هم ممکن است نظرات و سلایق متفاوتی داشته باشند.

- مسأله پنجم اینکه بنا به بحث قبلی، هر چه تفاوت بین فردی فراگیران بیشتر باشد (یعنی مقدار خطای تصادفی و خطای آزمونگران کمتر)، پایایی افزایش خواهد داشت. اکنون این مفهوم قابل درک می‌شود که چرا سیف (۱۳۹۴) در این‌باره بیان می‌کند که اگر یک آزمون را در مورد گروهی از افراد که از نظر توانایی مورد سنجش تجانس کمتری با یکدیگر دارند اجرا کنیم، آزمون پایایی بیشتری نشان خواهد داد. همین موضوع در مثال امتحان آسیب‌شناسی که پیش‌تر عنوان شد، صادق است.

در بخش‌های بعدی با معرفی هر یک از ابزارهای ارزیابی ضمن اشاره به مطالعات صورت گرفته، به بیان عوامل موثر با پایایی هر ابزار خواهیم پرداخت که در جدول ۱۷-۲ به صورت خلاصه به آنها اشاره شده است.

توصیه‌های کلی که برای بهبود پایایی در ارزیابی فراگیر وجود دارد، شامل موارد زیر است:

- تنوع ابزارهای ارزیابی
- برگزاری آزمون‌های مکرر
- افزایش تعداد سؤالات
- همگون کردن سؤالات با اهداف دوره
- طراحی سؤالات از تمام اهداف و موضوعات

افتراق پایایی و عینیت

نکته‌ای که در اینجا باید به آن توجه داشت افتراق مفهوم پایایی از عینیت[1] در ارزیابی است، اگرچه این دو موضوع از لحاظ مفهومی بسیار با یکدیگر قرابت دارند. عینیت در واقع به معنای مشخص بودن قواعد و معیارهای اندازه‌گیری به منظور کسب توافق بین صاحب‌نظران در انتخاب پاسخ درست است (امین و کو ۲۰۰۳). برخلاف دیدگاه مرین‌بور[2] که مفهوم عینیت را در مقابل ذهنیت[3] در نظر می‌گیرد (مرین بور ۲۰۱۵)، از نظر گلدشمیت[4] در واقعیت طیفی از این دو مفهوم وجود دارد. او معتقد است که محققان حوزه آموزش پزشکی از ترکیبی از داده‌های عینی و ذهنی بهره‌مند می‌شوند (گلدشمیت ۲۰۱۵).

با این همه، این قرابت مفهومی باعث شده بود که برای مدت‌ها، پایایی و عینی بودن آزمون معادل و هم‌ارز یکدیگر به کار روند و در خیلی از موارد عینی بودن آزمون، معیاری برای پایایی بالا محسوب شود اما تحقیقات اخیر این موضوع را

جدول ۱۷-۲: عوامل موثر بر پایایی آزمون

منبع تغییرپذیری	عوامل موثر بر پایایی آزمون
آزمودنی‌ها	میزان استعداد و توانایی ذهنی، انگیزش و علاقه، خستگی، ناهمگونی عملکرد
آزمون	طولانی بودن زمان آزمون، دشواری سؤالات، قدرت تمیز سؤالات، همگن بودن سؤالات، تعداد سؤالات
شرایط اجرای آزمون	دستورالعمل آزمون، شرایط فیزیکی محل از قبیل میزان نور، صدا و درجه حرارت و .... زمان آزمون
ارزیابان	تخصص و دانش مرتبط با محتوای ارزیابی، آشنایی با ابزار ارزیابی، آگاهی از معیارهای نمره‌دهی، تعداد ارزیابان

1. Objectivity
2. Merrienboer
3. Subjectivity
4. Goldszmidt

تأیید نمی‌کنند. به صورت مشخص در موردOSCE این نظر وجود داشت که مزیت عمده آن نسبت به سایر ابزارهای سنجش مهارت‌های بالینی، در عینی بودن است که تصور می‌شد پایایی را افزایش می‌دهد. در حالی‌که بعدها بر اساس نتایج تعداد زیادی مطالعه مشخص شد که پایایی آزمون OSCE به عوامل دیگری از جمله نمونه‌گیری خوب از محتوای بالینی مورد نظر و یا تعداد بیشتر ایستگاه‌ها بر می‌گردد. در واقع آنچه پایایی آزمون OSCE را بالا می‌برد، اولاً به خاطر تعداد بیشتر موضوعاتی است که مورد ارزیابی قرار می‌گیرند (نسبت به سایر روش‌های مرتبط) و ثانیاً ناشی از چرخش دانشجویان در ایستگاه‌های مختلف و مواجهه آنان با نمونه خوبی از بیماران و آزمونگران است. در واقع عینی بودن، ساختارمند بودن و چک لیست‌های استاندارد اثر کمتری بر پایایی آزمون OSCE دارند (ون‌درولوتن و سوانسون[1] ۱۹۹۱، سوانسون و همکاران ۱۹۹۵، نورمن و همکاران[2] ۲۰۰۶). البته این یافته‌ها محدود به آزمون OSCE نیستند. مطالعات نشان می‌دهند که ابزارهای ذهنی مانند امتحانات شفاهی، «مورد بالینی کامل» و Mini-CEX به شرطی که نمونه‌گیری مناسبی داشته باشند، پایایی مطلوبی به دست می‌دهند (واس و همکاران[3] ۲۰۰۱).

به صورت خلاصه، دلیلی ندارد که روش‌های ارزیابی ذهنی را به دلیل پایایی پایین کلاً کنار بگذاریم. روش‌هایی وجود دارد که می‌توان با به کار بستن آنها پایایی یک ابزار ارزیابی ذهنی را بهبود بخشید. مانند انتخاب آزمونگران متخصص، استفاده از چندین آزمونگر، توجیه و آموزش آزمونگران، انتخاب سؤالات متعدد، استفاده از بلوپرینت برای طرح سؤال و استفاده از ساختار و معیارهای نمره‌دهی مشخص. از طرف دیگر در صورت استفاده از روش‌هایی که به ساختارمندی و استاندارد بودن شهره هستند، نمی‌توانیم تصور کنیم که پایایی آزمون را تضمین کرده‌ایم. به عنوان مثال OSCE که با پنج ایستگاه برگزار می‌شود، هرچند که در هر ایستگاه آزمونگر با استفاده از یک چک‌لیست مشخص و عینی به سنجش عملکرد فراگیران می‌پردازد، ممکن است پایایی بالایی نداشته باشد.

روایی

مفهوم روایی

هرچند بعید به نظر می‌رسد اما این احتمال وجود دارد که آزمونی که هیچ چیز مفیدی را اندازه‌گیری نمی‌کند، از پایایی خوبی برخوردار باشد (بویل و فیشر[4] ۲۰۰۷). این موضوع به یکی دیگر از خصیصه‌های آزمون در موقعیت‌های اندازه‌گیری تحت عنوان روایی اشاره دارد. در واقع، بررسی پایایی ابزار اندازه‌گیری، تنها بخشی از فرایندی است که به منزله داشتن آزمون خوب باید صورت گیرد. بخش دیگر این فرایند بررسی روایی است.

روایی به این ویژگی می‌پردازد که آیا آزمون طراحی شده به اندازه‌گیری آنچه که ما قصد سنجش آن را داریم می‌پردازد یا در عمل چیز دیگری را می‌سنجد. بنابراین روایی اصطلاحی است که به هدفی که آزمون برای تحقق بخشیدن به آن تهیه شده است، اشاره می‌کند (امین و کو ۲۰۰۳). بر این اساس، آزمونی رواست که دقیقاً اهداف مورد نظر را ارزیابی کند و نه چیز دیگری را. به عنوان مثال، اگر هدف از برگزاری دوره آموزشی مهارت‌های ارتباطی، ارتقاء مهارت‌های مصاحبه فراگیران باشد، در این حالت ابزار طراحی شده برای ارزیابی باید به سنجش همین توانایی در آزمودنی‌ها بپردازد. آزمون کتبی برای سنجش چنین توانایی در داوطلبان قاعدتاً به عنوان یک آزمون با روایی پایین در نظر گرفته می‌شود به این دلیل که به ارزیابی دانش محتوایی آنان می‌پردازد نه مهارت‌های مصاحبه. در مقابل، یک نمونه آزمون با روایی بالا در این زمینه مشاهده مستقیم مصاحبه فراگیر در مواجهه با بیمار استانداردشده است. به طور خلاصه می‌توان گفت که روایی هر آزمون با توجه به محتوا و یا هدف مورد انتظار تعیین می‌شود (امین و کو ۲۰۰۳).

1. Swanson
2. Norman et al.
3. Wass et al.
4. Boyle & Fisher

روش‌های مختلفی برای اندازه‌گیری روایی آن وجود دارد که در بخش هشتم کتاب به آنها پرداخته خواهد شد. در اینجا یک نکته کلی در خصوص روایی آزمون ذکر می‌کنیم. همانند پایایی، روایی به عنوان یک ویژگی ذاتی ابزارهای ارزیابی تلقی نمی‌گردد (امین و کو ۲۰۰۳). این مسأله تا حدی تحت تاثیر هدف اجرای آزمون، نتایج حاصل از آن و تفسیر ارزیابان قرار دارد. در توضیح این مسأله توجه به دو نکته حائز اهمیت است:

- اولاً یک آزمون ممکن است برای یک موقعیت ارزیابی از روایی بالایی برخوردار باشد اما در سایر موقعیت‌ها روا محسوب نشود (امین و کو ۲۰۰۳، هاینز[1] ۲۰۰۴). در مثال بالا، در شرایطی که هدف آزمون صرفاً ارزیابی دانش شناختی داوطلبان از اصول مهارت‌های ارتباطی باشد، آزمون کتبی نیز یک ابزار ارزیابی روا به شمار می‌آید. اما اگر هدف آزمون سنجش مهارت‌های برقراری ارتباط فراگیران در برخورد با بیماران باشد، در این شرایط دیگر استفاده از آزمون‌های کتبی کارساز نیست.
- ثانیاً یک ابزار اندازه‌گیری ممکن است برای اندازه‌گیری یک خصیصه ویژه روا باشد در حالی که برای سنجش همان خصیصه بر روی جامعه دیگر از اعتبار خوبی برخوردار نباشد. به عنوان مثال، یک آزمون ممکن است برای سنجش استدلال بالینی کارآموزان پزشکی از روایی بالایی برخوردار باشد، در حالی که همان آزمون برای سنجش توانایی‌های استدلال دستیاران مناسب و روا تلقی نشود. زیرا اهداف یادگیری این دو گروه با یکدیگر متفاوت است.

بنابراین در نظر گرفتن این نکته اهمیت دارد که روایی یک ویژگی ثابت آزمون نیست که همیشه وجود داشته باشد یا نداشته باشد؛ بلکه آن را باید در زمینه و شرایط مورد استفاده خودش بررسی کرد. تعدادی از عوامل موثر بر روایی در جدول ۱۸-۲ آمده است. برای بهبود روایی ارزیابی فراگیر چندین توصیه به صورت زیر ارائه شده است:

- تنوع ابزارهای ارزیابی
- استفاده از بلوپرینت برای طرح سؤال
- استفاده از دستورالعمل‌های طراحی سؤال

طراحی سؤالات مناسب برای ارزیابی سطوح شناختی بالا

انواع روایی

به طور کلی به انواع متفاوتی از روایی در متون اشاره شده است (التون و جانستون[2] ۲۰۰۲). در یک تقسیم‌بندی کلی گفته می‌شود که روایی ارزیابی را می‌توان از دو منظر روایی درونی و بیرونی[3] مورد بررسی قرار داد:

روایی بیرونی با این امر سروکار دارد که نتایج به دست آمده از آزمون تا چه حد قابل تعمیم به گروه مشابه‌ای از

جدول ۱۸-۲: عوامل موثر بر روایی آزمون

منبع	منابع موثر بر روایی
آزمودنی‌ها	ترکیب آزمون شوندگان از لحاظ توانایی مورد سنجش، ویژگی‌های روانی آزمون شوندگان از قبیل سطح انگیزش یا علاقه
آزمون	نمونه‌گیری از محتوای مورد سنجش، کیفیت سؤال‌های آزمون، تعداد سؤالات آزمون، ترتیب قرار گرفتن سؤال‌های آزمون به دنبال یکدیگر، طراحی سؤالات مناسب برای ارزیابی سطح شناختی بالا
شرایط اجرای آزمون	تهیه بلوپرینت آزمون، دستورالعمل آزمون
ارزیابان	میزان تسلط بر محتوای مورد سنجش، آشنایی با اصول طراحی سؤالات

1. Haines
2. Elton & Johnston
3. Internal & External

فراگیران در نمونه‌ای بزرگ‌تر است. این روایی به جنبه پژوهشی ارزیابی اشاره می‌کند. به عنوان مثال، برای ارزیابی مهارت برقراری ارتباط کارآموزان پزشکی دانشکده، ۲۰ نفر از آنان را انتخاب می‌کنید و از ایشان آزمونی به صورت مشاهده مستقیم به عمل می‌آورید. چنانچه از نتایج حاصله بتوانید نتیجه‌گیری کنید که وضعیت کل کارآموزان پزشکی در دانشکده از نظر مهارت‌های ارتباطی چگونه است، در این شرایط می‌توان گفت که ارزیابی از روایی بیرونی برخوردار است. روایی بیرونی بیشتر تحت تاثیر انتخاب نمونه مناسب برای شرکت در ارزیابی است تا ابزار ارزیابی و ویژگی‌های آن. از آنجا که بحث ما روایی ابزار است، در این کتاب بر روایی درونی تمرکز می‌کنیم. بنابراین هرجا صحبت از روایی شد، نوع درونی مدنظر است. روایی درونی، که ذکر شده مهمترین نوع روایی در فرایند ارزیابی دوره‌های آموزشی است، به معنای آن است که آزمون تا چه حد پیامدها و اهداف در دوره آموزشی را ارزیابی می‌نماید (براون و همکاران[1] ۲۰۱۳). از چند راه می‌توانیم به درک این موضوع برسیم که همان انواع روایی را تشکیل می‌دهند و تحت نام روایی محتوایی[2]، روایی صوری[3]، روایی معیاری[4] و روایی سازه[5] شناخته شده‌اند. البته اقوال در این رابطه متفاوت است و شاید دقیق‌تر این باشد که بگوییم تقسیم‌بندی واقعی و مرزبندی بین انواع روایی وجود ندارد و همه آنها در واقع به جنبه‌های مختلف مفهوم روایی از زوایای متفاوت می‌پردازند. در ادامه به شرح جنبه‌های مختلف روایی خواهیم پرداخت.

۱) روایی محتوایی

روایی محتوایی بدین معنی است که یک ابزار ارزیابی، تا چه اندازه قلمرو محتوایی دوره آموزشی را در بر می‌گیرد. روایی محتوایی یکی از مشخصه‌های کلیدی یک ابزار ارزیابی است که نشان می‌دهد که آیا ابزار ارزیابی دقیقاً همان محتوایی را که قصد ارزیابی آن را داریم، می‌سنجد یا خیر. بر طبق این تعریف، آزمونی رواست که نمونه سؤال‌های آن، معرف کل سؤالاتی باشد که می‌توان از محتوا یا موضوع آموزشی مورد نظر طراحی نمود. نکته حائز اهمیت در این خصوص آن است که در یک آزمون امکان اینکه از تمام مباحث، سؤال پرسیده شود، وجود ندارد. به همین دلیل در یک آزمون، به اجبار تنها تعدادی محدودی از سؤالات انتخاب می‌شوند. آنچه در عمل انتظار می‌رود آن است که این تعداد سؤال، نمونه خوب و مناسبی از کل اهداف آموزشی باشد به طوری که نمره حاصل از آنها، نمایانگر وضعیت واقعی دانشجو باشد. یکی از روش‌های اطمینان از کسب روایی محتوا، استفاده از جدول مشخصات آزمون یا بلوپرینت است که در همین فصل مورد بحث قرار گرفت.

به طور کلی آنچه حائز اهمیت است، در نظر گرفتن این نکته است که روایی محتوا خصوصیتی است که همزمان با تدوین آزمون در آن تنیده می‌شود. به علاوه هر چند روایی ویژگی ذاتی ابزار ارزیابی نیست، باید در نظر داشت که در صورتی که ابزار ارزیابی قابلیت پوشش حجم بیشتری از محتوا را داشته باشد، قاعدتاً شانس آن برای کسب روایی محتوایی بالاتر فراهم خواهد بود. به عنوان مثال، از آنجا که امتحانات چند گزینه‌ای امکان نمونه‌گیری وسیع‌تری از اهداف را فراهم می‌سازند، به همین دلیل می‌توانند از روایی بالایی برخوردار باشند.

به منظور بهبود روایی محتوایی آزمون برداشتن سه گام ضروری است:

- تعیین موضوعات و مشکلاتی که داوطلب باید در مواجهه با آن‌ها توانمند شود. این مشکلات ممکن است از طریق نظرات تیم خبرگان و صاحب‌نظران یا بر اساس مطالعاتی که به مشاهده و بررسی آنچه که پزشکان در محیط‌های واقعی عملکرد با آن مواجه می‌شوند، استخراج شود.
- تعریف مجموعه وظایفی که در ارتباط با آن موضوع یا مشکل وجود دارد و داوطلب باید در آن توانمند باشد. به عنوان مثال، فرض کنید که مشکل درد قفسه سینه به عنوان یکی از تظاهرات شایع بالینی است که پزشکان عمومی در

1. Brown et al.
2. Content validity
3. Face validity
4. Criterion validity
5. Construct validity

محیط‌های کاری با آن مواجه هستند. در این شرایط مجموعه وظایف مورد انتظار شامل اخذ شرح حال از بیمار مبتلا به آنژین، تفسیر نوار قلب، داشتن توانایی در احیا قلبی- ریوی و آموزش به بیمار در مورد نحوه استفاده از داروهای قلبی یا رژیم غذایی باشد اما اداره بیماری که دچار عوارض دارویی شده است، در حیطه وظایف مورد انتظار از وی نیست.

- تدوین بلوپرینت آزمون که معمولاً به سادگی با توجه به دو بُعد فوق‌الذکر قابل اجرا است. همان‌طور که پیشتر نیز اشاره شد، روایی محتوایی، نمایشگر اهداف یادگیری در ارزیابی است و در عمل از طریق بلوپرینت قابل دستیابی است. بلوپرینت به فرایندی اشاره دارد که در آن محتوای آزمون به دقت و بر اساس اهداف یادگیری، مشخص می‌شوند و همچنین اهدافی که قرار است در آزمون مورد ارزیابی قرار گیرند و وزن نسبی آنها تعیین می‌گردند. به طور خلاصه می‌توان گفت، بلوپرینت یک راه آسان برای تعریف مجموعه سؤالاتی است که در آزمون ارائه می‌شوند.

۲) روایی صوری

روایی صوری که به روایی ظاهری و نمادی نیز مشهور است، یک شاخص مقدماتی و حداقلی از روایی محتوایی است. روایی صوری این مطلب را مد نظر دارد که سؤال‌های آزمون تا چه حد در ظاهر شبیه به موضوعی هستند که برای اندازه‌گیری آن تهیه شده‌اند. به طور خلاصه می‌توان گفت روایی صوری به معنای منطقی بودن، جالب بودن و تناسب ظاهری ابزار اندازه‌گیری است. در صورت عدم توجه به روایی صوری این خطر وجود دارد که آزمون از سوی آزمودنی‌ها جدی تلقی نشود (بویل و فیشر ۲۰۰۷).

به عنوان مثال، یک آزمون فیزیولوژی را در نظر بگیرید که برای دانشجویان پزشکی طراحی شده است اما سؤالاتی دارد که ظاهراً از مکانیسم‌های عملکردی بدن حیوانات پرسش می‌کند. هرچند ممکن است از نظر پایه علمی فیزیولوژیک، مکانیسم‌های موردنظر در بدن انسان و حیوان یکسان باشند، این آزمون برای دانشجویان پزشکی که در آینده با بدن انسان سر و کار دارند، فاقد روایی صوری است. در این صورت آزمون‌شونده ممکن است علاقه‌ای به جواب دادن سؤال‌های آزمون از خود نشان ندهد زیرا ممکن است چنین تصور کند که آزمون به شغل آینده او ارتباطی ندارد.

البته برخی معتقد هستند که وجود روایی صوری تنها در پاره‌ای از مواقع آن مفید است (سیف ۱۳۹۴) و روایی صوری مبنا و پایه درستی برای قضاوت در خصوص ارزش ابزارهای اندازه‌گیری نیست (بویل و فیشر ۲۰۰۷). با این حال توجه به روایی صوری آزمون موجب بهبود مقبولیت آن از سوی آزمودنی‌ها می‌شود.

۳) روایی ملاکی

روایی ملاکی به صورت «میزان ارتباط بین نمرات حاصل از یک آزمون با نمرات حاصل از یک آزمون یا وسیله اندازه‌گیری دیگر» تعریف می‌شود (سیف ۱۹۹۰). برای این منظور عملکرد هر فرد در یک آزمون با عملکرد وی در آزمون دیگری که معمولاً معیار یا ملاک نامیده می‌شود، مقایسه می‌شود. اگر دو اندازه‌گیری در یک زمان انجام شوند، عبارت روایی همزمان به کار می‌رود و اگر فاصله زمانی داشته باشند، چون از نتایج آزمون اول برای پیش‌بینی نتایج آزمون دوم (یعنی آزمون معیار) استفاده می‌شود، به آن روایی پیش‌بینی می‌گویند. در هر دو حال، ضریب همبستگی بین نمره‌های حاصل نمایانگر روایی است.

روایی پیش‌بینی به معنای آن است که ارزیابی به ما اطلاعاتی درباره رفتار آینده فراگیران ارائه می‌دهد (فلاچیکو ۲۰۰۴). بنابراین روایی پیش‌بینی اشاره به این مطلب دارد که سؤالات آزمون برای محتوای مورد نظر تا چه حد قابلیت پیش‌بینی عملکرد و دانش آزمودنی‌ها را در موقعیت‌های دیگر و یا حوزه محتوایی دیگر دارد. به چند مثال در این رابطه توجه کنید:

- عملکرد دانشجو در برابر بیمارنما به عنوان پیش‌بینی‌کننده عملکرد واقعی او در مواجهه با بیماران درمانگاه. فرض بر این است که اگر آزمودنی‌ها عملکرد خوبی در برابر بیمارنما داشته باشند، پزشک خوبی برای بیماران واقعی نیز خواهند

بود. در واقع اگر آزمونی از آن‌ها در برابر بیمارنما بگیریم که بتواند همبستگی خوبی با نمرات آتی آنها در برابر بیمار واقعی نشان دهد، آزمون بیارنما از روایی معیاری بالایی برخوردار است.

- رابطه نمرات آزمون آناتومی سیستم عصبی با عملکرد دانشجویان در طول سال‌های بالینی در درک تظاهرات بالینی حوادث مغزی عروقی.
- ارتباط بین عملکرد فراگیر در آزمون جامع علوم‌پایه با عملکرد وی آزمون جامع پیش‌کارورزی.
- ارتباط عملکرد موفقیت‌آمیز دانشجویان سال اول پرستاری در آزمون تشریحی در مورد اصول مراقبت از بیمار با عملکرد پرستاران فارغ‌التحصیل در محیط بالین برای ارائه صحیح و موثر مراقبت‌های درمانی.

در حقیقت در آموزش پزشکی روایی پیش‌بینی به معنای آن است که ما به دنبال طراحی ابزارهایی باشیم که بتوان بر اساس آن‌ها عملکرد فراگیران را در محیط بالین پیش‌بینی نمود تا اینکه صرفاً به ارزیابی دانسته‌های آنان در یک زمینه خاص پرداخت.

۴) روایی سازه

روایی سازه در واقع بیان‌کننده ارتباط بین نظریه و متدلوژی موضوع مورد ارزیابی و نوع ارزیابی است. به عبارت دیگر، در روایی سازه تاکید بر این است که تکنیک‌های ارزیابی باید بر پایه مفهومی باشند که آزمون قصد ارزیابی آن را دارد (امین و کو ۲۰۰۳). همچنین مهم است که نمرات حاصل از اجرای آزمون در راستای مفاهیم و سازه‌های نظری موضوع مورد سنجش باشند. روایی سازه یک ابزار نمایانگر آن است که ابزار تا چه اندازه یک سازه یا خصیصه را طبق مبنای نظری آن می‌سنجد.

مفهوم روایی سازه، چگونگی ایجاد آن و چگونگی اطمینان از حصول به آن بیشتر هنگام ساخت پرسشنامه‌ها مورد بحث و بررسی قرار می‌گیرد. در اینجا برای تقریب ذهن، مثالی از کاربرد روایی سازه در ساخت ابزار ارزیابی فراگیر ارائه می‌کنیم:

فرض کنید که طراح آزمون در نظر دارد برای سنجش مهارت حل مسأله فراگیران یک ابزار تعاملی تدوین کند. او تصمیم می‌گیرد با استفاده از اطلاعاتی که با حضور در محیط‌های بالینی به دست آورده است، تعدادی سناریو طراحی کند. داوطلبان در ابتدا با یک شکایت بالینی روبرو می‌شوند و سپس از آنها خواسته می‌شود از طریق کار با شبیه‌ساز به سؤالات مرتبط درباره انجام معاینه فیزیکی، اخذ شرح حال بیمار، ارائه تشخیص‌های افتراقی و ... پاسخ دهند. برای سنجش عملکرد دانشجو، نمره پاسخ‌های او به سؤالات جمع می‌شود.

این آزمون طبق نظریه حل مسأله، یک آزمون روا نیست. طبق نظریه‌های کنونی مربوط به طرح‌واره‌ها و الگوهای ذهنی، آنچه باعث افتراق متخصصان از دانشجویان از نظر مهارت حل مسأله می‌شود، این است که متخصصان در مقایسه با دانشجویان قادر هستند از طریق جمع‌آوری اطلاعات کمتر تشخیص‌های صحیح‌تری ارائه دهند. به عبارت دیگر، افراد متخصص در جمع‌آوری اطلاعات «کارآمدتر» عمل می‌کنند نه اینکه لزوماً از مهارت بیشتری برخوردار باشند. در حالی که سیستم نمره‌دهی طراحی‌شده، قطعاً نمی‌تواند بین کارآمدی یک متخصص و یک دانشجو تمیز قائل شود. بنابراین دلیل قانع کننده‌ای برای تردید در مورد روایی سازه این ابزار ارزیابی وجود دارد به این خاطر که نمره دانشجو با نظریه‌های زیربنایی موضوع مورد ارزیابی مطابقت ندارد.

گاهی برای یک مفهوم واحد، چندین سازه را می‌توان به صورت اجزای تشکیل‌دهنده آن در نظر گرفت که اگرچه تا حدی به یکدیگر ارتباط دارند اما در عین حال مفاهیم نسبتاً جداگانه‌ای هستند. این جنبه از روایی سازه نیز بیشتر هنگام طراحی پرسشنامه‌ها مطرح است. به عنوان مثال پس از طراحی ابزار سنجش جو آموزشی در دانشگاه داندی (DREEM) طراحان با انجام تحلیل عاملی، ۵ سازه برای مفهوم کلی جو آموزشی استخراج کردند. به این معنا که اگر پرسشنامه‌ای بخواهد جو آموزشی را بسنجد در صورتی واجد روایی سازه است که نشان دهد ۵ سازه آن را مورد سنجش قرار داده است. هرچند که مفهوم روایی سازه چندان برای بحث آزمون‌ها و ارزیابی فراگیر رایج نیست اما کم و بیش همان مفاهیم در اینجا

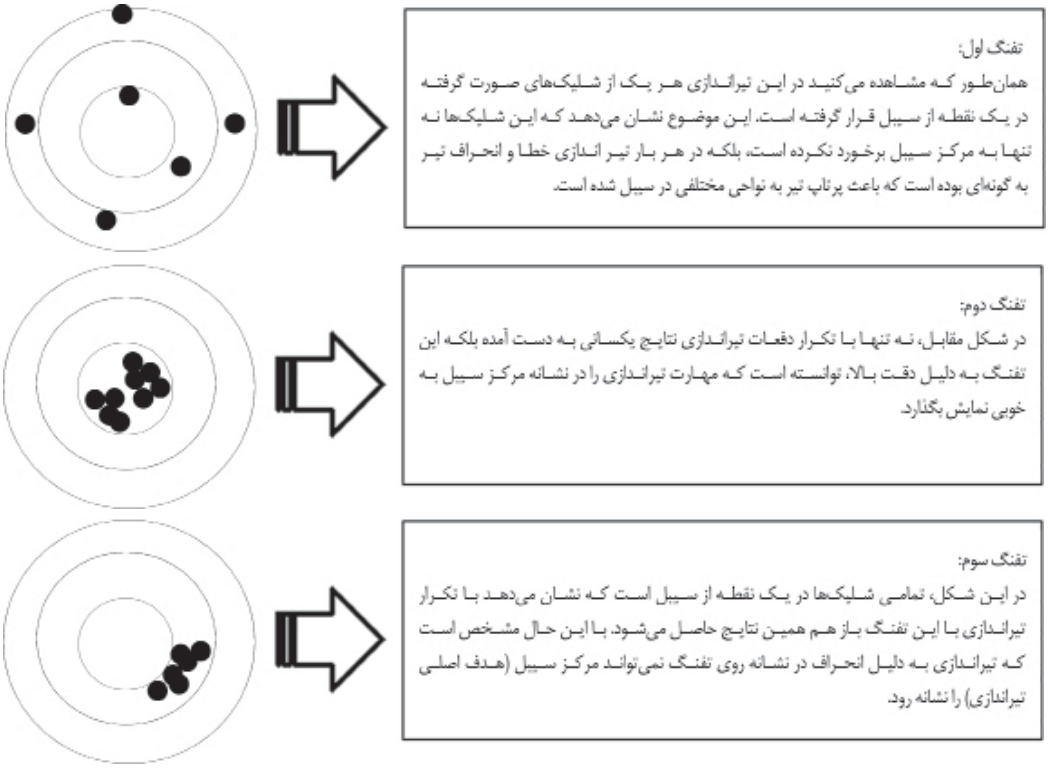
نیز وجود دارد. به عنوان مثال، برای آزمون OSCE در مقطع پیش‌کارورزی می‌توان سازه‌های متفاوتی متصور شد که همان موضوعات و حیطه‌های توانمندی‌ها باشند.

ارتباط بین پایایی و روایی

روایی و پایایی به طور نزدیکی با هم در ارتباط هستند اما به دلیل پیچیدگی‌های این ارتباط، شرح آن قدری دشوار است. پایایی پیش‌نیاز یک آزمون روا و معتبر است و روایی به شدت در یک آزمون غیرپایا تحت تاثیر قرار می‌گیرد. در نقطه مقابل آن، یک آزمون می‌تواند پایایی بالایی (ثبات) داشته باشد، بدون اینکه روا و معتبر باشد. به عبارت دیگر، در صورتی که آزمونی از پایایی بالایی برخوردار باشد نمی‌توان با قطعیت کامل گفت که آن آزمون روا هم هست. جهت درک بهتر این ارتباط شاید شرح مثال زیر کارساز باشد. در شکل ۴-۲ زیر نتایج حاصل از تیراندازی یک نفر (با فرض مهارت بالا) با استفاده از سه تفنگ را مشاهده می‌کنید.

همان‌طور که از مثال فوق به وضوح می‌توان درک کرد، استفاده از ابزار مناسب جهت ارزیابی عملکرد باعث می‌شود که با اطمینان خاطر بتوان گفت که اندازه‌گیری تا چه حد منطبق با اهداف مورد انتظار است و علاوه بر آن، با تکرار اندازه‌گیری همان نتایج اولیه حاصل می‌گردد. مورد اول نشان‌دهنده اهمیت روایی آزمون و مورد دوم مطرح کننده پایا بودن ابزار ارزیابی است.

به عنوان مثالی دیگر، در یک دوره هدف از آزمون، سنجش دانش پایه آزمودنی‌ها در ارائه مشاوره موثر به والدین در



شکل ۴-۲ نتایج حاصل از تیراندازی با استفاده از سه تفنگ (ارتباط بین پایایی و روایی)

مورد ریسک بروز نقص‌های ژنتیکی در حاملگی‌های بعدی است. پرسیدن این سؤال که شایعترین نقص کروموزمی در سندرم داون چیست، اگر چه به دلیل مشخص بودن پاسخ از پایایی بالایی برخوردار است اما به نظر می‌رسد که روایی چندان خوبی نداشته باشد و هدف مورد نظر را ارزیابی نکند. فراگیران باید اطلاعات دیگری فراتر از دانستن صرف این دانش ساده داشته باشند (امین و کو ۲۰۰۳). ارتباط بین روایی و پایایی در مثال زیر بهتر درک می‌شود. تصور کنید هدف آزمون این است که فراگیران علل متداول اختلالات تنفسی را در نوزاد تازه متولدشده تشخیص بدهند. در جدول ۱۹-۲ تعدادی از ابزارهای ارزیابی برای سنجش این هدف و روایی و پایایی آنها نشان داده شده است. قابلیت اجرا فاکتور دیگری است که در قسمت بعدی به آن خواهیم پرداخت.

همان‌طور که در مثال بالا مشاهده می‌کنید، با توجه به آن که هدف آزمون سنجش مهارت داوطلبان در تشخیص اختلالات تنفسی در نوزادان تازه متولد شده است، هر چه آزمون در سنجش این مهارت با واقعیت منطبق‌تر باشد، روایی بالاتری خواهد داشت. به طور کلی ذکر این نکته ضروری است که آزمون پایا به بهبود روایی آن نیز می‌انجامد و آزمونی که پایا نباشد، احتمالاً از روایی پایینی برخوردار خواهد بود.

تقوا و همکاران ۱۳۸۶

این پژوهشگران پژوهشی با عنوان پایایی و روایی نخستین آزمون OSCE روانپزشکی در ایران انجام دادند. این پژوهش با هدف بررسی پایایی و روایی آزمون OSCE در حوزه روانپزشکی و در میان دستیاران انجام شده است. پس از بررسی و بحث‌های فراوان میان استادان روانپزشکی کشور، نه ایستگاه ۱۲ دقیقه‌ای طراحی گردید. در هر ایستگاه دو ارزیاب ممتحن جداگانه به تکمیل چک لیست‌های از پیش طراحی شده پرداختند. در هشت ایستگاه از بیمار استانداردشده استفاده شد. روایی ظاهری و محتوایی، روایی ساختاری، روایی هم‌زمان و پایایی بین دو ارزیاب، ثبات درونی و پایایی بین ایستگاه‌ها بررسی شد. روایی ظاهری و محتوایی آزمون پس از برگزاری جلساتی با حضور استادان و دبیر دانش‌نامه تخصصی کشور و استخراج کلیه هدف های عملی و به بحث گذاشتن آن در نشست های گوناگون به دست آمد. روایی ساختاری با مقایسه نمرات OSCE دستیاران با نمره های استادان مستقیم ایشان و روایی هم زمان با بررسی همبستگی نمره های OSCE با آزمون دانش‌نامه کتبی، شفاهی، ارتقا و ورودی به رشته روانپزشکی بررسی گردید، که نشان دهنده روایی مناسب این آزمون بود. بررسی پایایی آزمون نیز به کمک پایایی بین دو ارزیاب، ثبات درونی و پایایی بین ایستگاه ها نشان داد که این آزمون از پایایی مناسب برخوردار است.

جدول ۱۹-۲: روایی، پایایی و قابلیت اجرای سه ابزار مختلف برای اندازه‌گیری یک هدف (برگرفته از امین و کو ۲۰۰۳)

شکل آزمون	روایی	پایایی	قابلیت اجرا
آزمون بالینی: دانشجویان یک نوزاد مبتلا اختلال تنفسی را مشاهده کنند و تشخیص محتمل را بر اساس شرح‌حال بیماری و معاینه فیزیکی مطرح نمایند.	+++	++	+
آزمون تشریحی تغییریافته: به دانشجویان یک سناریوی بالینی در مورد نوزاد مبتلا به اختلال تنفسی همراه با نتایج آزمایشگاهی و گرافی سینه ارائه شود.	++	++	+++
آزمون چند گزینه‌ای: دانشجویان مجموعه‌ای از سؤالات چندگزینه‌ای در مورد دانش مربوط به اختلالات تنفسی در نوزادان دریافت خواهند کرد.	+	+++	+++

تاثیر آموزشی

از نظر استادان، ارزیابی قسمتی از فرایند آموزشی است که عمدتاً در انتهای دوره اجرا می‌گردد اما از دیدگاه فراگیران، ارزیابی احتمالاً به عنوان اصلی‌ترین جزء برنامه درسی در نظر گرفته می‌شود که اغلب بقیه قسمت‌های آموزش را تحت‌الشعاع قرار می‌دهد به گونه‌ای که برای اغلب دانشجویان تازه آغاز فرایند یادگیری است.

این مفهوم که ارزیابی، محرک و نیروی پیش‌برنده یادگیری است، به عنوان یکی از اصول مهم آموزش در نظر گرفته می‌شود و اشاره به این موضوع دارد که رویکرد به ارزیابی در یک دوره آموزشی به شدت فعالیت‌های یادگیری فراگیران را تحت تاثیر خود قرار می‌دهد (سائر[1] ۱۹۷۱). شایان ذکر است که این مفهوم در متون تحت نام «روایی پیامدی»[2] هم مورد اشاره قرار گرفته است. همچنین عبارت «پیامد برگشتی»[3] برای اشاره به اثراتی که ارزیابی بر فرایند یادگیری فراگیران دارد، مطرح گردید. به این معنا که ارزیابی مشخص می‌کند فراگیران چه مطالبی را چگونه و با چه روشی یاد بگیرند.

به طور خلاصه، هر آزمون علاوه بر اینکه دانش و مهارت و توانمندی فراگیران را می‌سنجد، خودش بر روی یادگیری آن‌ها اثر دارد. پیامدهای حاصل از ارزیابی می‌توانند مثبت و سازنده باشند به گونه‌ای که فراگیران را به یادگیری مناسب تشویق کند. این نتیجه در صورتی حاصل می‌گردد که ارزیابی بر اساس آنچه که فراگیران واقعاً باید یاد بگیرند، تنظیم گردد. این تاثیر از جنبه رویکرد اتخاذی فراگیران برای یادگیری مطالب درسی، عمق یادگیری مطالب و حتی مدت زمانی که صرف یادگیری می‌کنند، قابل بررسی است (بلاکسهام و بیود[4] ۲۰۰۷).

توجه به این نکته ضروری است که در برخی مواقع پیامد برگشتی می‌تواند نامطلوب باشد و باعث سوق دادن فراگیران به یادگیری سطحی یا یادگیری مطالبی شود که اهمیت چندانی ندارند (کروکس[5] ۱۹۸۸، فردریکسن و کولینز[6] ۱۹۸۹). در واقع، بسته به اینکه ارزیابی چگونه و در چه زمینه‌ای انجام شود، می‌تواند اثر آموزشی منفی به دنبال داشته باشد (بیگز ۲۰۰۶). به عنوان مثال، فراگیران از طریق امتحانات قبلی حدس می‌زنند که در امتحانات آتی چه سؤالاتی مطرح خواهد شد و سعی می‌کنند هنگام مطالعه به پاسخ آن سؤالات توجه بیشتری نمایند. این یکی از پیامدهای ارزیابی است که باعث می‌شود که فراگیران به یادگیری سطحی مطالب برای اخذ نمره بیشتر در امتحان روی بیاورند. همچنین این مسأله که عموم آزمون‌های دوره پزشکی شامل سؤالات چندگزینه‌ای هستند که صرفاً بر ارزیابی دانش فراگیران متمرکز هستند، موجب می‌شود فراگیران نیازی به حضور در بخش و بالین بیمار پیدا نکنند، توانمندی‌هایی چون معاینه و شرح‌حال را در خود ارتقا ندهند و نسبت به مهارت برقراری ارتباط و تعهد حرفه‌ای بی توجه باشند.

به صورت کلی، تاثیر آموزشی ارزیابی از چهار جنبه قابل بررسی است:

- محتوای آزمون می‌تواند موجب یادگیری شود. به عنوان مثال، در صورتی که قصد داریم فراگیران در انتهای دوره مهارت حل مسأله کسب کنند، نباید سؤالات حفظ کردنی از آن‌ها بپرسیم.
- شکل آزمون نیز می‌تواند موجب یادگیری شود. به عنوان مثال، طرح سؤالات تشریحی در آزمون منجر به هدایت فراگیران به یادگیری عمیق‌تر مطالب درسی می‌شود (دیویدسون[7] ۲۰۰۲) یا برگزاری OSCE باعث می‌شود دانشجویان به کسب مهارت‌های عملی توجه نشان دهند.
- آزمون بر اساس اطلاعاتی که به فراگیران می‌دهد، موجب یادگیری می‌شود. شاید بتوان گفت که بازخورد دادن به دانشجو بر اساس نتایج آزمون ملموس‌ترین تاثیر مورد انتظار از برگزاری آزمون است.

1. Syer
2. Consequential validity
3. Backwash
4. Blaxham & Boyd
5. Crooks
6. Frederiksen and Collins
7. Davidson

- برنامه‌ریزی آزمون، تاریخ و حتی زمان اجرای آزمون در طول دوره آموزشی نیز بر روند یادگیری فراگیران تاثیرگذار است.

در صورتی که نتایج مورد انتظار یادگیری در فرایند ارزیابی به خوبی منعکس گردد، فعالیت‌های آموزشی استاد و فعالیت‌های یادگیری فراگیران هر دو به سمت یک هدف مشابه حرکت خواهند کرد. بدین ترتیب فراگیران در هنگام آماده شدن برای ارزیابی، نتایج مورد انتظار یادگیری را فرا خواهند گرفت. دستیابی به این موضوع به نظر آسان می‌رسد، اما همچنان تفکرات قدیمی در خصوص ارزیابی وجود دارد که باعث پیچیده کردن این مسأله شده است.

با توجه به مطالب فوق، در نظر گرفتن این نکته ضروری است که اهداف ارزیابی باید کاملاً در راستای اهداف آموزشی دوره باشند. در غیر این صورت، آنچه برای دانشجویان به صورت غالب در می‌آید، اهداف ارزیابی است. یعنی دانشجویان بیشتر از اینکه به دنبال یادگیری اهداف آموزشی باشند، به دنبال یادگیری مسائلی خواهند بود که در آزمون مورد ارزیابی قرار می‌گیرند. بنابراین هنگام بازنگری یک کوریکولوم و اصلاح آن، ایجاد تغییر در نحوه ارزیابی دانشجویان می‌تواند تأثیر آشکارتر و مستقیم‌تری به همراه داشته باشد تا تغییر روش‌های آموزشی. این در حالی است که معمولاً آخرین حوزه‌ای که مورد بازنگری قرار می‌گیرد، ارزیابی دانشجو است.

به طور خلاصه توصیه‌های زیر برای بهبود تأثیر آموزشی روش ارزیابی به کار می‌رود:

- تنوع روش‌های ارزیابی
- طراحی سؤال با هدف سنجش سطوح بالای شناختی
- توجه به ارزیابی‌های تکوینی مکرر و همراه با بازخورد مؤثر در طول دوره در کنار آزمون‌های تراکمی
- در اختیار گذاشتن کلید سؤالات پس از برگزاری آزمون، ارائه بازخورد و بحث در مورد آن‌ها

مقبولیت

در حوزه ارزیابی هم مانند بسیاری از جنبه‌های آموزش آنچه در اکثر مواقع مبنای تصمیم‌گیری است، شواهد و یافته‌های پژوهشی نیستند. بلکه نظرات، ارزش‌ها و باورهای فراگیران، استادان و مؤسسات به شدت در این رابطه تأثیرگذار هستند. بنابراین، اینکه ابزار تا چه حد مورد استقبال و پذیرش دست‌اندرکاران قرار می‌گیرد، از مسائلی است که باید به آن توجه کرد. در واقع استادان هم مانند دانشجویان از الگوهای رفتاری خاصی پیروی می‌کنند که اتفاقاً گاهی به صورت عمیق و ناخودآگاه بروز می‌کنند و بی‌توجهی به آنها مشکلاتی در پی خواهد داشت. به عنوان مثال، معمولاً امتحان‌های بسیار ساختارمند خیلی خوشایند آزمونگران نیست یا اینکه گاهی برخورد رودررو با فراگیران را ترجیح می‌دهند.

از همین رو، برای انتخاب و استفاده از ابزارهای ارزیابی، باید نظرات و عقاید دست اندرکاران را جویا شویم و از آن‌ها در مورد شیوه ارزیابی بازخورد بگیریم. در عین حال، برای ایجاد تغییرات در سیستم ارزیابی، با در اختیار گذاشتن اطلاعات و شواهد کافی و مستدل از قبل ایشان را آماده و توجیه کنیم. مسلم است که مقبولیت یا قابلیت پذیرش[1] به میزان زیادی بر قابلیت اجرای ارزیابی تاثیرگذار است.

هزینه و قابلیت اجرا

در مورد محدودیت منابع، نیاز به توضیح اضافه وجود ندارد. مسلماً ارزیابی خوب، هزینه‌بر است. طراحی سؤال، آموزش آزمونگران، اجرای آزمون، تصحیح پاسخ‌ها، تحلیل آزمون، و بازخورد به دانشجویان و دست‌اندرکاران و مؤسسه فعالیت‌هایی هستند که مستلزم صرف هزینه، زمان و نیروی انسانی هستند. با این همه، توجه به سه نکته حائز اهمیت است:

1. Accessibility

- **سرمایه‌گذاری روی ارزیابی، سرمایه گذاری برای آموزش و یادگیری است.** به عبارت دیگر، ارزیابی خوب با یادگیری خوب همراه است و آن را تسهیل می‌کند (بروان ۲۰۰۴). این موضوع مطرح‌کننده نقش ارزشمندی است که ارزیابی می‌تواند در ارتقاء کیفیت یادگیری فراگیران بازی کند. توجه به این موضوع از آن جهت اهمیت دارد که باید ارزیابی را به گونه‌ای سازمان‌دهی کرد که در دستیابی به اهداف آموزشی نقش یک کاتالیزور را بازی کند. به این معنا که روند دستیابی به پیامدها و اهداف آموزشی را سرعت بخشد.
- **فایده‌ای که از تغییر روش ارزیابی حاصل می‌شـــود، احتمالاً بیشتر از تغییر روش‌های آموزشی است.** در مقایســـه این دو باید گفت که اگرچه هزینه‌ای که برای روش‌های ارزیابی صرف می‌شـــود به مراتب بیشتر است، اما هزینه کردن برای آن از ســـوی موسســـات آموزشـــی و حتی اعضای هیأت علمی بیشتر مورد قبول واقع می‌شود. در نتیجه هنگام تخصیص منابع باید این مسأله را هم در نظر داشت که تا چه حد با صرف منابع می‌توان به ایجاد تغییر در یادگیری فراگیران کمک کرد.
- **سرمایه‌گذاری روی ارزیابی، اطمنیان از کیفیت عملکرد فارغ التحصیلان آتی را در پی خواهد داشت.** شاید بتوان این گونه گفت که سرمایه‌گذاری در بحث ارزیابی صحیح از عملکرد فراگیران، زمینه‌ساز پیاده‌سازی آزمون‌ها با کیفیت بالاتر است که به نوبه خود بر سنجش دقیق‌تر عملکرد فارغ‌التحصیلان نیز تاثیر گذار خواهد بود. این موضوع به ویژه از سوی مؤسسات اعتبارسنجی جهت ارائه و یا تایید مجوز بسیار حائز اهمیت است.

از بعد قابلیت اجرا[1] توجه به این نکته ضروری است که حتی بهترین و یا با کیفیت‌ترین آزمون‌ها ممکن است قابلیت اجرا و پیاده‌سازی نداشته باشند و این مسأله لزوماً ارتباطی به هزینه ندارد. به عنوان مثال، اجرای ارزیابی ۳۶۰ درجه از یک طرف نیازمند طراحی فرم‌های گوناگون برای ارزیابی توسط استاد، دستیار، پرسنل، بیمار و ... است و از طرف دیگر، جمع‌آوری و جمع‌بندی فرم‌ها از منابع گوناگون موجب تحمیل حجم کاری فراوان می‌شود. بنابراین در انتخاب کارآمدترین ابزارهای ارزیابی، همواره سنجش قابلیت اجرا و هزینه اقتصادی نقش دو کفه ترازو را بازی می‌کنند که عدم توجه به هریک موجب برهم خوردن توازن تصمیم‌گیری می‌شود.

1. Feasibility

منابع

1. Amin Z, Chong YS, Khoo HE. Practical guide to medical student assessment: World Scientific; 2006.
2. Amin Z, Khoo HE. Basics in medical education: World Scientific; 2003.
3. Anastasi A, Urbina S. Norms and the meaning of test scores. Psychological testing. 1997:49-80.
4. Andrade H, Cizek GJ. Handbook of formative assessment: Routledge; 2010.
5. Bhola HS. Evaluating» Literacy for Development» Projects, Programs and Campaigns. Evaluation Planning, Design and Implementation, and Utilization of Evaluation Results. UIE Handbooks and Reference Books 3: ERIC; 1990.
6. Biggs J. Teaching for quality learning at university. 1999. SRHE & Open University Press, Buckingham. 2006.
7. Bloxham S, Boyd P. Developing Effective Assessment In Higher Education: A Practical Guide: A Practical Guide: McGraw-Hill Education (UK); 2007.
8. Boyle J, Fisher S. Educational testing: a competence-based approach: John Wiley & Sons; 2008.
9. Brown GA, Bull J, Pendlebury M. Assessing student learning in higher education: Routledge; 2013.
10. Crooks TJ. The impact of classroom evaluation practices on students. Review of educational research. 1988;58(4):438-81.
11. Davidson RA. Relationship of study approach and exam performance. Journal of Accounting Education. 2003;20(1):29-44.
12. Elton L, Johnston B. Assessment in Universities: a critical review of research. 2002.
13. Epstein RM, Hundert EM. Defining and assessing professional competence. Jama. 2002;287(2):226-35.
14. Falchikov N. Involving students in assessment. Psychology Learning & Teaching. 2004;3(2):102-8.
15. Falchikov N, Boud D. Student self-assessment in higher education: A meta-analysis. Review of Educational Research. 1989;59(4):395-430.
16. Frederiksen JR, Collins A. A systems approach to educational testing. Educational researcher. 1989;18(9):27-32.
17. George J, Cowan J. 4 handbook oftechniquesfor formative assessment Mapping the student›s learning experience. London: Kogan Page; 1999.
18. Gregory RJ. Psychological testing: History, principles, and applications: Allyn & Bacon; 2004.
19. Goldszmidt M, Chahine S, Cristancho S, Watling C, Lingard L. On the value of the 'subjective'in studies of human behavior and cognition. Perspectives on medical education. 2015;4(1):49-50.
20. Haines C. Assessing students› written work: marking essays and reports: Routledge; 2004.
21. Nevo D, Nevo D. School-based evaluation: A dialogue for school improvement: Pergamon Oxford; 1995.

22. Norman G. Postgraduate assessment–reliability and validity. Trans J Coll Med S Afri. 2003;47:71-5.
23. Pangaro L, ten Cate O. Frameworks for learner assessment in medicine: AMEE Guide No. 78. Medical teacher. 2013;35(6):e1197-e210.
24. Rowntree D. Assessing students: How shall we know them?: Taylor & Francis; 1987.
25. Schuwirth L, Colliver J, Gruppen L, Kreiter C, Mennin S, Onishi H, et al. Research in assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. Medical teacher. 2011;33(3):224-33.
26. Shumway JM, Harden RM, Europe AfMEi. The assessment of learning outcomes for the competent and reflective physician: AMEE; 2003.
27. Smith SR. AMEE Guide No. 14: Outcome-based education: Part 2-Planning, implementing and evaluating a competency-based curriculum. Medical teacher. 1999;21(1):15-22.
28. Swanson DB, Case SM, van der Vleuten CP, Boud D. Strategies for student assessment. The challenge of problem based learning. 1991:260-73.
29. Swanson DB, Norman GR, Linn RL. Performance-based assessment: Lessons from the health professions. Educational Researcher. 1995;24(5):5-11.
30. Syer CA, Shore BM. Science Fairs: What Are the Sources of Help for Students and How Prevalent Is Cheating? School Science and Mathematics. 2001;101(4):206-20.
31. Tavakol M, Dennick R. Post-examination analysis of objective tests. Medical teacher. 2011;33(6):447-58.
32. Tavakol M, Dennick R. Post-examination interpretation of objective test data: Monitoring and improving the quality of high-stakes examinations: AMEE Guide No. 66. Medical teacher. 2012;34(3):e161-e75.
33. van Merrienboer JJ. What people say# what people do. Perspectives on medical education. 2015;4(1):47-8.
34. Wass V, McGibbon D, Van der Vleuten C. Composite undergraduate clinical examinations: how should the components be combined to maximize reliability? Medical Education. 2001;35(4):326-30.
35. Whitcomb ME. Competency-based graduate medical education? Of course! But how should competency be assessed? Academic Medicine. 2002;77(5):359-60.
36. Wiliam D, Black P. Meanings and consequences: a basis for distinguishing formative and summative functions of assessment? British Educational Research Journal. 1996;22(5):537-48.
37. Wright BD, Stone MH. Best Test Design. Rasch Measurement. 1979.
۳۸. سیف، ع ا. اندازه‌گیری، سنجش و ارزشیابی آموزشی. نشر دوران، ۱۳۹۴.
۳۹. ذوالفقاری ب، ادیبی ن، درخشان‌فر ث، تن‌ساز م. آزمون‌های پیشرفت تحصیلی در علوم پزشکی. انتشارات مدیریت مطالعات و توسعه آموزش پزشکی. ۱۳۷۹.
۴۰. دلاور ع. مقدمه‌ای بر نظریه‌های اندازه‌گیری (روانسنجی)، ۱۳۸۹.

۴۱. عسگری ع. روی‌آوردهای نوین در روان‌سنجی، قسمت دوم: مبانی و مفاهیم نظریه سؤال پاسخ. روانشناسی تحولی، ۱۳۸۶؛ ۳(۱۲):۳۶۷-۳۷۰
۴۲. تقوا آ، رسولیان م، پناغی ل، و همکاران. پایایی و روایی نخستین آزمون ساختاریافته عینی بالینی (OSCE) روانپزشکی در ایران. مجله روانپزشکی و روانشناسی بالینی ایران، ۱۳۸۶؛ ۱۳(۱): ۱۷-۲۴

فصل ۳

چارچوب‌های رایج در ارزیابی فراگیر

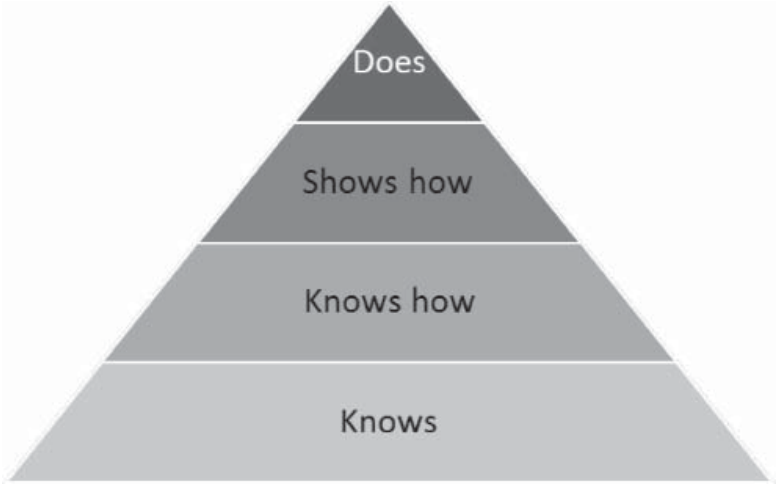
مقدمه

در هر سیستم ارزیابی، مفاهیم زیربنایی مرتبط با اهداف و پیامدهای ارزیابی انتخابی، روش‌های موجود برای اخذ قضاوت و تحلیل و مقایسه با استانداردها تحت عنوان چارچوب‌های ارزیابی مطرح می‌شوند. اثرات بالقوه زیادی برای استفاده از چارچوب‌های ارزیابی برشمرده‌اند. شاید بتوان گفت که اولین اثر استفاده از چارچوب‌های ارزیابی در ارتباط با هدایت استادان در مشاهداتی است که از عملکرد فراگیران خواهند داشت. طراحی بلوپرینت آزمون که بر اساس آن می‌توان سؤالات امتحانی را انتخاب نمود و تهیه چک‌لیست‌های ارزیابی عملکرد نیز می‌تواند بر اساس چنین چارچوب‌هایی صورت گیرد. هر چند باید به این موضوع اعتراف کرد که فراگیران مطالب زیادی را خارج از آنچه که در برنامه‌های درسی آنان لحاظ شده فرا می‌گیرند، با این حال چارچوب‌های ارزیابی منعکس‌کننده ارزش‌ها و انتظارات مؤسسه از آنان هستند. موفقیت در به کارگیری این چارچوب‌ها به میزان زیادی به وضوح مطالب ارائه شده، سهولت استفاده از آن‌ها و پاسخگویی به ارزش‌های ذی‌نفعان مربوط می‌شود. بر همین اساس در این فصل سعی خواهد شد تعدادی از چارچوب‌هایی را که در ارزیابی کاربرد دارند، معرفی کنیم. البته تعدادی از این چارچوب‌ها که به طور کلی تحت عنوان ارزیابی مبتنی بر توانمندی مورد بحث قرار می‌گیرند و در فصل دوم همین بخش ارائه شدند، در این قسمت تکرار نمی‌شوند.

هرم میلر

در طول دهه‌های گذشته، ارزیابی مهارت‌های عملی فراگیران علوم پزشکی مخصوصاً در محیط‌های کاری[1] به عنوان یک جزء ضروری ارزیابی عملکرد در نظر گرفته شده است. به دنبال گسترش موج علاقه به ارزیابی توانمندی‌ها در محیط‌های کاری، جرج میلر[2] در سال ۱۹۹۰ الگویی برای ارزیابی توانمندی‌های بالینی پیشنهاد کرد که به «هرم میلر»[3] شهرت دارد. این هرم صلاحیت بالینی را به چهار سطح تقسیم می‌کند: «می‌داند»[4]، «می‌داند چگونه»[5]، «نمایش می‌دهد چگونه»[6] و «انجام می‌دهد»[7] (شکل ۱-۳).

1. Workplace
2. George Miller
3. Miller's Pyramid
4. Know
5. Knows how
6. Shows
7. Does



شکل ۱-۳: هرم میلر

داوطلبی که در سطح «می‌داند» است، در ابتدای راه است و دانسته‌های وی شامل حقایق[1] است. «می‌داند چگونه» در رابطه با مفهوم‌سازی و درک عمیق‌تر دانسته‌ها است. در یک سطح بالاتر زمانی که داوطلب «نمایش می‌دهد چگونه» یعنی مهارت خود را در محیطی شبیه‌سازی شده اجرا می‌کند. در بالاترین سطح هرم یعنی «انجام می‌دهد»، از داوطلب انتظار می‌رود وظیفه مشخص‌شده را در شرایط واقعی به صورت ماهرانه و با صلاحیت انجام دهد (بیگز ۲۰۰۶).

به صورت کلی می‌توان گفت که هرم میلر یک چارچوب مفید برای تشخیص تفاوت‌ها و شباهت‌های بین ابزارهای ارزیابی فراهم می‌کند و استفاده از آن برای انتخاب ابزار ارزیابی آزمون مفید است:

- **می‌داند:** پایین‌ترین سطح هرم میلر شامل روش‌هایی می‌شود که به ارزیابی دانسته‌های فراگیران در زمینه توانمندی مورد انتظار می‌پردازند. در واقع این‌گونه می‌توان تصور کرد که یادگیری دانش، شالوده شکل‌گیری توانمندی‌ها در نظر گرفته می‌شود. به عنوان مثال، طراحی آزمون چندگزینه‌ای که اصول اخلاقی و حفظ اسرار بیماران را می‌سنجد، پایه‌ای برای تصمیم‌گیری در خصوص دانسته‌های دانشجویان در زمینه تعهد حرفه‌ای فراهم می‌سازد.
- **می‌داند که آن عمل را چگونه باید انجام دهد:** برای عمل به عنوان یک پزشک، اگرچه برخورداری از پایه دانش خوب ضروری به نظر می‌رسد، اما کافی نیست. دانستن اینکه چطور از دانش موجود در زمینه جمع‌آوری داده‌ها، تحلیل و ارزیابی یافته‌ها و در نهایت تصمیم‌گیری در خصوص وضعیت بیمار استفاده شود، به همان اندازه اهمیت دارد. به عنوان مثال، آزمونی که از طریق طرح یک معضل اخلاقی سعی در ارزیابی توانایی استدلال فراگیران دارد، اطلاعاتی را در مورد تفکر اخلاقی آنان در مواجهه با موقعیت‌های پیش رو در بالین ارائه می‌دهد.
- **چگونگی انجام آن را نشان می‌دهد:** دانستن چگونگی انجام یک عمل نیز هیچ تضمینی برای مهارت در انجام آن به شمار نمی‌رود. فراگیران ممکن است از پایه دانشی خوبی برخوردار باشند و حتی از چگونگی بکاربستن آن در موقعیت‌های مختلف اطلاع داشته باشند، اما در مواجهه با بیماران نتوانند عملکرد مناسبی نشان دهند. به عنوان مثال، استفاده از بیمارنما که با یک چالش اخلاقی مواجه شده است، این فرصت را در اختیار فراگیر می‌گذارد که مهارت مدیریت یک چالش حرفه‌ای را در عمل به نمایش بگذارد.
- **عملاً آن را انجام می‌دهد:** نقدی که همواره به روش‌های ارزیابی سنتی وارد می‌شود، این است که اطلاعات

1. Factual

حاصل از ارزیابی در یک محیط کنترل‌شده و شبیه‌سازی شده نمی‌تواند پیش‌بینی کننده عملکرد فراگیران در محیط‌های واقعی باشد. بنابراین نتایج حاصل از این آزمون‌ها نمی‌تواند مستقیماً به محیط‌های بالینی تعمیم داده شود. بنابراین بالاترین سطح هرم ارزیابی میلر اختصاص به روش‌هایی دارد که به ارزیابی عملکردهای معمول پزشکان در موقعیت‌های حرفه‌ای‌شان می‌پردازد. به عنوان مثال، استفاده از سیستم ثبت رویدادهای مهم می‌تواند تصویری از عملکردهای واقعی دانشجویان را در محیط‌های بالین از منظر تعهد حرفه‌ای ارائه نماید.

توجه به این نکته ضروری است که ابزارهای ارزیابی گوناگون برای ارزیابی سطوح مختلف صلاحیت کاربرد دارند. به عنوان مثال در حالی‌که سؤالات چند گزینه‌ای برای آزمودن سطوح یک و دو مفید است، کارایی آنها برای ارزیابی سطح سه و چهار محدود است. به صورت مشابه، آزمون OSCE اگرچه برای ارزیابی سطح «نمایش می‌دهد» بسیار مناسب است اما اگر شخص بخواهد سطح وسیعی از دانش را در سطح «می‌داند» مورد ارزیابی قرار دهد استفاده از آن چندان کاربردی ندارد. جدول ۱-۳ روش‌های ارزیابی متداولی را که در هر از سطوح هرم میلر قابل استفاده هستند، معرفی می‌کند.

هنگام انتخاب ابزار بر اساس هرم میلر توجه به این نکته ضروری است که ممکن است به نظر برسد شکل هرم‌مانند آن دال بر این مفهوم است که روش‌های ارزیابی سطوح بالاتر از ارزش بیشتری برخوردار هستند. اما واقعیت امر آن است که معیار ارزشمندی یک ابزار تطبیق آن با اهداف آزمون است. به عنوان مثال، در صورتی که سنجش دانسته‌های فراگیران مدنظر است، همان‌طور که پیش‌تر نیز اشاره شد، استفاده از آزمون چندگزینه‌ای در مقایسه با OSCE بهتر خواهد بود.

همچنین توجه به این نکته حائز اهمیت است که صلاحیت بالینی قابل تفکیک نیست و دانش و صلاحیت بالینی یک ویژگی کلی است. به‌همین دلیل زمانی که در حال برنامه‌ریزی یک نظام برای ارزیابی صلاحیت بالینی هستیم، مهم است که حداقل یک یا دو ابزار ارزیابی برای هر سطح از این هرم انتخاب کنیم تا توانایی داوطلب به صورت جامع و کامل مورد ارزیابی قرار گیرد. در عین حال، هیچ روش ارزیابی منحصر به فردی، برای به دست آوردن اطلاعات مورد نیاز برای قضاوت در مورد مقوله پیچیده‌ای همچون ارائه خدمات حرفه‌ای به وسیله یک پزشک موفق وجود ندارد (میلر ۱۹۹۰). این موضوع مطرح کننده نقش به‌سزایی است که طراحی نظام ارزیابی در سنجش صلاحیت پزشک بازی می‌کند. جهت کسب اطلاعات بیشتر در این خصوص به بخش نهم کتاب مراجعه نمایید.

تاکسونومی بلوم

آنچه که رالف تایلر[1] در سال ۱۹۴۹ تحت عنوان «منطق تایلری»[2] ارائه داد، زمینه‌ساز حرکت آموزش مبتنی بر پیامد شد. این نظریه، زیربنای اصلی برای طرح چهار سؤال ساده اما مهم را فراهم ساخت:

- موسسات آموزشی باید در جستجوی دستیابی به چه مقاصد آموزشی باشند؟
- تجربیات یادگیری برای دستیابی به این اهداف چه هستند؟

جدول ۱-۳: ابزارهای ارزیابی متناسب با چهار سطح هرم میلر

سطح ارزیابی	ابزار ارزیابی
می‌داند و می‌داند چگونه	آزمون شفاهی، سؤالات تشریحی، سؤالات کوته‌پاسخ، سؤالات چندگزینه‌ای، سؤالات «جورکردنی گسترده» ، آزمون «ویژگی‌های کلیدی»
نمایش می‌دهد چگونه	OSCE، مورد بالینی کامل، مورد بالینی کوتاه
انجام می‌دهد	DOPS، Mini-CEX ،چک لیست، ارزیابی ۳۶۰ درجه، لاگ‌بوک، کارپوشه

1. Ralph Tyler
2. Tyler Rationale

- چگونه می‌توان تجربیات یادگیری را به طور اثربخش سازمان‌دهی کرد؟
- چگونه می‌توان تعیین کرد که آیا این مقاصد تحقق یافته‌اند یا نه؟

سؤال اول و چهارم تایلر جرقه اولیه برای طرح ایده چارچوب‌های ارزیابی را فراهم ساخت (پنگارو و تن‌کیت ۲۰۱۳). از زمان تایلر به بعد، دانشمندان بسیاری به بسط این مفهوم پرداختند. از جمله دانشمندان پیشرو در این حیطه بنجامین بلوم بود. بلوم نیز همانند تایلر معتقد بود که نوشتن هدف، فرایند آموزش را روشن می‌کند و از جمله اینکه به طراحی اثربخش آزمون کمک زیادی خواهد کرد. او تحت تاثیر تفکرات استاد خود رالف تایلر، در سال ۱۹۵۶ درصدد سازمان‌دهی اهداف آموزشی بر اساس درجه پیچیدگی آنان برآمد (مهرمحمدی ۱۳۹۱). ماحصل تلاش‌های او در این زمینه منجر به ارائه شکل جدیدی از طبقه بندی اهداف در سه حیطه شناختی[1]، عاطفی[2] و روانی- حرکتی[3] شد. متعاقب آن سه گروه تشکیل شد تا هر یک روی یکی از حیطه‌ها کار کنند و سلسه مراتب مربوط به آن حیطه را استخراج کنند. نتیجه کار بلوم و همکارانش چاپ دو کتابچه بود که سلسه مراتب حیطه شناختی و حیطه عاطفی را مشخص می‌کردند اما این پژوهشگران نتوانستند در زمینه روانی–حرکتی به ارائه تاکسونومی بپردازند.

تقسیم‌بندی حیطه شناختی طبق پیشنهاد بلوم و همکاران به این صورت بود: یادآوری، درک، کاربرد، تحلیل، ترکیب و ارزشیابی (بلوم و همکاران ۱۹۵۶). هر یک از سطوح یادگیری که بالاتر قرار دارد، به نحوه اجرای سطوح ماقبل خود وابسته است. این تقسیم‌بندی بعدها توسط شاگردان بلوم اصلاح شد و امروزه به این شکل مطرح می‌شود: یادآوری، درک، کاربرد، تحلیل، ارزشیابی و خلاقیت (اندرسون و کراتهول[4] ۲۰۰۱). بیشترین مورد استفاده این تاکسونومی در مشخص کردن سطح شناختی آزمون‌ها مخصوصاً سؤالات چندگزینه‌ای است. پیشنهاد گروه بعدی که روی حیطه عاطفی کار کرده بود، به این شکل بود: توجه، پاسخ، ارزش‌گذاری، سازمان‌دهی ارزش‌ها و درونی شدن ارزش‌ها (کراتهول و همکاران ۱۹۶۴). در خصوص حیطه روانی–حرکتی پژوهشگرن دیگری به پیشنهاد جارچوب پرداخته‌اند. یکی از معروف‌ترین آنها که توسط الیزابت سیمپسون ارائه شد فراگیری یک هدف سایکوموتور را شامل این موارد می‌داند: تقلید، اجرای مستقل، دقت، هماهنگی حرکات و در نهایت، عادی شدن رفتار (سیمپسون[5] ۱۹۷۲).

تاکسونومی SOLO[6]

بیگز و کولیز[7] در سال ۱۹۸۲ ساختاری به نام تاکسونومی SOLO ارائه دادند که چارچوبی برای تمایز بین مراحل مختلف یادگیری فراهم می‌سازد. در تاکسونومی SOLO که شامل پنج فاز یا سطح یادگیری است، انتخاب افعال مناسب برای توصیف هر یک از مراحل یادگیری یک اصل اساسی است. در جدول ۲–۳ افعال مورد استفاده در هر فاز یادگیری ارائه شده است.

- **پیش‌ساختاری**[8]: در این سطح فراگیران به جمع‌آوری حجم وسیعی از اطلاعات غیرساختارمند و بی‌ربط با موضوع مورد نظر می‌پردازند.
- **تک‌ساختاری**[9]: در این سطح فراگیران تنها بخشی از مفاهیم را در ارتباط با موضوع یادگیری ارائه می‌دهند. به عنوان مثال، در این فاز فراگیر ممکن است تنها به توصیف فعالیت‌های یادگیری و عوامل موثر بر آن بپردازد، بدون آنکه به منابع آموزشی توجه داشته باشد.

1. Cognitive domain
2. Affective domain
3. Psychomotor
4. Anderson & Krathwohl
5. Simpson
6. Structure of Observed Learning Outcomes (SOLO) taxonomy
7. Biggs & Collis
8. Pre-structural
9. Uni-structural

- **چندساختاری**[1]: در این سطح فراگیران از رویکردهای سطحی برای یادگیری حقایق استفاده می‌کنند و از رویکردهای عمیق برای درک معانی موضوعات مختلف استفاده می‌کنند. بدین ترتیب دانشجویان قادر به درک کل موضوع هستند اما نمی‌توانند به درک جزئیات موضوع نائل آیند.
- **منطقی**[2]: در این فاز فراگیران از طریق شرح جزئیات موضوع، بین قسمت‌های مختلف در جریان یک روند منطقی ارتباط برقرار می‌کنند.
- **انتزاعی**[3]: در این سطح پاسخ‌های ارائه شده بسیار فراتر از آنچه که باید باشند، بیان می‌شوند در حالی‌که پاسخ‌های منطقی و محدود قابل پیش‌بینی هستند. به عبارت دیگر، در حالی‌که پاسخ‌های مرتبط با موضوع بیان می‌شود و بین مفاهیم مختلف انسجام کلی وجود دارد، فراگیر قادر است که به راه‌حل‌های خارج از چارچوب مورد انتظار نیز اشاره کند.

در پایان این فصل توجه شما را به جدول ۳-۳ که ویژگی‌های چارچوب‌های فوق‌الذکر را خلاصه کرده است، جلب می‌کنیم:

جدول ۲-۳: بعضی از افعال مربوط به پیامدهای مورد انتظار یادگیری در تاکسونومی SOLO

سطح	فعل
پیش‌ساختاری	-
تک‌ساختاری	حفظ کردن، شناسایی کردن، تشخیص دادن، حساب کردن، تعریف کردن، پیدا کردن، برچسب زدن، مسابقه دادن، نامیدن، نقل قول کردن، خواندن، نظم دادن، نوشتن، تقلید کردن
چندساختاری	طبقه‌بندی‌کردن، شرح دادن، فهرست کردن، گزارش دادن، بحث کردن، نشان دادن، انتخاب کردن، محاسبه‌کردن
منطقی	به کار بردن، ادغام کردن، تجزیه و تحلیل کردن، توضیح دادن، پیش بینی کردن، نتیجه گیری‌کردن، خلاصه کردن، بررسی کردن، استدلال کردن، انتقال دادن، ایجاد کردن، توصیف کردن، مقایسه کردن، سازمان‌دهی کردن، بحث کردن، ساختن، بررسی کردن، ترجمه کردن، تفسیر کردن، حل مشکل
انتزاعی	نظریه‌سازی، فرضیه‌سازی، تعمیم دادن، منعکس کردن، تولید کردن، ایجاد کردن، ابداع کردن، حل کردن با اصول اولیه

1. Multi-structural
2. Relational
3. Extended abstract

جدول ۳-۳: مرور چارچوب‌های رایج در ارزیابی

چارچوب	سطوح یا طبقات	توضیح
هرم میلر	می‌داند	دانش: اطمینان از اینکه دانشجو، دستیار یا پزشک می‌دانند که برای انجام عملکرد حرفه‌ای به طور موثر چه مطالبی موردنیاز است.
	می‌داند چگونه	دانش کاربردی: اطمینان از اینکه دانشجو، دستیار یا پزشک مهارت جمع‌آوری اطلاعات، تحلیل و تفسیر داده‌ها را برای تشخیص منطقی و ارائه برنامه درمانی کسب کرداند.
	نمایش می‌دهد چگونه	مهارت: اطمینان از اینکه دانشجو، دستیار یا پزشک قادر هستند در محیط‌های کنترل شده انجام عمل را به نمایش بگذارند.
	انجام می‌دهد	عملکرد: اطمنیان از اینکه دانشجو، دستیار یا پزشک می‌توانند عمل مو ردانتظار را در محیط‌های واقعی انجام دهند.
تاکسونومی بلوم	حیطه شناختی	دانش، درک و فهم، کاربرد، تجزیه و تحلیل، ترکیب، ارزشیابی
	حیطه نگرشی	دریافت و توجه، واکنش، ارزش‌گذاری، سازمان‌دهی ارزشها، تبلور و درونی‌سازی
تاکسونومی SOLO	پیش‌ساختاری	دانشجو حجم وسیعی از اطلاعات را بدون هیچ ارتباط و ساختار مشخص دریافت می‌کند
	تک‌ساختاری	دانشجو قادر است ارتباط ساده و واضحی را بین اطلاعات مختلف برقرار سازد.
	چندساختاری	دانشجو قادر است ارتباطات بین اطلاعات مختلف را شناسایی کند اما توانایی استفاده از فراشناخت را ندارد.
	منطقی	دانشجو قادر است تاثیر معنادار ارتباط بین موضوعات مختلف را با یکدیگر مشاهده کند.
	انتزاعی	دانشجو قادر است یادگیری های خود را در موقعیت‌های جدید به کار گیرد و یا راه حل‌های فراتر ارائه نماید.

منابع

1. Anderson LW, Krathwohl DR. (Eds.). A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives. New York: Longman. 2001
2. Battistone MJ, Milne C, Sande MA, Pangaro LN, Hemmer PA, Shomaker TS. The feasibility and acceptability of implementing formal evaluation sessions and using descriptive vocabulary to assess student performance on a clinical clerkship. Teaching and learning in medicine. 2002;14(1):5-10.
3. Biggs J. Teaching for quality learning at university. 1999. SRHE & Open University Press, Buckingham. 2006.
4. Biggs JB, Collis KF. Evaluation the quality of learning: the SOLO taxonomy (structure of the observed learning outcome): Academic Press; 1982.
5. Bloom BS. Taxonomy of Educational Objectives: The Classification of Educational Goals. New York: David McKay Company, Inc. 1965
6. Frank JR. The CanMEDS 2005 physician competency framework: Better standards, better physicians, better care: Royal College of Physicians and Surgeons of Canada; 2005.
7. General Medical Council. 2003. Tomorrow's Doctors. Recommendations on undergraduate medical education. UK: GMC.
8. Hemmer PA, Papp KK, Mechaber AJ, Durning SJ. Evaluation, grading, and use of the RIME vocabulary on internal medicine clerkships: results of a national survey and comparison to other clinical clerkships. Teaching and learning in medicine. 2008;20(2):118-26.
9. Krathwohl DR., Bloom BS, Masia BB. Taxonomy of Educational Objectives. The Classification of Educational Goals, Handbook II: Affective Domain. New York: David McKay Company, Inc. 1964
10. Miller GE. The assessment of clinical skills/competence/performance. Academic medicine. 1990;65(9):S63-7.
11. Pangaro L. A new vocabulary and other innovations for improving descriptive in-training evaluations. Academic Medicine. 1999;74(11):1203-7.
12. Pangaro L, ten Cate O. Frameworks for learner assessment in medicine: AMEE Guide No. 78. Medical teacher. 2013;35(6):e1197-e210.
13. Simpson EJ. The classification of educational objectives in the psychomotor domain. The Psychomotor Domain. 1972; 3:43-56. Gryphon House.

فصل | ٤ |

نظریه‌های اندازه گیری

مقدمه

نگاهی به روند تغییرات صورت گرفته در علم اندازه‌گیری حاکی از بروز پیشرفت‌های قابل توجه در قرن ۱۹ میلادی است. دیدگاه‌های مختلفی در جهت تحلیل نتایج پس از اندازه‌گیری مطرح است که در این فصل به شرح آن‌ها خواهیم پرداخت.

به طور کلی، نظریه کلاسیک آزمون به عنوان اولین نظریه منسجم در اوایل قرن ۲۰ شکل گرفت (کورویل[1] ۲۰۰۴). از لحاظ تاریخی، تلاش‌های اسپیرمن در دهه ۱۹۸۰، پایه‌های تکوین نظریه کلاسیک آزمون را فراهم ساخت و سپس فیشر (۱۹۲۵) در کتاب معروف خود به نام «روش‌های آماری برای محققان» مبانی آن را تحکیم کرد (دلاور ۱۳۸۹). زمان زیادی از ظهور این نظریه نگذشته بود که تلاش برای توسعه و تدوین ابزارهای اندازه‌گیری بر طبق اصول نظریه کلاسیک به سرعت گسترش یافت. اوج تکامل معرفی این نظریه را می‌توان در کتاب‌های «مبانی نظری آزمون‌های روانی» و «نظریه‌های آماری نمرات آزمون‌های روانی» مشاهده کرد (همبلتون و ون‌درلیندن[2] ۱۹۸۲).

گرچه نظریه کلاسیک مدت زمان طولانی زیربنای تصمیم‌گیری در علم روانسنجی بود، با این حال برخی مطالعات، محدودیت‌هایی را در این نظریه و در آزمون‌های ساخته‌شده بر اساس آن نشان دادند (مگنو[3] ۲۰۰۹). بنابراین هم‌زمان با بسط این مدل، ضعف‌های جدی آن بیش‌تر آشکار شد و روان‌سنجان و متخصصان آزمون‌سازی را بیش از پیش به‌سمت مدل‌های نوین اندازه‌گیری سوق داد (بوک[4] ۱۹۹۷). به طوری که از نیمه دوم قرن بیستم به بعد، به‌تدریج زمینه برای ارائه نظریات جدید فراهم شد و افرادی مانند لرد (۱۹۵۲)، راش[5] (۱۹۵۸)، رایت[6] (۱۹۶۸)، همبلتون (۱۹۷۹) در این مسیر گام‌های مؤثری برداشتند.

در بحث ارزیابی فراگیر که به نوعی اندازه‌گیری در آموزش محسوب می‌شود، حفظ کیفیت روش‌های ارزیابی به اندازه کیفیت فرایندهای یاددهی–یادگیری اهمیت دارد. امتحانات پس از اجرا باید با استفاده از مشخصه‌های روانسنجی به منظور شناسایی، کنترل و بهبود کیفیت فرایند ارزیابی بررسی و تحلیل شوند. به این ترتیب از نتایج تحلیل پس‌آزمون می‌توان به عنوان بازخوردی نه تنها برای بهبود روایی و پایایی فرایند ارزیابی استفاده نمود، بلکه همچنین با کمک اطلاعات حاصل می‌توان به بهبود کیفیت برنامه درسی و استراتژی‌های آموزشی دست یافت (توکل و دنیک[7] ۲۰۱۱؛ ۲۰۱۲). تحلیل پس از

1. Courville
2. Hambleton & Van der Linden
3. Magno
4. Bock
5. Rash
6. Wright
7. Dennick

آزمون، که به کمک نظریه‌های اندازه‌گیری امکان‌پذیر است، امکان شناسایی سؤالات نادرستی که ممکن است به کاهش کیفیت آزمون منتهی شوند را فراهم می‌سازد (توکل و دنیک ۲۰۱۱). به دلیل اهمیت و جایگاه آموزش پزشکی، لازم است اطلاعات حاصل از امتحانات پزشکی به منظور شناسایی، کنترل و بهبود خطاهای احتمالی با استفاده از مشخصه‌های روانسنجی بررسی و تحلیل شوند. با استفاده از نتایج حاصل از تحلیل پس آزمون می‌توان ارزیابی دقیق‌تری از عملکرد داوطلبان به دست آورد.

با توجه به کاربرد نظریه‌های اندازه‌گیری در تحلیل پس‌آزمون، در این فصل ابتدا کلیات مربوط به نظریه کلاسیک آزمون ارائه می‌شود و در ادامه مروری اجمالی در خصوص نظریه‌های تعمیم‌پذیری و سؤال پاسخ صورت می‌گیرد. برای کسب اطلاعات بیشتر درباره جزئیات عملی مربوط به این نظریه‌ها به بخش هشتم کتاب که به تحلیل آزمون اختصاص یافته است، مراجعه کنید.

نظریه کلاسیک آزمون

نظریه کلاسیک آزمون بر مدل‌های آماری متمرکز است که نمره آزمودنی را مورد توجه قرار می‌دهند. بر طبق اصول نظریه کلاسیک، نمره داوطلب در هر آزمون از مجموع نمره واقعی[1] عملکرد وی و نمره خطا[2] به دست می‌آید. از آنجا که نمره واقعی و خطا به عنوان متغیرهای پنهان[3] و غیرقابل مشاهده در نظر گرفته می‌شوند، از این رو، این نظریه نمرات آزمون را به نمره واقعی ربط می‌دهد. ارتباط مشخصه‌های آماری سؤال با مشخصه‌های آماری آزمون مانند میانگین، انحراف معیار و اعتبار آزمون به خوبی شناخته شده است و از آن‌ها در ساخت آزمون‌ها استفاده می‌شود. اما نکته شایان توجه در این خصوص آن است که این مشخصه‌ها وابسته به نمونه هستند و همین امر ارزش آن‌ها را کاهش می‌دهد. به عبارت دیگر، این مشخصه‌ها تنها زمانی مفید هستند که نمونه، معرف جامعه‌ای باشد که آزمون برای آن ساخته می‌شود. بالتبع اگر نمونه معرف جامعه نباشد، از ارزش این مشخصه‌ها کاسته می‌شود.

در نظریه کلاسیک آزمون این فرض مطرح است که داده‌های مورد استفاده در تحلیل پس‌آزمون شامل منابعی از خطای اندازه‌گیری هستند که می‌توانند بر نمرات مشاهده‌شده داوطلبان تاثیرگذار باشند. خطای آزمون ممکن است در هر یک از مراحل ساخت آزمون، اجرا، نمره‌دهی و تفسیر عملکرد رخ دهد.

در زیر تعدادی از فاکتورهای دخیل در بروز خطاهای مرتبط با عملکرد آزمون‌گران لیست شده است:

- درک ناقص اهداف ارزیابی
- تفسیر نادرست معیارهای ارزیابی
- کاربرد متناقض معیارهای ارزیابی
- تبعیض جنسیتی/ نژادی
- عدم آموزش در خصوص اصول ارزیابی
- تنوع بین آزمون‌گران از حیث تجربه و تخصص
- تاثیر ذهنیت آزمون‌گران در نمره‌دهی

علاوه بر موارد فوق باید گفت که مقداری از تغییرات در نمره داوطلبان به منابع خطای مرتبط با خود آنها برمی‌گردد. البته آنچه که در بروز خطا در اندازه‌گیری تاثیرگذار است، ظرفیت درونی[4] داوطلبان نیست بلکه واکنش‌ها در ارتباط با بیماری، استرس و یا نبود آموزش مناسب و میزان آمادگی آنان برای ارزیابی است. در زیر تعدادی از عوامل ایجاد کننده

1. True score
2. Error score
3. Latent variable
4. Intrinsic capacity

خطا در ارزیابی در ارتباط با عملکرد آزمون شوندگان ارائه شده است:

- استرس
- بیماری
- نبود آموزش مناسب
- آموزش متناقض
- محیط یادگیری نامناسب
- کمبود منابع مناسب
- کمبود فرصت تمرین
- کمبود خواب

تعدادی از عوامل ایجاد کننده خطا در فرایند تهیه و یا تفسیر آزمون و همچنین فرایندهای موثر بر شرایط آزمون دخیل است:

- سؤالات مبهم
- آزمون بسیار طولانی (خستگی) یا بسیار کوتاه
- سؤالات غیرروا
- آزمون با سؤالات غیرهمگن
- آزمون بسیار دشوار یا بسیار آسان
- نبود راهنمای آزمون
- سروصدای زیاد محل آزمون
- گرم بودن یا سرد بودن محل آزمون
- نبود وقت کافی
- سطح روشنایی محل آزمون

بر طبق اصول نظریه کلاسیک به منظور بهبود کیفیت آزمون‌ها، منابع خطای احتمالی باید به حداقل ممکن برسند و یا در صورت امکان حذف شوند. در صورت حذف و یا کاهش خطاهای اندازه‌گیری می‌توان انتظار داشت که نمره مشاهده شده معرف نمره واقعی فراگیران باشد.

به علاوه، در این نظریه میزان خطاهای اندازه‌گیری از طریق محاسبه پایایی آزمون تعیین می‌شود. پایایی آزمون می‌تواند به صورت اندازه‌گیری خطاهای همراه با نمره به دست آمده و یا از طریق ثبات درونی آزمون تفسیر شود. برای کسب اطلاعات بیشتر از رویکردهای متداول در تخمین پایایی آزمون در نظریه کلاسیک، به بخش هشت کتاب مراجعه کنید.

مزایای نظریه کلاسیک

- **آشنایی ارزیابان با مفاهیم پایه‌ای آن:** به طور تقریبی اکثر مقیاس‌هایی که در حال حاضر مورد استفاده قرار می‌گیرند، بر اساس اصول و قوانین این نظریه طراحی شده‌اند. به همین دلیل استادان و ارزیابانی که درگیر مسایل اندازه‌گیری هستند، آشنایی بیشتری با اصول نظریه کلاسیک دارند. از جمله تحلیل‌های آماری به کار گرفته شده در نظریه کلاسیک می‌توان به میانگین، انحراف معیار، ضریب دشواری، ضریب تمیز، آلفای کرونباخ، ضریب پایایی کودر–ریچاردسون، ضریب همبستگی دورشته‌ای–نقطه‌ای و خطای معیار اندازه‌گیری اشاره کرد.
- **قابلیت اجرای بالا:** روش‌های مورد استفاده در این نظریه از قابلیت اجرای بالایی برخوردار هستند. به عنوان مثال

بسیاری از نرم‌افزارهای آماری، قابلیت اجرای تحلیل‌های مربوط به این نظریه را دارند.

- **عدم نیاز به بهینه بودن تک‌تک سؤالات آزمون:** در این نظریه، الزامی بر بهینه بودن تک‌تک سؤالات وجود ندارد. گاهی اوقات طراحی سؤالاتی که قادر باشند به تنهایی و به شکل بهینه، متغیر زیربنایی را اندازه‌گیری کنند کار بسیار دشواری است. از این رو در این نظریه سعی می‌شود تا با افزودن سؤالات مختلف، اکثر خطاهای مربوط به تک تک سؤالات تقلیل یافته و یا حتی در مواردی خنثی گردد. از این رو بر طبق اصول نظریه کلاسیک، در شرایطی که همبستگی بین سؤال‌های آزمون پایین باشد، افزودن سؤالات اضافی می‌تواند به رفع این مشکل کمک کرده و به لحاظ نظری باعث افزایش اعتبار آزمون شود (دی ویلز[1] ۲۰۰۶).
- **کم بودن حجم نمونه مورد نیاز:** در نظریه کلاسیک در مقایسه با نظریه سؤال پاسخ برای تحلیل‌ها، حجم نمونه کمتری مورد نیاز است. اگرچه این موضوع باعث می‌شود که نتایج حاصل از آزمون وابسته به آزمودنی‌ها بوده و در نمونه‌های مختلف متفاوت باشد.

محدودیت‌های نظریه کلاسیک

- **تمرکز بر تحلیل کل آزمون:** از آنجا که تمرکز اصلی نظریه کلاسیک بر تحلیل آزمون و شناسایی خطاهای اندازه‌گیری آن است، به همین دلیل اطلاعات کمی در خصوص عملکرد فردی آزمودنی‌ها در ارتباط با سؤالات آزمون فراهم می‌شود.
- **دشواری طرح تعداد زیادی سؤال در یک آزمون:** از آن جا که طرح تعداد زیادی سؤال برای رسیدن به دقت قابل قبول در آزمون مورد نیاز است، از این رو آزمون‌های ساخته شده در این نظریه دارای سؤالات زیاد و گاهی مشابه هستند. در برخی موارد، تلاش برای طراحی سؤالاتی که با سایر سؤالات آزمون همبستگی بالایی داشته باشند، می‌تواند صرفاً به شباهت‌های ظاهری بین سؤالات منجر گردد. در چنین شرایطی، نه تنها متغیر مورد اندازه‌گیری، بلکه ویژگی‌های نامناسبی از سؤال‌ها از قبیل مشکلات مربوط به ساختار و گرامر زبانی نیز می‌تواند در تمامی سؤالات آزمون رخ دهد. به سخن دیگر، در این حالت نمره واقعی داوطلب در واقع آمیزه‌ای از ویژگی‌های زیربنایی متغیر مورد اندازه‌گیری و ویژگی‌های صوری که مدنظر و هدف آزمون نبوده است خواهد بود که تشخیص آن‌ها از یکدیگر ناممکن است. به همین دلیل، با استفاده از روش‌های نظریه کلاسیک آزمون به سختی می‌توان بین میزان تاثیر ویژگی‌های مربوط به متغیر مورد اندازه‌گیری و ویژگی‌های ظاهری آن‌ها تمایز قایل شد.
- **وابستگی پارامترها به عملکرد آزمودنی‌ها:** در این نظریه، برآورد پارامترها وابسته به تک‌تک افراد مورد مطالعه است. تعیین ضریب دشواری، تمیز و پایایی مبتنی بر همبستگی به دست آمده از نمرات تک تک داوطلبان است. در این حالت اجرای آزمون در نمونه‌های مختلف داوطلبان که دارای واریانس متفاوت است، بدون شک به نتایج یکسان منتهی نخواهد شد. به عبارت دیگر، پارامترهای برآورد شده با استفاده از این نظریه، به گروه آزمودنی که آزمون بر روی آن‌ها اجرا می‌شود، وابسته است. بنابراین ضرایب دشواری و تمیز یک سؤال واحد اگر بر اساس این نظریه محاسبه شوند، از یک گروه به گروه دیگری از داوطلبان تغییر خواهند کرد. به عنوان مثال، در صورتی که داوطلبان از لحاظ متغیر مورد سنجش در سطحی بالایی از توانایی قرار داشته باشند، ضریب دشواری به دست آمده به عدد یک و در حالی که گروه مورد نظر در سطح پایین‌تر از حد متوسط باشند، مقدار این ضریب به صفر نزدیک می‌شود. همان‌طور که پیشتر نیز اشاره شد هر چه ضریب دشواری به عدد یک نزدیک‌تر باشد، این موضوع حاکی از آن است که سؤالات آزمون برای این گروه از داوطلبان ساده است و بالعکس (جهت کسب اطلاعات بیشتر در این خصوص به بخش هشت کتاب مراجعه کنید).

1. DeVellis

- **نمرات مشاهده شده تحت تاثیر سطح دشواری آزمون:** در نظریه کلاسیک، نمرات مشاهده شده با تغییر در سطح دشواری آزمون افزایش یا کاهش می‌یابند. اهمیت این موضوع زمانی احساس می‌شود که قصد داشته باشیم عملکرد داوطلبان را با یکدیگر مقایسه نماییم که فرم‌های متفاوتی از یک آزمون و یا حتی بخش‌های متفاوتی در درون یک آزمون دریافت می‌کنند. در این شرایط قطعاً بحث عادلانه بودن ارزیابی با چالش‌هایی همراه خواهد بود، به این دلیل که نمره مشاهده شده افراد بر حسب دشواری یا سادگی آزمون دستخوش تغییر می‌شود.
- **یکسان بودن خطای اندازه‌گیری برای تمامی آزمودنی‌ها:** نقطه ضعف دیگر نظریه کلاسیک پذیرش یکسان بودن میزان خطای اندازه‌گیری برای تمامی آزمودنی‌هاست. این فرض یکی از مشهورترین و رایج‌ترین جنبه‌های نظریه کلاسیک است. به عنوان مثال، در صورت اجرای یک آزمون دشوار، خطای استاندارد اندازه‌گیری برای آزمودنی‌های ضعیف در مقایسه با آزمودنی‌های قوی، یکسان در نظر گرفته می‌شود در حالی‌که در عمل ممکن است این طور نباشد.
- **دشواری تصمیم‌گیری در مورد عملکرد دانشجویان با نمره یکسان:** اگر نمره آزمودنی‌های مختلف در آزمون یکسان، اما الگوی پاسخ‌دهی آنان به سؤالات آزمون متفاوت از یکدیگر باشد، در این شرایط تصمیم‌گیری در مورد توانایی آن‌ها، از منظر سطح دشواری سؤالات، قطعاً کار بسیار دشواری خواهد بود (توکل و دنیک ۲۰۱۲). به سخن دیگر، از آنجا که نمره یکسان آزمودنی‌ها در یک آزمون به این معنا نیست که آن‌ها به سؤالات مشابهی پاسخ صحیح داده‌اند، با در نظر گرفتن فرض متفاوت بودن سطح دشواری سؤالات مختلف، تصمیم‌گیری در مورد عملکرد آن‌ها غیر ممکن خواهد بود (همبلتون و جونز[1] ۱۹۹۳).
- **دشواری تهیه آزمون‌های موازی:** بر طبق نظریه کلاسیک آزمون، آزمون‌های موازی به آزمون‌هایی اطلاق می‌شود که محتوای یکسانی را می‌سنجند و در آن آزمودنی‌ها نمره واقعی یکسانی کسب می‌کنند. از این رو آخرین نقطه ضعف نظریه کلاسیک آزمون به تعریف آزمون‌های موازی بر می‌گردد. در ارتباط با این تعریف باید اذعان کرد در عمل طراحی دو آزمون به طور کاملاً یکسان دشوار و در اغلب مواقع غیرممکن است و این در حالی است که تخطی از این مفروضه نتایج مربوط به نظریه کلاسیک را زیر سؤال می‌برد. در واقع در این شرایط، طراحی آزمون‌های ناموازی که موازی انگاشته می‌شوند، منجر به برآورد نادرستی از اعتبار آزمون، خطای استاندارد اندازه‌گیری و مدت زمان مورد نیاز برای رسیدن به حد مطلوبی از پایایی آزمون خواهد شد (همبلتون ۱۹۸۹).

نظریه تعمیم‌پذیری

یکی از نظریات مهم و معتبر در سنجش و اندازه‌گیری، نظریه تعمیم‌پذیری است که عمدتاً برای بررسی میزان پایایی اندازه‌گیری‌های رفتاری به کار می‌رود. این نظریه بهترین راه دستیابی به دقت اندازه‌گیری را بیان می‌کند. نظریه تعمیم‌پذیری با استفاده از ترکیب نظریه کلاسیک و روش تحلیل واریانس، سعی در برآورد ضرایب پایایی دارد.

همان‌طور که پیشتر هم اشاره شد، در نظریه کلاسیک، پایایی به ثبات درونی آزمون اشاره دارد. به عنوان مثال، اگر آزمون شوندگان با سؤالات مشابه و تحت شرایط یکسان اما در زمان‌های گوناگون مورد آزمون مجدد قرار گیرند، نتایج بدست آمده باید کم و بیش مشابه باشند. بر طبق نظریه کلاسیک، سؤالات و شرایط ممکن است با خطاهایی همراه باشند که بالتبع بر نتایج آزمون تاثیرگذار است. بنابراین از آنجا که روش‌های تخمین پایایی در نظریه کلاسیک از قبیل آلفای کرونباخ و کودر–ریچاردسون نمی‌توانند منابع بالقوه خطای اندازه‌گیری همراه با سؤالات و شرایط آزمون را از یکدیگر افتراق دهند، قادر نیستند میزان آنها را تخمین بزنند. به عبارت دیگر، نظریه کلاسیک نمی‌تواند در آن واحد تمام منابع خطا

1. Hambleton & Jones

را در نظر بگیرد و سهم هر یک از آن‌ها را در تولید خطای اندازه‌گیری برآورد نماید.

با توجه به موارد فوق نظریه تعمیم‌پذیری یا جی-تئوری در پاسخ به این چالش اساسی توسط کرونباخ و همکاران در سال ۱۹۷۲ ارائه شد و بعدها توسط برنان[1] گسترش داده شد (داگلاس[2] ۲۰۰۶). کرونباخ و همکاران با ارائه این نظریه در تلاش بودند تا با شناسایی، تخمین و تمیز وجوه مختلف آزمون، بتوانند تصویر واضح‌تری از منابع خطای اندازه‌گیری به طراحان آزمون ارائه نمایند. به طور خلاصه، این نظریه که در واقع بسط نظریه کلاسیک است، در شناخت منابع خطا بسیار کارایی دارد (کاردینت و همکاران[3] ۲۰۱۱). در نظریه کلاسیک می‌توان تنها یک منبع خطا را در یک زمان واحد مشخص نمود. این در حالی است که در نظریه تعمیم‌پذیری، میزان هر یک از منابع خطا تخمین زده می‌شود و سپس روشی برای بهینه‌سازی مقدار پایایی فراهم می‌شود (وب و شولسون[4] ۲۰۰۵).

مزایای نظریه تعمیم پذیری

- **مشخص کردن سهم منابع خطای مختلف:** با استفاده از نظریه تعمیم‌پذیری می‌توان منابع خطا را شناسایی نموده و از یکدیگر افتراق داد و حتی میزان تاثیر هر یک از آن‌ها را برآورد کرد. نظریه تعمیم پذیری به ارزیاب کمک می‌کند تا منابع خطای آزمون را به قسمت‌های مختلف تقسیم نماید. این تقسیم منابع خطا می‌تواند به شفاف شدن بهتر واریانس مورد مطالعه کمک نماید. بر همین اساس بر طبق این نظریه، در ابتدا منابع متعددی که تصور می‌شود بر نمره مشاهده شده تاثیرگذار است، بررسی می‌گردد و سپس سهم هر یک در واریانس آزمون مشخص می‌شود.
- **برآورد پایایی و نحوه بهبود آن:** نظریه تعمیم‌پذیری این امکان را برای ارزیابان فراهم می‌سازد که با استفاده از اطلاعات حاصل از آنالیز آماری بتوانند به سؤالات این چنینی پاسخ بدهند: آیا پایایی آزمون را می‌توان با افزایش تعداد عملکردهای مورد سنجش و یا افزایش تعداد قضاوت‌های صورت گرفته بهبود بخشید؟ یا ترکیبی از دو مورد فوق در افزایش پایایی آزمون موثر خواهد بود؟ آیا نمرات حاصل از آزمون به اندازه کافی پایاست که بتوان بر اساس آن در خصوص عملکرد داوطلبان تصمیماتی اتخاذ نمود؟ آیا تغییر در ساختار آزمون از قبیل استفاده از راهنمای آزمون، تدوین روبریک[5] ارزیابی، آموزش مصححان و ... می‌تواند به پایایی بیشتر آزمون بیانجامد؟
- **امکان تعمیم نتایج به سایر شرایط آزمون:** نظریه تعمیم‌پذیری چارچوبی را در اختیار ارزیابان قرار می‌دهد که به کمک آن می‌توانند پایایی آزمون را در موقعیت‌های مختلف آزمون پیش‌بینی کنند. در واقع می‌توان آنچه را که در صورت طراحی مجدد آزمون (مثلا با تعداد سؤال متفاوت) اتفاق می‌افتد، به تصویر کشید. قابلیت‌های این نظریه امکان مطالعه شرایط مختلف و مقایسه پایایی نتایج در شرایط مختلف را فراهم می‌آورد.
- **کاربرد آن در موقعیت‌های مختلف:** موارد کاربرد نظریه تعمیم پذیری بسیار زیاد است و در حوزه‌های مختلف علمی و مخصوصاً در پژوهش‌های مربوط به بررسی و تحلیل ابعاد مختلف اندازه‌گیری‌ها، از این نظریه استفاده می‌شود. جی-تئوری همچنین امکان طراحی و تحلیل آزمون‌های مختلف را فراهم می‌سازد.

محدودیت‌های نظریه تعمیم پذیری

- **اجرایی نبودن استفاده از نظریه تعمیم‌پذیری در شرایط واقعی:** از جمله مشکلات کاربرد جی-تئوری این است که در بررسی شرایط اندازه‌گیری، منابع مختلف خطا باید در ارتباط با هم بررسی و تحلیل شود. این بدان معناست که به عنوان مثال برای سنجش سهم واریانس مربوط به ارزیابان، آزمودنی‌ها و ایستگاه‌ها باید عملکرد کلیه ارزیابان در

1. Brennan
2. Douglas
3. Cardinet et al.
4. Webb & Shavelson
5. Rubric

طول ایستگاه‌های مختلف برای آزمودنی‌ها تحلیل گردد. بی‌شک، محدودیت‌های عملی زیادی برای انجام این امر وجود دارد. به دلیل این محدودیت‌ها، معمولاً از جی-تئوری در مطالعات پایلوت (تعداد معدود ایستگاه‌ها و ارزیابان) استفاده می‌شود و سپس نتایج حاصل به شرایط واقعی ارزیابی تعمیم داده می‌شود (لاوسون[1] ۲۰۰۶).

- **پیچیدگی و کمبود نرم‌افزارهای تحلیل:** این نظریه از دید وب و شیولسون به دو دلیل کمتر مورد استفاده قرار گرفته است (وب و شولسون ۲۰۰۵). یکی به دلیل پیچیدگی آن و دیگری عدم گسترش کافی نرم‌افزارهایی که بتوان از طریق آنها از نظریه تعمیم‌پذیری استفاده کرد.

توضیحات تفصیلی در خصوص کاربرد و جزئیات عملی جی-تئوری در بخش هشتم کتاب مورد بحث قرار خواهد گرفت.

نظریه سؤال-پاسخ

با آنکه بسیاری از متخصصان روانسنجی، شروع حرکت نوین اندازه‌گیری را به دهه ۱۹۶۰ و بر پایه کارهای اولیه جورج راش[2] نسبت می‌دهند، با این حال منشأ اصلی این تحولات به سال‌های پیش از آن بر می‌گردد. به لحاظ تاریخی ریشه‌های اولیه نظریه سؤال پاسخ را می‌توان در مطالعات بینت و سایمون[3] در سال ۱۹۱۶ مشاهده کرد. پیشگامان دیگری مانند ثراندیکو ترستون[4] در دهه ۱۹۲۰، مدل‌هایی را وارد اندازه‌گیری کردند که منطق آن با منطق نظریه کلاسیک تفاوت‌های داشت (دلاور ۱۳۸۹). بعدها افرادی از قبیل ترمن و مریل[5] با طراحی منحنی که در محور افقی آن سن تقویمی آزمودنی‌ها و در محور عمودی آن نسبت پاسخ صحیح قرار داشت، سعی در ترسیم منحنی ارتباط دهنده این دو متغیر داشتند. نمودارهایی که به این صورت به دست آمد چیزی است که ما امروزه منحنی ویژگی سؤال می‌نامیم (بیکر و کیم[6] ۲۰۰۴). به دنبال این موضوع لاولی[7] در سال ۱۹۴۳ مطالعات بسیاری در زمینه تعیین پارامترهای منحنی ویژگی سؤال بر اساس روش بیشینه درست‌نمایی[8] انجام داد. تلاش‌های وی نشان داد که بسیاری از شاخص‌های نظریه کلاسیک را می‌توان بر اساس تابعی از پارامترهای سؤال بازگو نمود. بعد از لاولی، لرد[9] تلاش عظیمی برای گسترش کار وی انجام داد. با مرور زمان مشخص شد که به جای نظریه کلاسیک که به نمره آزمون متکی است، باید نظریه‌ای طراحی کرد که بر ویژگی‌های سؤالات تشکیل دهنده آن متکی باشد و در برآورد پارامترها ثبات داشته باشد. سرانجام نظریه اندازه‌گیری نوین به نام سؤال پاسخ در دهه ۱۹۵۰ به عنوان یک جایگزین برای نظریه کلاسیک معرفی شد. لرد در سال ۱۹۵۲ سعی کرد که در قالب کارهای پژوهشی، کاربرد مفاهیم نظریه سؤال پاسخ برای رسیدن به ثبات در شاخص‌های روانسنجی سؤالات را بررسی نماید (سپاسی ۱۳۸۲). با این حال، نخستین بار بایرن‌بام در سال ۱۹۵۷ مجموعه مقالات فنی در این زمینه نوشت که به معرفی مدل‌های منطقی آزمون و روش‌های برآورد پارامترهای این مدل پرداخت. سپس جورج راش در سال ۱۹۶۰ کتابی در این خصوص منتشر کرد و در نهایت در دهه ۱۹۹۰ با پژوهش‌های لرد و رایت[10] توجه قابل ملاحظه‌ای به نظریه سؤال پاسخ معطوف شد به گونه‌ای که امروزه شاهد کاربرد روزافزون آن در تحلیل آزمون‌ها هستیم (هومن و عسگری ۱۳۸۱). بنابراین نتیجه تلاش‌های گسترده متخصصان علوم روانسنجی، توسعه شکل جدیدی از نظریه اندازه‌گیری به نام نظریه سؤال پاسخ

1. Lawson
2. Georg Rasch
3. Binet & Simon
4. Thorndike & Thurstone
5. Terman & Merill
6. Kim
7. Lawly
8. Maximum Likelihood
9. Lord
10. Right

بود که در طول چندین سال گذشته، همه کارهایی را که در گذشته درباره تهیه آزمون و زمینه‌های روانسنجی وابسته به آن انجام شده بود، تحت‌الشعاع خود قرار داد.

در حال حاضر، نظریه سؤال پاسخ یکی از قدرتمندترین ابزارهایی است که برای تهیه و تجزیه و تحلیل آزمون‌ها به کار می‌رود و آن چنان گسترده و فراگیر شده است که برخی معتقد هستند دوره نظریه کلاسیک و در نتیجه برخی از مفاهیم مرتبط با آن از جمله پایایی نتایج ارزیابی به سرآمده است و استفاده از اصول نظریه سؤال پاسخ جایگزین آن شده است.

نظریه سؤال پاسخ، یک نظریه جامع آماری درباره تاثیر ویژگی‌های سؤالات آزمون، آزمودنی‌ها و چگونگی سنجش توانایی‌هایی است. این نظریه به بررسی این مطلب می‌پردازد که چگونه عملکرد افراد در آزمون و سؤال‌ها به توانایی‌هایی که به وسیله سؤالات آزمون اندازه‌گیری می‌شود، ارتباط پیدا می‌کند. نظریه سؤال پاسخ با استفاده از مدل‌های ریاضی پیچیده‌تر از آنچه در نظریه کلاسیک به کار می‌رود، یک تابع آماری به دست می‌دهد که با استفاده از آن می‌توان احتمال پاسخ درست به یک سؤال را به عنوان تابعی از توانایی[1] آزمودنی و همچنین برخی ویژگی‌های سؤال معرفی کرد. به عبارت دیگر، در نظریه سؤال پاسخ فرض بر این است که به عنوان مثال احتمال پاسخ درست دادن به یک سؤال با افزایش دانش داوطلب افزایش می‌یابد یا خیر (سیف ۱۳۹۴).

به طور کلی نظریه سؤال–پاسخ سعی دارد تا نحوه پاسخ هر آزمودنی را به هر سؤال آزمون ترسیم نماید (همبلتون و همکاران ۱۹۹۱). در ساده‌ترین شکل، نظریه سؤال پاسخ بر پایه این فرض قرار دارد که احتمال آنکه شخص تصادفی *j* با سطح توانایی θ_j به سؤال تصادفی *i* و با سطح دشواری b_i پاسخ صحیح بدهد، بستگی به توانایی او و سطح دشواری سؤال دارد. به عبارت دیگر، اگر فردی از توانایی بالایی در زمینه خاص برخوردار باشد، احتمالاً به سؤال آسان پاسخ صحیح خواهد داد. برعکس آن نیز صدق می‌کند، اگر آزمودنی از توانایی پایین برخوردار است و سؤال نیز دشوار است، احتمالاً این فرد به سؤال پاسخی نادرست خواهد داد.

ویژگی‌های نظریه سؤال–پاسخ

- برخلاف نظریه کلاسیک که در آن نمره خام آزمون‌شونده از مجموع نمرات در پاسخ به همه سؤال‌های آزمون به دست می‌آید، در نظریه سؤال پاسخ به جای محاسبه نمره خام آزمون، عملکرد داوطلب در هر سؤال به شکل منفرد محاسبه و بررسی می‌گردد. دلیل این مطلب آن است که مفاهیم نظریه سؤال پاسخ مبتنی بر تک تک سؤال‌های آزمون است تا بر پایه مجموعه‌ای از پاسخ‌ها که به عنوان نمره یک آزمون به کار می‌رود.
- بر خلاف نظریه کلاسیک، نمره آزمون‌شونده در نظریه سؤال–پاسخ برابر با مجموع پاسخ‌های درست او نیست، بلکه هم تابع تعداد پاسخ‌های آزمون‌شونده و هم تابع ویژگی‌های سؤال‌های آزمون است (هوگان[2] ۲۰۰۷). به عنوان مثال در صورتی که در یک آزمون، دو آزمون‌شونده به تعداد مساوی به سؤالات درست پاسخ داده باشند، اما یکی از آن‌ها به سؤال‌های آسان و دیگری به سؤالات دشوار پاسخ صحیح داده باشند، نمره آن‌ها با یکدیگر مساوی نخواهد بود. بر طبق نظریه سؤال– پاسخ داوطلبی که به سؤال‌های دشوارتر پاسخ داده است، نمره بالاتری خواهد گرفت.
- در نظریه سؤال– پاسخ هدف اصلی اجرای آزمون روی آزمودنی‌ها، تعیین موقعیت آن‌ها در مقیاس توانایی است. بر طبق این نظریه، از طریق تعیین توانایی هر داوطلب دو هدف حاصل می‌گردد. اول آن که می‌توان داوطلبان را بر حسب این‌که دارای چه میزان از توانایی مورد انتظار هستند ارزیابی نماییم. دوم آن که از این طریق می‌توان عملکرد داوطلبان را با یکدیگر مقایسه نمود.
- شاید مهم‌ترین تمایز بین نظریه‌های کلاسیک و نظریه‌های جدید آزمون، ثبات پارامترهای سؤال و توانایی در

1. Ability
2. Hogan

نظریه سؤال پاسخ باشد. این ویژگی منجر به وابسته نبودن پارامترهای افراد به سؤال‌های موجود در آزمون و وابسته نبودن پارامترهای سؤال‌ها به نحوه توزیع توانایی آزمودنی‌ها می‌شود. بنابراین در نظریه سؤال پاسخ بر خلاف نظریه کلاسیک، اطلاعات مربوط به شاخص‌های سؤال مانند ضریب‌های دشواری و تمیز در نمونه‌های مختلف آزمون‌شوندگان یکسان خواهد بود. دلیل این امر آن است که الگوی ریاضی مورد استفاده برای محاسبه پارامترهای سؤال در نظریه سؤال پاسخ بر اساس توانایی آنان به دست می‌آید نه از روی نمره کل آزمون شوندگان (سیف ۱۳۹۴). بنابراین در صورت اجرای آزمون در دو گروه متفاوت، اطلاعات حاصل از یک نمونه از آزمون شوندگان با اطلاعات حاصل از نمونه دیگر آزمون شوندگان صرف نظر از میانگین توانایی گروه‌های آزمون شونده برابر خواهد بود.

- در نظریه سؤال-پاسخ، از تابع آگاهی[1] آزمون (که در فصل هشت مفصلاً مورد بحث قرار خواهد گرفت) به منظور محاسبه خطای اندازه‌گیری و برآورد پایایی آزمون استفاده می‌شود. آگاهی آزمون، تابعی از میزان توانایی آزمودنی براساس سؤالات آزمون است؛ بنابراین آگاهی آزمون براساس سطوح مختلف توانایی، تغییر می کند.

انواع مدل‌های نظریه سؤال- پاسخ

با نگاهی به منابع موجود به سادگی می‌توان دریافت که طبقه‌بندی‌های متفاوتی از مدل‌های سؤال پاسخ ارائه شده است. در ادامه به صورت مختصر به انواع این مدل‌ها اشاره خواهیم کرد.

- **دسته‌بندی بر اساس ابعاد اندازه‌گیری:** مدل‌های سؤال-پاسخ را می‌توان به دو خانواده مدل‌های تک‌بعدی[2] و چندبعدی[3] تقسیم کرد (تیسن و اورلاندو[4] ۲۰۰۱). در مدل‌های اولیه نظریه سؤال پاسخ، چنین فرض می‌شد که عامل زیربنایی عملکرد آزمودنی در تست یا آزمون، تک بُعدی است. بدین معنی که تنها یک توانایی می‌توان برای توصیف روابط بین سؤال‌ها و عملکرد فرد در آزمون به کار برد. اما با توسعه این نظریه و نیز سهولت پیاده‌سازی مدل‌های پیچیده‌تر، مدل‌ها مبتنی بر عوامل زیربنایی متعدد (چندبعدی) برای تحلیل آزمون‌ها نیز ارائه شدند که تحت عنوان نظریه سؤال-پاسخ چندبُعدی شناخته می‌شوند. به عبارت دیگر در مدل‌های چندبُعدی فرض بر آن است که داده‌ها از خصایص چندوجهی یا چندگانه به دست می‌آیند. در هر حال، به دلیل پیچیدگی فزاینده مدل‌های چندبُعدی، در عمل بیشتر از مدل‌های تک‌بُعدی استفاده می‌شود.
- **دسته‌بندی بر اساس نحوه نمره‌گذاری سؤالات:** مدل‌های نظریه سؤال- پاسخ بر پایه نحوه نمره‌گذاری سؤالات نیز طبقه‌بندی می‌شوند. یعنی بر اساس این که سؤالات به صورت پاسخ‌های صحیح-غلط ارزش‌گذاری شوند یا طیفی از پاسخ‌های صحیح. به عنوان مثال، یک سؤال چندگزینه‌ای در واقع یک مدل دوارزشی[5] است، حتی اگر دارای چهار یا پنج گزینه باشد، زیرا فقط به‌گونه صفر و یک (غلط-صحیح) نمره‌گذاری می‌شود. انواع دیگر مدل‌ها، برای داده‌های چندارزشی[6] به کار می‌روند که در آن مقدار نمره هر پاسخ متفاوت در نظر گرفته می‌شود. به عنوان مثال، آزمون ویژگی‌های کلیدی می‌تواند در مدل چند ارزشی جای گیرد، در صورتی که برای هر یک از پاسخ‌های سؤال، وزن نمره متفاوتی مشخص شود (برای کسب اطلاعات بیشتر در این خصوص به بخش چهارم، آزمون‌های استدلال بالینی مراجعه کنید).
- **دسته‌بندی بر اساس تعداد پارامترهای برآورده شده سؤال:** مدل‌های سؤال-پاسخ بر اساس تعداد پارامترهای برآوردشده نیز دسته‌بندی می‌شوند. در اینجا توجه به این نکته ضروری است که در تمامی مدل‌های نظریه سؤال-

1. Information function
2. Unidimensional
3. Multidimensional
4. Thissen & Orlando
5. Dicotomous
6. Polytomous

پاسخ، پارامتر توانایی آزمودنی و در نتیجه آن پارامتر دشواری برآورد می‌شوند اما تخمین پارامتر تمیز و احتمال حدس زدن به ترتیب مبنای مدل دو و سه پارامتری هستند. بدین ترتیب، در مدل تک پارامتری مقدار حدس و قدرت تمیز سؤال‌ها برابر فرض می‌شوند و تنها پارامتر دشواری برآورد می‌شود درحالی‌که در مدل دو پارامتری فرض بر این است که در سؤالات احتمال حدس زدن وجود ندارد، اما بر اساس جایگاه و قدرت تمیز سؤال از یکدیگر متغیر هستند. سرانجام در مدل سه پارامتری، هر سه پارامتر دشواری، ضریب تمیز و حدس لحاظ می‌شود. در همین رابطه باید گفت که پاره‌ای از مواقع مشاهده می‌شود که حتی آزمودنی‌های خوب به برخی از سؤالات خیلی ساده پاسخ غلط می‌دهند. این امر ممکن است از بی دقتی آنان یا وجود اطلاعاتی ورای آنچه که مورد نظر طراح سؤال بوده است، ناشی شود. جهت حل این مشکل برخی از روانسنجان از جمله مک دونالد[1] (۱۹۶۷) و بعدها بارتن[2] و لرد (۱۹۸۱) مدلی ارائه دادند که در برآورد توانائی آزمودنی‌ها این عامل را در نظر می‌گیرد و بنابراین به مدل منطقی چهار پارامتری شهرت یافته است (همبلتون ۱۹۸۹). هر چند در عمل، این مدل تنها از لحاظ نظری جالب و مورد توجه قرار گرفت، زیرا که پیشنهاد کنندگان این مدل نتوانسته‌اند فواید مترتب بر این مدل را تبیین کنند.

انتخاب مناسب‌ترین مدل

یکی از سؤالات چالش‌برانگیز در نظریه سؤال پاسخ آن است که بهترین مدل برای تحلیل داده‌ها چیست؟ پاسخ به این سؤال نیازمند در نظر گرفتن فاکتورهای متعددی است که در اینجا به چند مورد اشاره می‌شود:

- اگر سؤالات قدرت تمیز برابری نداشته باشند، بهتر است که از مدل‌های دو و سه پارامتری به جای مدل تک پارامتری استفاده شود.
- اگر نوع سؤالات آزمون طوری است که احتمال حدس زدن پاسخ توسط آزمودنی‌ها محتمل است، بهتر ست مدل سه پارامتری استفاده شود. برعکس، در سؤالات تشریحی و آزمون‌های سنجش نگرش، عامل حدس چندان نمی‌تواند تاثیر داشته باشد. بنابراین استفاده از مدل سه پارامتری برای این سؤالات معمول نیست.
- مسأله دیگر که در بررسی تاثیر پارامتر حدس زدن در پاسخ دادن به سؤالات کمک می‌کند، این است که آزمودنی‌ها با کمترین توانایی تا چه میزان به سؤالات بسیار دشوار در آزمون پاسخ داده‌اند. اگر آن‌ها به سؤالات بسیار دشوار پاسخ نداده باشند، می‌توان از پارامتر حدس چشم‌پوشی نمود. اما اگر این دسته از دانشجویان به سؤالات بسیار دشوار، بیش از حدانتظار پاسخ داده باشند، باید پارامتر حدس در مدل گنجانده شود که در این صورت مدل سه پارامتری بر مدل‌های یک و دو پارامتری ارجحیت دارد.
- روش دیگر برای انتخاب بهترین مدل استفاده از تابع آگاهی است. بر اساس این روش، مدلی بهترین برآورد از توانایی افراد را ارائه خواهند نمود که در نقطه میانگین توزیع توانایی افراد، بیشترین آگاهی و کمترین خطا را داشته باشد. برای انتخاب بهترین مدل متناسب با داده‌های مشاهده شده (برازش مدل) می‌توان از مقایسه سؤالات نیز استفاده نمود.

مزایای نظریه سؤال- پاسخ

- **مستقل بودن پارامتر توانایی از آزمون:** نظریه کلاسیک و جی-تئوری بر تحلیل آزمون و منابع بالقوه خطای اندازه‌گیری متمرکز می‌شوند و اطلاعات مختصری در مورد توانایی و ظرفیت‌های آزمون شوندگان با آزمون و سؤالات آن ارائه می‌دهند (ریکاو و مارکولایدس[3] ۲۰۱۱). برخلاف آن، در نظریه سؤال پاسخ هدف اصلی، اندازه‌گیری ارتباط بین توانایی آزمون‌شوندگان و سطح دشواری و تمیز سؤالات به صورت مستقل از یکدیگر است.

1. McDonald
2. Barton
3. Raykov & Marcoulides

- **مســـتقل بودن پارامترهای ســـؤال از نمونه آزمودنی‌ها:** از آنجا که در نظریه ســـؤال پاســـخ، ارتباط هر یک از پارامترهای آزمون با توجه به ســـطوح مختلف توانایی آزمون شـــوندگان تعیین می‌شـــود، محاسبه پارامترهای سؤال مســـتقل از نمونه‌های گوناگون فراگیـــران خواهد بود. در این حالت در صورت اجـــرای آزمون در نمونه‌های مختلف داوطلبان، پارامترهای برآورده شـــده از یک سؤال، مشابه با ســـطح توانایی مورد انتظار خواهد بود. از این رو، چنانچه طراح آزمون بخواهد آزمونی تدوین کند که با ویژگی‌های جامعه آزمودنی مورد مطالعه ارتباط نداشـــته باشد، این کار تنها از طریق نظریه سؤال پاسخ امکان پذیر است (همبلتون و همکاران ۱۹۸۹). به عنوان مثال در صورتی که قصد داریم آزمونی طراحی کنیم که بتوانیم بر اساس نتایج حاصل از آن، دانشجویان ضعیف را برای شرکت در یک برنامه اصلاحی شناســـایی کنیم و یا اگر قصدمان تعیین دانشـــجویان برتر به منظور اعطای بورس تحصیلی اســـت، نیاز به طراحی چنین آزمون‌هایی وجود دارد.

محدودیت‌های نظریه سؤال- پاسخ

- **پیچیدگی مدل‌های نظریه سؤال- پاسخ:** از نظر عملی، بدون توجه به کاربرد، سؤال‌های فنی پیچیده‌تری در این نظریه در مقایسه با مدل‌های کلاسیک مطرح می‌شود. بنابراین می‌توان گفت از لحاظ فنی، مدل‌های نظریه سؤال پاسخ پیچیده هستند (توکل و دنیک ۲۰۱۲). پیچیدگی محاسبات ریاضی زیربنایی نظریه سؤال پاسخ، یکی از موانع عمده‌ای است که کاربرد این نظریه را با مشکل روبرو ساخته است. هر چند استفاده از نرم‌افزارهای پیشرفته آماری امروزی، انجام این محاسبات را امکان پذیر نموده است اما دشواری درک و یادگیری این نظریه هنوز به قوت خود باقی است.
- **مشکلات برازش مدل:** مشکل دیگر، برازش[1] مدل است. تمام مزایای نظریه سؤال پاسخ در صورتی حاصل می‌شود که مدل با داده‌ها برازش داشته باشد، مسأله‌ای که هنوز جواب روشنی برای آن وجود ندارد. مقصود از برازش آن است که داده‌های حاصل از اندازه‌گیری و ارزیابی تا چه اندازه با مدل‌های موجود در نظریه سؤال-پاسخ هماهنگی و مطابقت دارد. روش‌های گوناگونی برای برازش داده‌های تجربی با مدل وجود دارد اما هنوز به طور کامل روشن نیست که چگونه باید مشکلات برازش مدل را حل نمود، به ویژه مشکلاتی که مربوط به ابعاد مختلف آزمون است.
- **نیاز به تعداد زیاد آزمودنی:** یکی از موضوعات و نکات مهم در نظریه سؤال پاسخ، حجم نمونه آزمودنی‌ها و حجم نمونه سؤال است. این دو عامل به ویژه در مدل سه پارامتری می‌توانند بر برآورد پارامترهای سؤال و توانایی تاثیرات جدی داشته باشد. لرد طی یک مطالعه و ضمن مقایسه مدل‌های یک و دوپارامتری در برآورد نمره حقیقی آزمودنی‌ها نشان داد وقتی حجم نمونه کوچک باشد، پارامتر قدرت تمیز سؤال‌ها و حدس زدن سؤال‌ها را نمی‌توان به‌دقت تعیین کرد (لرد ۱۹۸۰). از این رو، در بعضی موقعیت‌های معین، محدود و با حجم نمونه کوچک‌تر از ۱۰۰ یا ۲۰۰ آزمودنی، برآورد نمره حقیقی در مدل راش (تک پارامتری) می‌تواند اندکی بهتر از مدل دو پارامتری باشد (حبیبی و همکاران ۱۳۹۱).

مقایسه کاربرد و جایگاه نظریه‌های اندازه‌گیری در آموزش پزشکی

کاربرد هر یک از نظریه‌های اندازه‌گیری به آزمون مورد نظر، هدف و موقعیت ارزیابی وابسته است. در ادامه نگاه اجمالی به شماری از قواعد مرتبط با کاربرد نظریه‌های اندازه‌گیری خواهیم داشت (لامبرت و شوورث ۲۰۱۱):

- استفاده از نظریه کلاسیک آزمون در آزمون‌هایی مانند آزمون‌های چندگزینه‌ای یا بازپاسخ مفید است. در نظریه

1. Fitness

کلاسیک آزمون، محاسبه پارامترهای مربوط به سؤال از قبیل شاخص تمیز و دشواری به سادگی از طریق اکثر نرم‌افزارهای آماری امکان‌پذیر است. به علاوه، تفسیر پارامترهای مربوط به سؤال آسان و قابل آموزش می‌باشد. شاخص پایایی به عنوان مثال آلفای کرونباخ، بر حسب همبستگی بین آزمون–بازآزمون تعیین می‌شود. بنابراین استفاده از اصول نظریه کلاسیک آزمون برای تخمین پایایی در آزمون‌های هنجار محور مناسب‌تر است.

- نظریه تعمیم‌پذیری از انعطاف‌پذیری بالایی برخوردار است و امکان حذف یا در نظر گرفتن منابع واریانس را در محاسبه فراهم می‌سازد. هرچند این کار مستلزم، شناخت کافی محقق از منابع مختلف واریانس و نحوه تعامل آن‌هاست. به علاوه، نرم‌افزارهای ابتدایی طراحی شده برای محاسبه پارامترهای سؤال خیلی کاربرپسند نبودند. به همین دلیل سال‌ها بعد، دو محقق در دانشگاه مک‌مستر[1] نرم‌افزاری جهت کاربری آسان‌تر محاسبات مربوط به این نظریه ارائه کردند. به طور کلی، نظریه تعمیم‌پذیری در مقایسه با نظریه کلاسیک از مزایای بیشتری برخوردار است، چرا که در نظریه کلاسیک پارامترهای گوناگون باید در ترکیب با یکدیگر محاسبه شوند (به عنوان مثال محاسبه آلفای کرونباخ، ضریب کاپا یا ICC برای محاسبه توافق بین مشاهده‌گران).
- نظریه سؤال–پاسخ تنها به وسیله افرادی که با اصول و مفاهیم کلی این نظریه آشنایی دارند قابل استفاده است. به علاوه، استفاده از بانک سؤالات و اجرای آن بر روی تعداد قابل ملاحظه‌ای از آزمون‌شوندگان پیش از اجرای آزمون اصلی مورد نیاز است. مجموع موارد فوق، منجر به کاربرد محدود این نظریه در عمل شده است.

1. McMaster

منابع

1. Baker FB, Kim SH. Item response theory: Parameter estimation techniques. 2nd ed. London, UK: Taylor & Francis; 2004.
2. Bock RD. A brief history of item theory response. Educational Measurement: Issues and Practice 1997; 16(4): 21-33.
3. Cardinet J, Johnson S, Pini G. Applying generalizability theory using EduG: Taylor & Francis; 2011.
4. Courville TG. An empirical comparison of item response theory and classical test theory item/ person statistics. Texas, US: Texas A&M University. 2004.
5. DeVellis RF. Classical test theory. Med Care 2006; 44(11 Suppl 3): S50-S59
6. Douglas ML. Applying generalizability objective structured clinical examination in a naturalistic environment. Joumal of Manipulative and Physiological Therapeutics6)29;2006):463-7.
7. Hambleton RK. Principles and selected application of item response theory. In: Linn RL, American Council on Education, editors. Educational measurement. 3rd ed. Washington, DC: American Council on Education; 1989.
8. Hambleton RK, Vander Linden, Wim J. Advance in Item Response Theory and Applications: An Introduction, Applied Psychplogical Measurement 1982;6(4), 373- 378.
9. Hambleton R, Jones R. Comparison of classical test theory and item response theory and their applications to test development. Educ Meas Issues Pract 1993; 12:38–47.
10. Hambleton RK, Swaminathan H, Rogers HJ. Fundamentals of item response theory. Newbury Park, CA: Sage. 1991.
11. Hogan T. Psychological testing: a practical introduction: John Wiley 2007.
12. Lawson DM. Applying generalizability theory to high-stakes objective structured clinical examinations in a naturalistic environment. Journal of manipulative and physiological therapeutics. 2006;29(6):463-7.
13. Lord FM. Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum. 1980.
14. Lord FM, Novick MR. Statistical Theories of mental test scores. Reading, MA: Addison-Wesley 1968.
15. Magno C. Demonstrating the difference between classical test theory and item response theory using derived test data. The International Journal of Educational and Psychological Assessment. 2009;1(1):1-11.
16. Raykova T, George A. Macrolides. Classical Item Analysis Using Latent Variable Modeling: A Note on a Direct Evaluation Procedure. Structural Equation Modeling: A Multidisciplinary Journal 2011;18(2):315-324

17. Schuwirth LW, van der Vleuten CP. General overview of the theories used in assessment: AMEE Guide No. 57. Medical teacher. 2011;33(10):783-97.
18. Tavakol M, Dennick R. Post-examination analysis of objective tests. Medical teacher. 2011;33(6):447-58.
19. Tavakol M, Dennick R. Post-examination interpretion of objective test data: Monitoring and improving the quality of high-stakes examinations – A commentary on two AMEE Guides. Med Teach 2012, 34(3):245–248.
20. Tavakol M, Dennick R. Post-e`xamination interpretion of objective test data: Monitoring and improving the quality of high-stakes examinations: AMEE Guide No 66. Med Teach 2012; 34: e161–e175.
21. Thissen D. Orlando M. (2001).Item response theory scored in two categories.In D.Thissen Wainer H. (Eds.), Test scoring (pp.73-140).Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
22. Webb NM, Shavelson RL. Generalizability Theory: Overview. Encyclopedia of Statistics in Behavioral Science 2005; 2:717–719

۲۳. حبیبی م، خدایی ا، ایزانلو م. نظریه های قدیم و جدید اندازه گیری در علوم رفتاری و پزشکی: مروری بر روش شناسی، مزایا و تنگناها، ۱۳۹۱؛ ۱۰(۴):۳۰۲–۳۱۵.
۲۴. دلاور ع. مقدمه ای بر نظریه‌های اندازه گیری (روانسنجی)، ۱۳۸۹.
۲۵. سپاسی ح. حسن س. مقایسه مفاهیم و مفروضه های نظریه کلاسیک و نظریه جدید سؤال–پاسخ در ساخت آزمونهای روانی و تربیتی، مجله علوم تربیتی و روانشناسی، ۱۳۸۲؛ ۱۰(۳–۴): ۱۷–۳۸
۲۶. سیف ع ا. اندازه گیری، سنجش و ارزشیابی آموزشی. نشر دوران، ۱۳۹۴.
۲۷. هومن ح، عسگری ع. پایه‌های اساسی نظریه سؤال– پاسخ (نظریه جدید روانسنجی)، انتشارات پیک فرهنگ، ۱۳۸۱.

۲

آزمون‌های کتبی بسته پاسخ

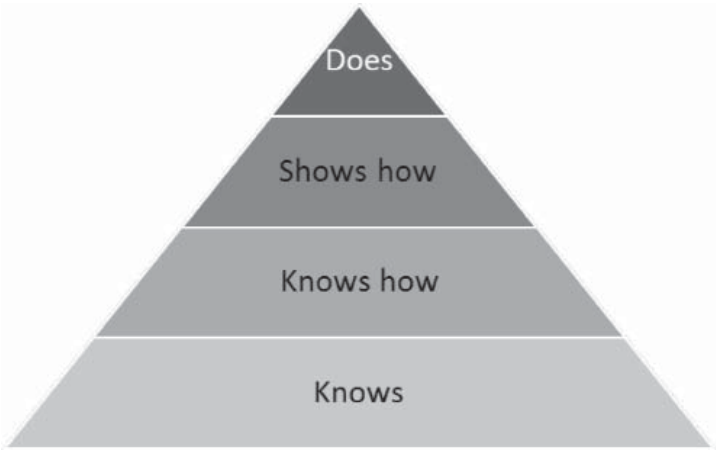
فصل | ۵ |

خانواده سؤالات بسته پاسخ

تاریخچه سؤالات بسته‌پاسخ

آنچه امروزه به عنوان سؤال تستی یا چندگزینه‌ای می‌شناسیم، در واقع جزئی از خانواده بزرگتر سؤالات کتبی است که برای ارزیابی حیطه شناختی[1] فراگیران به کار می‌روند. در مدل میلر دو سطح پایین هرم توسط سؤالات کتبی قابل سنجش هستند (شکل ۱–۵).

سؤالات کتبی به صورت کلی به دو دسته بازپاسخ[2] و بسته‌پاسخ[3] تقسیم می‌شوند. برای پاسخگویی به سؤالات بازپاسخ، خود دانشجو باید جواب را به صورت کلمات، عبارات و یا جملاتی تولید کند و در برگه پاسخنامه بنویسد. در حالی که در سؤالات بسته‌پاسخ، دانشجو فهرستی از گزینه‌ها را مشاهده می‌کند و باید جواب صحیح را از میان آنها تشخیص دهد.



شکل ۱–۵: هرم میلر

اســتفاده از سؤالات چندگزینه‌ای در حدود صد سال قبل آغاز شد. گفته می‌شود اولین بار یک روانشناس آمریکایی به نام

1. Cognitive
2. Constructed response, Uncued question, Open-ended question
3. Selected response, Cued questions, Closed-ended question

ادوارد تورندیک[1] که شـــهرت اصلی او در ارائه یک نظریه یادگیری بود، ســـؤال چندگزینه‌ای را ابداع کرد. با این حال، اولین مستندات در خصوص استفاده از سؤالات چندگزینه‌ای در امتحانات سطح بالا و مهم به فردریک کلی[2] نسبت داده می‌شود. او در حیطه ارزیابی فراگیران با دو مشکل مواجه شده بود. اولاً فرایند تصحیح برگه‌های امتحانی برای معلمان وقت‌گیر بود و ثانیاً این فرایند پرزحمت، بسیار وابسته به ذهنیت[3] معلم بود و منجر به نمرات قطعی و یکسان نمی‌شد. از نظر او، تدریس و آموزش زمانی اثربخش قلمداد می‌شـــدند که فراگیران عملکردی واحد در حد اســـتاندارد از خود نشان می‌دادند. در واقع، کِلی در عصر ماشـــین زندگی می‌کرد که استانداردسازی، یکسان‌سازی، کارایی و بازدهی بر مسائلی مانند تنوع، خلاقیت و فردیت ارجحیت داشتند. وی برای رفع مشکلات آزمون و در جهت کاهش پراکندگی پاسخ‌ها و عینیت بخشیدن به فرایند ارزیابی، برای سؤالات «گزینه»[4] در نظر گرفت. جالب اســـت سؤالاتی که او در آن زمان طراحی کرده بود، از لحاظ ظاهری با نوعی که امروزه تحت عنوان ســـؤال چندگزینه‌ای می‌شناســـیم، تفاوت چندانی ندارند. کلی در ســـال ۱۹۱۴ هنگامی که به عنوان رییس دانشکده آموزش کانزاس[5] مشغول به کار بود، آزمونی به نام Kansas Silent Reading Test را بنا نهاد. این آزمون از انواع اولیه تست‌های هوش به شمار می‌رفت که در آن از سؤالات چندگزینه‌ای استفاده شده بود.

طی سال‌های بعد، این شکل سؤال به سرعت در سطوح و رشته‌های مختلف مورد استفاده قرار گرفت. یکی از موارد دیگری که کاربرد سؤالات چندگزینه‌ای را وسعت بخشید، اجرای تست‌های هوش طی جنگ جهانی اول در ارتش امریکا بود. دانش آموخته دانشگاه استانفورد، آرتور اوتیس[6] که کار او تا حد زیادی تحت تاثیر روش فردریک کلی بود، به عنوان یکی از

پیشگامان استفاده از سؤالات تستی در این وادی شناخته می‌شود. پس از اتمام جنگ جهانی اول لازم بود که تعداد زیادی از افراد تحت آموزش و ارزیابی قرار بگیرند. از یک طرف دولت تنها دو سال تحصیل در دبیرستان را به صورت اجباری تعیین کرده بود که موجب افزایش تعداد فارغ‌التحصیلان شده بود و از طرف دیگر، کارخانه‌ها به دنبال جذب نیروی کار بودند. در نتیجه به یک روش سریع و موثر نیاز بود تا در مورد سطح دانش و مهارت تعداد زیادی از افراد به سرعت تصمیم‌گیری شود تا بتوان آنها را تقسیم کرد و در شغل و جایگاه مناسبی قرار داد.

درباره تاریخچه استفاده از سؤالات چندگزینه‌ای در موسسات آموزشی، می‌توان به سال ۱۹۲۶ اشاره کرد که در آزمون ورودی کالج، سؤالات تستی در آزمون استعداد تحصیلی[7] مورد استفاده قرار گرفتند. بعدها، در سال ۱۹۴۸ برای اولین بار آزمون پذیرش دانشکده حقوق[8] برگزار شد که شامل ۱۰ بخش بود و یک روز طول می‌کشید. قسمت عمده‌ای از این آزمون را سؤالات چندگزینه‌ای تشکیل می‌داد. با وجود این که طی سال‌ها تغییرات زیادی در این امتحان‌ها صورت گرفته است، همچنان بخش عمده‌ای از آن ها شامل سؤالات چندگزینه‌ای است.

با وجود همه این موارد که نشان دهنده اقبال عمومی نسبت به سؤال چندگزینه‌ای است، داستان آنجا به اوج می‌رسد که به صورت جسته و گریخته انتقاداتی به این شکل سؤال وارد شد. در واقع، تأکید زیاد بر سنجش محفوظات دانشجویان در کنار جو آموزشی معلم–محور مدارس، عملاً آموزش را به سمت و سویی برده بود که انگیزه و عملکرد فراگیران را تحت تاثیر قرار می‌داد. جالب اینجاست که خود فردریک کلی در سال ۱۹۲۸، زمانی که رییس دانشگاه ایداهو[9] بود، مقاله‌ای نوشت و در آن علاوه بر انتقاد از سیستم آموزشی، اعلام کرد سؤالات چندگزینه‌ای محدودیت‌هایی هم دارند. او معتقد بود که با این روش فقط قسمت مختصری از مطالبی را که آموزش داده شده است، می‌توان ارزیابی کرد و از آنجا که مواردی مانند تفکر، خلاقیت، توانایی حل مسأله و درک مطالب فراگیران با این روش سنجیده نمی‌شوند، مسایلی از این دست در سیستم آموزشی

1. Edward L.thorndike
2. Frederick J. Kelly
3. Subjectivity
4. Option
5. College of Education-University of Kansas
6. Arthur S. Otis
7. Scholastic Aptitude Tests (SAT)
8. Law School Admission Test (LSAT)
9. Idaho

مورد غفلت قرار می‌گیرند. او در ادامه، طرحی برای بازنگری داد که از جمله در آن توصیه کرد استفاده از سؤالات چندگزینه‌ای محدود شود و مواردی مانند تفکر نقاد مورد توجه قرار گیرند. متاسفانه در آن زمان گروه‌های مختلف انتقادات او را نپذیرفتند، معلمان دانشکده با نظرات او مخالفت کردند و در نهایت در سال ۱۹۳۰ از او خواسته شد تا استعفا دهد.

در هر حال، استفاده از سؤالات چندگزینه‌ای به عنوان یک روش ارزیابی همچنان ادامه پیدا کرد. ورود این شکل سؤال به حوزه امتحانات علوم پزشکی به حدود سال ۱۹۵۰ برمی‌گردد. قبل از آن به صورت معمول برای ارزیابی فراگیران از آزمون‌های تشریحی و شفاهی استفاده می‌شد که همیشه درجاتی از ذهنی بودن به همراه داشتند. در مقابل، سؤالات چندگزینه‌ای یک جواب مشخص داشتند و به علت عینی بودن مورد استقبال قرار گرفتند. مزیت مهم دیگری که به گسترش روزافزون این دسته از سؤالات کمک کرد، امکان گرفتن امتحان از تعداد زیاد دانشجویان در فرصت کم و نیز پوشش دادن محتوای زیادی از دوره بود. همچنین، امکان تصحیح آسان این سؤالات به کمک ماشین و تحلیل ساده نمرات باعث شد که آزمون‌های چندگزینه‌ای به سرعت در رشته‌ها و مقاطع مختلف تحصیلی مورد استفاده قرار گیرند.

در صفحات بعدی به توضیح شکل‌های مختلف سؤالات بسته‌پاسخ خوهیم پرداخت اما در اینجا به ذکر همین نکته بسنده می‌کنیم که در گذشته، طیف وسیعی از انواع آن مورد استفاده قرار می‌گرفتند. در حالی که امروزه این گونه نیست و اَشکال محدودی از این گروه استفاده گسترده و عام دارند. به عنوان مثال، در خصوص به کارگیری اشکال مختلف سؤالات بسته‌پاسخ در آزمون گواهینامه[1] پزشکی امریکا در سال ۱۹۸۵ ذکر شده است که دلیل این کار صرفاً ایجاد تنوع در یک آزمون طولانی بود. اما ۲۵ سال پس از اینکه این آزمون از آزمونی تشریحی به آزمونی با انواع مختلف سؤالات بسته‌پاسخ تبدیل شده بود، بورد ملی ارزیابان پزشکی[2] با بررسی مطالعات انجام شده، تصمیم گرفت تنوع سؤالات را کاهش دهد. در آن هنگام مقرر شد فقط هفت نوع سؤال (انواع A، B، C، G، K، X و M) طرح شوند. اواسط دهه ۱۹۸۰ مجدداً پژوهش‌ها مورد بررسی قرار گرفتند و توافق حاصل شد که ۴ نوع سؤال (انواع A، B، C و K) کافی هستند. اخیراً با بررسی‌های مجدد، انواع سؤالات این امتحان تنها به دو نوع (چندگزینه‌ای با بهترین پاسخ و جورکردنی گسترده) محدود شده است (کیس و سوانسون[3] ۲۰۰۲).

همچنین امروزه در انگلیس در آزمون‌های عضویت کالج‌های سلطنتی پزشکان[4]، عضویت کالج سلطنتی پزشکان عمومی[5] و بورد ارزیابی حرفه‌ای و زبان‌شناسی[6] که سابقاً شامل سؤالات «درست–نادرست متعدد» بودند، از سؤالات «پنج‌گزینه‌ای با بهترین پاسخ» یا «جورکردنی گسترده» استفاده می‌شود (اندرسون[7] ۲۰۰۴).

انواع سؤالات بسته‌پاسخ

از زمان معرفی سؤالات بسته‌پاسخ تاکنون، تغییرات زیادی در شکل ساختاری آنها پیشنهاد شده است. هرچند در حال حاضر، رایج ترین نوع سؤال بسته‌پاسخ، سؤال «چندگزینه‌ای با بهترین پاسخ»[8] می‌باشد، اَشکال متنوعی از آنها با نام‌های مختلف در دسترس هستند که البته بیشتر آنها امروزه توصیه نمی‌شوند و چندان مورد استفاده قرار نمی‌گیرند. برخی از آنها با حروف الفبا شناخته شده‌اند اما این حروف، مؤید مفهوم خاص یا خلاصه شده کلمه ویژه‌ای نیستند. برای کامل شدن بحث در اینجا به طور مختصر در مورد هر یک توضیحاتی ارائه می دهیم و در فصول بعدی، به ذکر جزئیات بیشتری در خصوص سه نوع متداول از آنها یعنی سؤال «چندگزینه‌ای با بهترین پاسخ»، سؤال «جورکردنی گسترده» و سؤال «درست–

1. Licensure
2. National Board of Medical Examiners (NBME)
3. Case & Swanson
4. Membership of the Royal Colleges of Physicians (MRCP)
5. Membership of the Royal College of General Practitioners (MRCGP)
6. Professional and Linguistic Assessments Board (PLAB)
7. Anderson
8. Single Correct Answer

نادرست» خواهیم پرداخت. البته انواع دیگری از سؤالات کتبی نیز وجود دارند که از لحاظ شکلی و ساختاری می‌توانند در دسته سؤالات بسته‌پاسخ قرار بگیرند. از جمله می‌توان به آزمون تدبیر مشکل بیمار[1]، آزمون ویژگی‌های کلیدی[2] و آزمون همخوانی با شرحنامه[3] اشاره کرد. در این سؤالات دانشجو به دنبال یک سناریوی کوتاه با فهرستی از پاسخ‌ها مواجه می‌شود و قرار نیست خودش جوابی تولید کند. از آنجا که عمدتاً هدف این نوع سؤالات بیشتر از اینکه سنجش محفوظات و اطلاعات بالینی دانشجو باشد، ارزیابی توانایی استدلال بالینی[4] است، در این کتاب قسمت جداگانه‌ای به آنها اختصاص یافته است و در بخش چهارم به صورت مفصل ارائه خواهند شد.

سؤال چندگزینه‌ای با بهترین پاسخ (نوع A)

در این کتاب، بحث اصلی در خصوص همین نوع سؤال است. بنابراین از این پس به اختصار این نوع سؤال را «سؤال چندگزینه‌ای» می‌نامیم و در مورد آن در فصل دوم همین بخش به صورت تفصیلی صحبت خواهیم کرد. در این نوع سؤال، چند گزینه وجود دارند که تنها یکی از آنها درست است. گزینه‌های دیگر انحرافی[5] هستند و دانشجو فقط باید یک گزینه صحیح را از بین گزینه‌های موجود (معمولاً چهار گزینه، گاهی سه یا پنج گزینه) انتخاب کند. صحت گزینه‌ها نسبی است یعنی در نگاه اول به نظر می‌رسد که همه گزینه‌ها یا مثلاً سه گزینه صحیح هستند اما در واقع، گزینه‌ها نسبت به یکدیگر صحیح‌تر یا غلط‌تر هستند و از دانشجو خواسته می‌شود تا صحیح‌ترین گزینه را انتخاب کند که به این حالت سؤال بهترین پاسخ[6] گفته می‌شود.

نمونه‌ای از سؤال چندگزینه‌ای با بهترین پاسخ

راننده ۲۵ ساله به دلیل صدمه ناحیه میانی قدام گردن بر اثر اصابت به میله فرمان خودرو حین تصادف رانندگی به اورژانس آورده شده است. وی هوشیار است، تعداد ضربان قلب او ۸۵ و تعداد تنفس او ۲۸ در دقیقه است. آمفیزم زیرجلدی، تنگی نفس، دیسفاژی و دیسفونی دارد. مناسب‌ترین اقدام برای این بیمار چیست؟ پیش‌کارورزی شهریور ۸۸

الف) باریم سؤالو اورژانس ب) انتقال به اتاق عمل ج) سی تی اسکن اورژانس د) آنژیوگرافی اورژانس

سؤال «درست–نادرست»

سؤال«درست–نادرست»[7] به صورت یک جمله است که دانشجو باید تشخیص دهد درست یا نادرست است. هرچند به نظر می‌رسد طراحی این نوع سؤال آسان است، طراحی نوع صحیح و بدون خطای آن که دانشجو را دچار سردرگمی نکند، مخصوصاً در حوزه علوم پزشکی کار دشواری است و همین مسأله اغلب باعث می‌شود سؤال بیش از حد ساده شود. به عبارت دیگر، سؤال «درست–نادرست» اغلب یا مبهم است یا بسیار آسان (آلبنز[8] ۱۹۹۳، شوورث و ون‌درولوتن[9] ۲۰۰۳). علاوه بر این، اگر جمله نادرست باشد و دانشجو به درستی این مسأله را تشخیص دهد، نمی‌توان مطمئن بود که آیا جواب صحیح را می‌داند یا با اطلاعات ناقص و در حالی که به یک جواب نادرست دیگر فکر می‌کند، جمله را نادرست تشخیص داده باشد (داونینگ[10] ۱۹۹۲، شوورث و ون‌درولوتن ۲۰۰۳).

در مورد سؤال «درست–نادرست» در فصل چهارم همین بخش به صورت تفصیلی صحبت خواهد شد.

1. Patient Management Profile (PMP)
2. Key Feature (KF)
3. Script Concordance Test
4. Clinical reasoning
5. Distractor
6. Best Answer
7. True/False
8. Albanese
9. Schuwirth & van der Vleuten
10. Downing

نمونه‌ای از سؤال «درست–نادرست»

عصب ۷ حاوی الیاف پاراسمپاتیک است. درست☐ نادرست☐

سؤال جورکردنی (نوع B)

در سؤال جورکردنی[1] دو لیست مجزا (یکی با حروف و یکی با اعداد) در اختیار دانشجو قرار می‌گیرد که باید آنها را به صورت متناظر با هم جور کند. این نوع سؤال، که تعداد سؤال‌ها و جواب‌های آن با هم برابر است، هرچند چندان نامناسب نیست اما به علت سوق دادن دانشجویان به سمت حفظ کردن مطالب و همچنین احتمال زیاد پاسخ حدسی چندان مورد استفاده قرار نمی‌گیرد.

نمونه‌ای از سؤال جورکردنی

برای هر یک از موارد زیر، یکی از شماره‌ها را که بیشترین ارتباط را با آن داراست، انتخاب کنید.

الف) کوارکتاسیون آئورت ب) مجرای شریانی باز ج) تترالوژی فالو د) حلقه عروقی آئورتی ه) آرترزی تریکوسپید

۱. آناستوموز سیستمیک–شریان ریوی
۲. شایع‌ترین نارسایی قلبی سیانوتیک مادرزادی
۳. درمان جراحی توسط آناستوموز
۴. علت احتمالی دیسفاژی در نوزادان و کودکان
۵. افزایش فشار خون در بازوها و کاهش فشار خون در پاها

سؤال نوع C

این سؤال شباهت‌هایی با دو نوع قبلی دارد. در این نوع سؤال، یک لیست از جملات وجود دارد که با اعداد مشخص می‌شوند و دو گزینه نیز وجود دارند که با حروف نشان داده می‌شوند. همچنین گزینه ج و گزینه د به ترتیب «هر دو گزینه الف و ب» و «هیچ‌کدام» هستند. دانشجو برای هر یک از جملات، باید یکی از گزینه‌ها را انتخاب کند.

هدف این سؤال، مقایسه علایم و نشانه‌ها و تشخیص و درمان دو بیماری است. مشکل اصلی این سؤال همان است که در سؤال «درست–نادرست» عنوان شد؛ یعنی تشخیص جملات کاملاً درست یا کاملاً غلط مخصوصاً در حیطه علوم پزشکی دشوار است. قضاوت در این باره نسبی است و کمتر به دانش صرف پزشکی بر می‌گردد. به همین دلیل استفاده از آن توصیه نمی‌گردد (کیس و سوانسون ۲۰۰۲).

نمونه‌ای از سؤال نوع C

برای هر یک از جملات که با شماره نشان داده شده‌اند، یکی از حروف زیر را که به بهترین شکل به آن ارتباط دارد، انتخاب کنید.

الف) پلاسمیدیوم فالسی پاروم ب) پلاسمیدیوم ویواکس ج) هر دو د) هیچ‌کدام

۱. درمان انتخابی حمله حاد، ترکیبی از پریماکین و کلروکین است.
۲. در مناطق اندمیک، حملات بالینی با تزریق کلروکین هفته‌ای یک بار کنترل می‌شود.
۳. توسط کلروکین درمان قطعی می‌یابد.
۴. عفونت توسط تزریق کلروکین هفته‌ای یک بار پیشگیری می‌شود.

1. Matching Question

سؤال نوع D

این سؤال نیز یک لیست عددی و یک لیست حرفی دارد. پنج وضعیت (یا بیماری) با عدد و سه گزینه با حرف نشان داده می‌شوند. یکی از سه گزینه صحیح است یعنی به چهار بیماری مرتبط می‌باشد.

دانشجو اولاً باید حرفی را انتخاب کند که چهار بیماری به آن مرتبط هستند. ثانیاً باید بیماری را که به آن حرف مرتبط نیست، مشخص کند. از آنجا که سؤال گیج‌کننده است، طراحی آن دشوار است و بین دانشجوی توانمند و غیرتوانمند افتراق قائل نمی‌شود، این نوع سؤال امروزه منسوخ شده است (کیس و سوانسون ۲۰۰۲).

نمونه‌ای از سؤال نوع D

چهار بیماری از پنج بیماری زیر منطبق با یکی از گزینه‌ها هستند. اولاً آن گزینه را مشخص کنید و ثانیاً بیماری نامرتبط را مشخص کنید.

الف) ائورینوفیلی
ب) پلاسماسیتوز
ج) لنفوسیتوز

۱. تریشینوز
۲. مالتیپل میلوما
۳. سندرم لوفلر
۴. بیماری هوچکین
۵. شیستوزوما

سؤال نوع E

این نوع سؤال از یک جمله دو قسمتی تشکیل شده است که قسمت دوم، توضیحی در مورد دلیل جمله است. این نوع سؤال به علت اینکه طراح ی دشواری دارد و برای دانشجو نیز سردرگم کننده است، امروزه مورد استفاده قرار نمی‌گیرد. اما هدف از طراحی آن، تحلیل روابط بود و تصور می‌شد مهارت استدلال و درک مفاهیم پایه برای پاسخگویی آن لازم است (کیس و سوانسون ۲۰۰۲). انتخاب گزینه‌ها مطابق دستورالعمل خاصی به صورت زیر انجام می‌شود:

نمونه‌ای از سؤال نوع E

جمله‌ای: شیر گاو نسبت به شیر مادر برای تغذیه نوزاد ارجح است. زیرا هم: شیر گاو نسبت به شیر مادر کلسیم بیشتری دارد.

گزینه مناسب را در مورد جمله فوق انتخاب کنید.

	جمله اول	جمله دوم	
الف)	درست	درست	استدلال صحیح است.
ب)	درست	درست	استدلال صحیح نیست.
ج)	درست	نادرست	
د)	نادرست	درست	
ه)	نادرست	نادرست	

سؤال «درست–نادرست متعدد» (نوع J یا X)

در سؤال درست–نادرست متعدد»[1] که نوع J یا X نامیده می‌شود، یک پایه مشترک به همراه چند گزینه یا جمله ارائه می‌شود و دانشجو باید درستی یا نادرستی هر یک را مشخص کند.

نمونه‌ای از سؤال «درست–نادرست متعدد»

در مورد هر یک از داروهای فشارخون زیر مشخص کنید که متعلق به دسته مهارکننده هایACE هستند یا خیر.

الف) آتنولول
ب) آمیلوراید
ج) انالاپریل
د) فوروزماید
ه) دیگوکسین
و) وراپامیل
ز) تریامترن
ح) پروپرانولول

1. Multiple True/False

این نوع سؤال، علاوه بر شکل بالا، به صورت دیگری نیز طراحی می‌شود که در قالب امتحانات چندگزینه‌ای معمول بسیار شایع است.

نمونه‌ای از سؤال «درست–نادرست متعدد»

کدام یک از داروهای زیر از مهارکننده‌های MAO است؟ دستیاری اسفند ۸۸

الف) هالوپریدول ب) فلوکستین ج) ایمی پرامین د) فنلزین

نمونه‌ای از سؤال «درست–نادرست متعدد»

در سندرم حاد رتروویرال کدام جمله صحیح است؟ پیش‌کارورزی شهریور ۸۰

الف) در همه بیماران مبتلا به HIV دیده می‌شود.
ب) تظاهرات آن همزمان با مثبت شدن آنتی بادی در مبتلایان به HIV است.
ج) درمان آن موجب تاخیر در عوارض وابسته به HIV می‌شود.
د) درمان اولیه آن سبب کاهش تعداد ویروس HIV نمی‌شود.

اما باید توجه کرد که سؤال «درست–نادرست» متعدد از لحاظ مفهومی و ساختاری با سؤال «چندگزینه‌ای با بهترین پاسخ» متفاوت است و نباید به جای آن به کار برده شود. این سؤالات روی «یک» موضوع متمرکز نمی‌شوند، باعث سردرگمی دانشجو می‌شوند و قادر نیستند بین دانشجویان توانمند و ضعیف به خوبی افتراق قایل شوند و استفاده از آنها در آزمون‌های حساس و مهم[1] توصیه نمی‌شود (کیس و سوانسون ۲۰۰۲، کمپبل[2] ۲۰۱۱).

اطلاعات بیشتر در مورد سؤال «درست–نادرست» در فصل چهارم همین بخش ارائه می‌شود.

سؤال چندپاسخی

در سؤال چندپاسخی[3] دو یا چند گزینه صحیح وجود دارند و سایر گزینه‌ها، انحرافی هستند.

نمونه‌ای از سؤال چندپاسخی

کدام یک از مشخصات ویروس است؟

✓ الف) می‌تواند باعث بیماری شود. ب) توسط خودش تکثیر می‌شود.
ج) از چندین سلول تشکیل شده است. ✓ د) در گیاهان و جانوران زندگی می‌کند.

نمره‌دهی این گونه سؤالات ممکن است به این شکل باشد که چنانچه دانشجو تمام گزینه‌های صحیح را علامت بزند و هیچ یک از گزینه‌های غلط را انتخاب نکند، نمره سؤال را کامل دریافت کند. اشکال این حالت این است که نمره دانشجویی که برخی از گزینه‌های درست را تشخیص داده است، با نمره دانشجویی که اصلاً جواب را بلد نبوده است، یکی می‌شود. در حالت دیگر، به ازای انتخاب هر یک از موارد صحیح و عدم انتخاب هر یک از گزینه‌های غلط، قسمتی از نمره به دانشجو داده می‌شود. به صورت کلی، استفاده از این نوع سؤال به دلیل دشواری تصحیح و نمره‌دهی، توصیه نمی‌شود. در ضمن، معمولاً این سؤالات سطح دشواری بالا و روایی و پایایی پایینی دارند (برتون و همکاران[4] ۱۹۹۱). البته می‌توان گفت از آنجا که این نوع سؤال، مجموعه‌ای از سؤالات «درست–نادرست» است، یک راه برای حل مشکلات تصحیح این است که به عنوان جایگزین به شکل زیر نوشته شوند (برتون و همکاران ۱۹۹۱).

1. High stake
2. Campbell
3. Multiple Response
4. Burton et al

شکل بهتر سوال چند پاسخی: سؤال «درست–نادرست متعدد»

کدام یک از مشخصات ویروس است؟

الف) می‌تواند باعث بیماری شود. درست ☐ نادرست ☐
ب) توسط خودش تکثیر می‌شود. درست ☐ نادرست ☐
ج) از چندین سلول تشکیل شده است. درست ☐ نادرست ☐
د) در گیاهان و جانوران زندگی می‌کند. درست ☐ نادرست ☐

سؤال با پاسخ‌های ترکیبی یا پیچیده (نوع K)

سؤال با پاسخ‌های ترکیبی یا پیچیده[1] به نوع K معروف است. در این نوع سؤال، تعدادی پاسخ مطرح می‌شود که برخی از آنها صحیح هستند. گزینه‌های سؤال، ترکیب‌های مختلفی از این پاسخ‌ها هستند که دانشجو باید ترکیب صحیح را انتخاب کند. اِشکال این سؤال که معمولاً ترکیبی از چند سؤال «درست–نادرست» است، این است که معمولاً به دانشجویان با دانش ناکافی برای انتخاب ترکیب درست، سرنخ می‌دهد و همچنین تمام مشکلات مربوط به سؤالات چند پاسخی را نیز دارد. به همین دلیل استفاده از آنها توصیه نمی‌شود (برتون و همکاران ۱۹۹۱، آلبنز ۱۹۹۳).

نمونه‌ای از سؤال ترکیبی

عدم تعادل مایعات که به صورت ادم مشخص می‌شود، معمولاً همراه است با:

۱) واکنش‌های آلرژیک ۲) نارسایی احتقانی قلب ۳) سوختگی شدید ۴) کمبود پروتئین
الف) فقط ۴ ب) ۱ و ۳ ج) ۲ و ۴ د) ۱و ۲ و ۳ و ۴

در اینجا هم بهتر است سؤال به صورت «درست–نادرست متعدد» نوشته شود (برتون و همکاران ۱۹۹۱).

شکل بهتر سؤال ترکیبی: از سؤال «درست–نادرست متعدد

عدم تعادل مایعات که به صورت ادم مشخص می‌شود، معمولاً همراه است با:

الف) واکنش‌های آلرژیک درست ☐ نادرست ☐
ب) نارسایی احتقانی قلب درست ☐ نادرست ☐
ج) سوختگی شدید درست ☐ نادرست ☐
د) کمبود پروتئین درست ☐ نادرست ☐

سؤال جورکردنی گسترده (نوع R)

سؤال جورکردنی گسترده[2] که به نوع R مشهور است، مشابه سؤال جورکردنی است اما در آن تعداد گزینه‌ها از تعداد پایه‌های سؤال بیشتر است. معمولاً یک تم محوری وجود دارد که پایه‌ها و گزینه‌ها حول آن نوشته می‌شوند. فهرستی از گزینه‌های همگون (بیش از چهار گزینه) در اختیار دانشجو گذاشته می‌شود. سپس چند پایه سؤال (کمتر از تعداد گزینه‌ها) که حول محور مشترک و معمولاً به صورت مورد بالینی هستند، ارائه می‌شوند.

سؤال «جورکردنی گسترده» این قابلیت را دارد که با طراحی پایه‌های خوب و مناسب، میزان درک و قدرت تحلیل دانشجو را مورد سنجش قرار دهد و به علت تعدد گزینه‌ها احتمال حدس زدن دانشجو در آن کاهش پیدا می‌کند.

در مورد این سؤال در فصل سوم مفصلاً بحث خواهد شد.

1. Combined Response or Complex Response
2. Extended Matching Question (EMQ), Extended Matching Item (EMI)

نمونه‌ای از سؤال «جورکردنی گسترده»

محور: خستگی
الف) لوسمی حاد — ب) آنمی بیماری مزمن — ج) بیماری احتقانی قلب
د) توبرکلوز — ح) اسفروسیتوز ارثی — ز) فقر آهن
ط) کمبود ویتامین B_{12} — ه) عفونت ویروس اپشتن بار — و) هایپوتیرو‌ییدی

برای هر یک از بیماران زیر کدام یک از تشخیص‌های فوق محتمل‌تر هستند؟
۱. خانم ۱۹ ساله با خستگی، تب و گلودرد از هفته قبل به شما مراجعه کرده است. در معاینه لنفانوپاتی گردنی و اسپلنومگالی دارد و درجه حرارت بدنش ۳۸/۳ درجه سانتی‌گراد است. جواب آزمایش اولیه او به صورت زیر است:
Leukocyte count: 5000/mm3 (80% lymphocytes, with many lymphocytes with atypical features).
AST: 200 U/L.
Bilirubin: normal
ALT: normal
۲. دختر ۵۱ ساله با درد شکم و خستگی که دو هفته قبل شروع شده است، مراجعه کرده است. در معاینه متوجه رنگ پریدگی، کبودی و تندرنس در ناحیه ستون مهره‌ها و هر دو فمور می‌شوید. جواب آزمایش خونی او به صورت زیر است:
Hemoglobin concentration: 7.0 g/dL
Leukocyte count of 2000/mm3
Platelet count of 15,000/mm3

سؤال انتخاب چندتایی

سؤال انتخاب چندتایی[1] از لحاظ ساختاری مشابه سؤال «جورکردنی گسترده» است. با این تفاوت که در سؤال «جورکردنی گسترده» برای هر پایه یک گزینه به عنوان بهترین جواب در نظر گرفته می‌شود اما اینجا هر یک از پایه‌ها می‌تواند بیش از یک گزینه را به خود اختصاص دهد.

نمونه‌ای از سؤال انتخاب چندتایی

الف) کلسیم — ب) فلوراید — ج) اسید فولیک — د) آهن — و) ویتامین B_1 — ز) ویتامین B_6
ح) ویتامین B_{12} — ط) ویتامین C — ی) ویتامین D — ک) ویتامین E — ه) ویتامین A

برای هر یک از کودکان زیر ویتامین یا مواد معدنی مناسب را انتخاب کنید.
۱. نوزاد یک ماهه برای معاینه روتین به درمانگاه آورده شده است. معاینه فیزیکی طبیعی است و منحصرا با شیر مادر تغذیه شده است (دو مورد).
۲. دختر شش ساله مبتلا به سیستیک فیبروزیس که تحت درمان دارویی نیست (سه مورد).

سؤال منفی

در سؤال منفی[2] یکی از گزینه‌ها کاملاً یا نسبتاً غلط است و از دانشجو خواسته می‌شود آن را تشخیص دهد.

نمونه‌ای از سؤال منفی

در مورد جسم خارجی تراکئوبرونکیال و مری کدام گزینه غلط است؟ — پیش‌کارورزی شهریور ۸۰

الف) شایع ترین محل جسم خارجی مری، ناحیه کریکوفانژیال است.
ب) بهترین وسیله برای خارج کردن جسم خارجی تراکئوبرونکیال، برونکوسکوپ قابل انعطاف است.
ج) در جسم خارجی تراکئوبرونکیال، نمای رادیوگرافیک ممکن است کاملا نرمال باشد.
د) جسم خارجی مری نیز مانند جسم خارجی تراکئوبرونکیال به محض تشخیص باید خارج شود.

1. Pick N
2. Negative question

شیوه معمول‌تر نگارش سؤال منفی به صورت جمله‌ای است که در انتهای آن «به جز» آورده می‌شود. باید توجه کرد که این تغییر شیوه نگارش تفاوتی در ماهیت سؤل ایجاد نمی‌کند و کماکان سؤال منفی محسوب می‌شود.

نمونه‌ای از سؤال منفی

در ارتباط با آناتومی عدسی چشم تمام عبارات درست هستند؛ به جز: پیش‌کارورزی شهریور ۸۵

الف) یک لایه اپی تلیالی در زیر کپسول خلفی قرار دارد.
ب) با افزایش سن بر ضخامت عدسی افزوده و از ضخامت ارتجاعی آن کاسته می‌شود.
ج) خط‌های سوچورال در قسمت جلو به شکل ۷ مستقیم و در قسمت عقب به صورت ۷ دیده می‌شوند.
د) زنول ها در محل اکواتور به عدسی می‌چسبند.

البته شاید هر یک از انواع سؤالات کتبی بسته‌پاسخ قابل تبدیل به سؤال منفی باشند و نتوان آن را یک طبقه جدا محسوب کرد. اما در هر حال، استفاده از سؤالات منفی به صورت کلی توصیه نمی‌شود. در یک مطالعه مروری از بین ۴۶ منبع در زمینه سنجش آموزش که ۳۵ مطالعه به سؤال منفی پرداخته بودند، ۳۱ مطالعه ذکر کرده بودند نباید از این نوع سؤال استفاده شود (هالادینا و داونینگ[1] ۱۹۸۹). علت این موضوع آن است که در اکثر مواقع، برای اینکه مشخص شود آیا اهداف آموزشی محقق شده‌اند یا نه، بهتر است آنچه دانشجو می‌داند صحیح است شناسایی شود، نه مواردی که می‌داند اشتباه است. به عبارت دیگر، صرف اینکه دانشجو بداند جوابی غلط است، به معنای آن نیست که حتماً جواب درست را هم می‌داند.

علی‌رغم تاکید دستورالعمل‌ها مبنی بر عدم استفاده از سؤال منفی، طراحان به دلیل سهولت طراحی همچنان از این نوع سؤال طرح می‌کنند. زیرا به جای اینکه دنبال سه گزینه انحرافی باشند، فقط لازم است یک گزینه غلط بنویسند. بعضی از مطالعات ذکر کرده‌اند که استفاده از نوع مثبت یا منفی سؤال، تأثیری در عملکرد دانشجویان نداشته است اما برخی از تحقیقات آنها را موجب سردرگمی دانشجو و سختی بی مورد دانسته‌اند. همچنین، مطالعاتی هم به این نتیجه رسیده‌اند که سؤال منفی، نمره دانشجویان را به صورت غیرواقعی زیاد می‌کند زیرا طراح که خود می‌داند سؤال پیچیده شده است، گزینه درست را به گونه‌ای می‌نویسد که به راحتی قابل تشخیص باشد (بولند و همکاران ۲۰۱۰).

شایان ذکر است که موارد معدودی از اهداف آموزشی نیز وجود دارند که در آنها واقعاً اهمیت دارد دانشجو بداند که چه چیزی صحیح نیست، چه کاری را نباید انجام دهد، چه اقدامی در اولویت نیست یا کدام داروها را نباید با یکدیگر تجویز کرد. به عنوان مثال، پزشکی که قرار است با توجه به وضعیت بیمار برای او تصمیم‌گیری کند، گاهی لازم است حتماً کاری را انجام ندهد. در حقیقت در این موارد، دانستن این موضوع که چه اقداماتی نباید انجام شود، ضروری است و می‌توان با احتیاط از سؤال منفی استفاده کرد. اما حتماً باید توجه شود که قسمت منفی در پایه سؤال باشد و نه در گزینه‌ها و همچنین موضوع مورد پرسش کاملاً برجسته و مشخص شود.

نمونه‌ای از سؤال منفی قابل قبول

مرد ۹۵ ساله‌ای مبتلا به سندرم احتقانی قلب است. در بررسی آزمایشگاهی مشخص می‌شود میزان پتاسیم سرم وی ۲/۵ mmol/L است. تجویز تمام داروهای زیر در این بیمار مجاز است؛ به جز:

الف) کاپتوپریل ب) اسپیرنولاکتون ج) کارودیلول د) دیگوکسین

1. Haladyna & Downing

جمع‌بندی انواع سؤالات چندگزینه‌ای

همان‌طور که احتمالاً متوجه شده‌اید در برخی از انواع سؤالات چندگزینه‌ای، دانشجو باید گزینه «کاملاً صحیح» را انتخاب کند. سؤالات نوع X، نوع J یا K، نوع C و نوع E از این دست هستند. باید دقت شود که وقتی می‌گوییم کاملاً صحیح، به معنای واقعی کلمه می‌خواهیم که «کاملاً صحیح» باشد (کیس و سوانسون ۲۰۰۲). در حالی که برای سؤال نوع A و R بهترین پاسخ از بین گزینه‌های موجود مدنظر است.

برای بهتر نشان دادن منظور، پاسخ‌های سؤال «درست–نادرست» را با سؤال چندگزینه‌ای که در آن «بهترین پاسخ» انتخاب می‌شود، مقایسه می‌کنیم. این موضوع به صورت شکل زیر قابل نمایش است:

از آنجا که در سه فصل این بخش از کتاب به سه نوع سؤال «چندگزینه‌ای با بهترین پاسخ»، «جورکردنی گسترده» و «درست–نادرست» می‌پردازیم، در جدول ۱–۵ خصوصیات این سه نوع سؤال را خلاصه می‌کنیم:

جدول ۱–۵: مقایسه ویژگی‌های سؤالات «چندگزینه‌ای با بهترین پاسخ»، «جورکردنی گسترده» و «درست–نادرست»

جدول ۱–۵ : مقایسه ویژگی‌های سؤالات «چندگزینه‌ای با بهترین پاسخ»، «جورکردنی گسترده» و «درست–نادرست»

ویژگی	درست–نادرست	چندگزینه‌ای با بهترین پاسخ	جورکردنی گسترده
احتمال پاسخ حدسی	۵۰ درصد	۲۵ درصد (۴ گزینه‌ای)	۱۲/۵ درصد (هشت گزینه‌ای)
تمیز بین دانشجویان قوی و ضعیف	کم	خوب	خوب
سطح شناختی مورد ارزیابی	یادآوری محفوظات و جزئیات	سطوح بالاتر و یادآوری محفوظات	سطوح بالاتر و یادآوری محفوظات
دشواری طراحی سؤال بدون خطا	زیاد	متغیر	متغیر
تشویق به یادگیری	کم	خوب	خوب
ارزیابی مطابق اهداف (روایی)	کم (حتی در صورت نمونه‌گیری خوب)	بالا (در صورت نمونه‌گیری خوب)	بالا (در صورت نمونه‌گیری خوب)

مزایا و محدودیت‌های سؤالات بسته‌پاسخ

در اینجا به صورت کلی به مزایا و محدودیت‌های خانواده سؤالات بسته‌پاسخ می‌پردازیم و سپس در فصول بعدی حسب مورد به صورت خاص در مورد ویژگی‌های سه نوع سؤال (چندگزینه‌ای با بهترین پاسخ، جورکردنی گسترده و درست–نادرست) صحبت خواهیم کرد.

مزایای سؤالات بسته‌پاسخ

- پوشش وسیع محتوا: این یک واقعیت است که در آزمون، امکان طرح سؤال از تمام مطالب وجود ندارد و همیشه باید نمونه‌ای که نماینده کل محتوا باشد، از بلوپرینت[1] دوره انتخاب شود. هر چقدر این نمونه، مطالب بیشتری از دوره را پوشش دهد (به عبارتی نمونه بزرگتری باشد) و بیشتر با اهداف دوره مطابقت داشته باشد، روایی محتوایی[2] آزمون بیشتر است. سؤالات کتبی بسته‌پاسخ می‌توانند برای سنجش طیف وسیعی از اهداف آموزشی مورد استفاده قرار گیرند و به همین دلیل از آنها در رشته‌ها و موضوعات مختلف استفاده می‌شود. این سؤالات اگرچه به راحتی برای ارزیابی سطوح پایین حیطه شناختی مانند یادآوری[3] و بازشناسی[4] به کار می‌روند اما اگر خوب طراحی شوند، می‌توانند سطوح پیچیده‌تر مانند درک و فهم، کاربرد اطلاعات، تحلیل داده‌ها و استدلال بالینی را نیز بسنجند. از طرف دیگر، از آنجا که پاسخ به سؤالات بسته‌پاسخ زمان زیادی نمی‌برد، این امکان وجود دارد که دانشجو در زمان محدود به تعداد زیادی سؤال از مطالب متفاوت پاسخ دهد. بنابراین، می‌توان گفت که به صورت بالقوه، روایی محتوایی در آزمون‌های بسته‌پاسخ خوب است.
- عینیت[5]: یک سؤال بسته‌پاسخ خوب، تحت تاثیر عواملی که هدف اصلی سنجش نیستند، قرار نمی‌گیرد. مثلاً در یک آزمون تشریحی، جدا از اینکه سطح سواد دانشجو چقدر باشد، نحوه نگارش یا دستخط او ممکن است در نمره‌ای که می‌گیرد، تاثیر بگذارد. یا در امتحانات شفاهی، ممکن است سوگیری‌ها و خطاهای شخصی استاد در نمره دخیل شوند. در حالی که این موارد تا حد زیادی در سؤال بسته‌پاسخ کنترل می‌شوند و به همین علت عینیت بالایی وجود دارد.
- تصحیح آسان: در این دسته از آزمون‌ها با توجه به اینکه جواب ثابت است و از طریق علامت زدن مشخص می‌شود، به راحتی پاسخنامه با استفاده از رایانه و توسط نرم افزار خوانده می‌شود. به همین دلیل در صورتی که تعداد فراگیران زیاد باشد، این روش قابلیت اجرای بالایی دارد.
- تحلیل ساده سؤالات: در این سؤالات، شاخص‌هایی مانند تمیز و دشواری به راحتی قابل محاسبه هستند. بنابراین، طراح سؤال به راحتی می‌تواند به بسیاری از اشکالات سؤال پی ببرد و آنها را رفع کند. به عنوان مثال، می‌تواند گزینه‌های انحرافی را که خوب عمل نکرده‌اند، تشخیص دهد و آنها را با گزینه‌های دیگر جایگزین نماید.
- ایجاد بانک سؤالات: جمع آوری سؤالات با پاسخ‌ها و شاخص‌هایشان به راحتی صورت می‌گیرد و این امکان را ایجاد می‌کند که سؤالات مناسب‌تر برای آزمون‌های بعدی مورد استفاده قرار گیرند.
- امکان خودآموزی و یادگیری مستقل: دانشجو می‌تواند از این نوع سؤال برای سنجش میزان یادگیری خود استفاده کند و همچنین برآوردی از میزان آمادگی خود برای امتحان به دست آورد.
- شفافیت: ارائه اطلاعات دقیق و شفاف در مورد امتحان مانند شکل سؤالات، ضرایب، نحوه نمره‌دهی و ... آسان است.

---

1. Blueprint
2. Content Validity
3. Recall
4. Recognition
5. Objectivity

محدودیت‌های سؤالات بسته‌پاسخ

- سؤالات بسته‌پاسخ معمولاً فقط قادرند حیطه دانشی را اندازه‌گیری کنند. در مواردی که نگرش یا مهارت عملی فرد باید ارزیابی شود، استفاده از سؤالات بسته‌پاسخ به تنهایی کافی نیست. علاوه بر این، در خود حیطه دانشی معمولاً سؤالات به سطوح پایین محدود می‌شوند. زیرا اولاً طراحی سؤالاتی که به ارزیابی سطوح بالاتر بپردازند و در عین حال، فاقد خطا باشند، بسیار دشوار است. ثانیاً مواردی مانند خلاقیت، تفسیر اطلاعات، سازمان‌دهی مطالب، مثال زدن و ... اساساً قابل سنجش با سؤال بسته‌پاسخ نیستند. به این ترتیب، همان‌طور که قبلاً ذکر شد، مهم است که در هر دوره آموزشی، نوع آزمون و سؤالات، مطابق با اهداف و بلوپرینت دوره انتخاب شوند. هماهنگی آزمون با اهداف دوره، منجر به بهبود روایی محتوایی آزمون می‌گردد.
- سؤالات بسته‌پاسخ معمولا با آنچه دانشجو در محیط واقعی با آن سر و کار خواهد داشت، فاصله دارند. هرچند تلاش می‌شود با طراحی سؤالات مبتنی بر سناریو، این فاصله کمرنگ شود، باز هم تمام ابعاد شرایط واقعی مواجهه با بیمار از جمله تصمیم‌گیری در شرایط عدم قطعیت یا اورژانسی قابل ارائه و ارزیابی به صورت کتبی نیستند. این مسأله بر روایی صوری[1] آزمون تاثیر می‌گذارد.
- از آنجا که در سؤالات بسته‌پاسخ، دانشجو جواب را می‌بیند، ممکن است پاسخ را به صورت حدسی انتخاب کند و به صورت شانسی گزینه صحیح را علامت بزند. میزان حدس زدن در سؤالات «چندگزینه‌ای با بهترین پاسخ»، نسبت به سؤالات «درست–نادرست» کمتر و نسبت به سؤالات «جورکردنی گسترده» بیشتر است. این پدیده مخصوصاً در سؤالاتی که خوب طراحی نشده‌اند و حاوی خطا هستند، رایج است و پایایی آزمون را تحت تأثیر قرار می‌دهد (نودیم[2] ۱۹۹۲). به همین دلیل راه‌کارهایی (مانند نمره منفی) برای کاهش حدس زدن مورد استفاده قرار می‌گیرد. البته شواهد نشان می‌دهند که این راه‌کارها پایایی آزمون را چندان بهبود نبخشیده‌اند (برتون ۲۰۰۲). در قسمت مربوط به «سؤالات رایج» در مورد این موضوع مفصلاً صحبت خواهد شد.
- این احتمال وجود دارد که سؤالات بسته‌پاسخ دانشجویان را به یادگیری سطحی و حفظ کردن مطالب تشویق کنند.
- پس از سال‌ها مواجهه با سؤالات بسته‌پاسخ، دانشجویان در شناسایی الگوها، سرنخ‌ها و تکنیک‌های پاسخ خبره می‌شوند. از این مسأله تحت عنوان «تکنیک تست‌زنی»[3] یاد می‌شود که البته خطاهای طراحی سؤال نیز در آن موثرند.
- احتمال اینکه دانشجو در آزمون بسته‌پاسخ گزینه‌های نادرست را یاد بگیرد، وجود دارد زیرا بیشتر از اینکه پاسخ‌های صحیح را ببیند، با پاسخ‌های غلط مواجه می‌شود. این موضوع، مخصوصاً وقتی که آزمون‌دهنده بعد از اتمام آزمون پاسخ‌های صحیح آن را چک نمی‌کند، ممکن است باعث به خاطر سپردن همان گزینه اشتباه شود و وی در آینده همان را به خاطر آورد.
- طراحی سؤال چندگزینه‌ای خوب و مطلوب طبق راهنماها و دستورالعمل‌های موجود، زمان‌بر و دشوار است. مخصوصاً طراحی گزینه‌های انحرافی خوب نیازمند تجربه و تمرین فراوان است.

1. Face Validity
2. Nnodim
3. Testwiseness

نکات مثبت و منفی سؤالات بسته‌پاسخ در جدول ۲-۵ خلاصه شده‌اند:

جدول ۲-۵: خلاصه نکات مثبت و منفی آزمون‌های کتبی بسته‌پاسخ

نکات مثبت	نکات منفی
امکان سنجش سطوح مختلف اهداف شناختی	تاکید عمده بر ارزیابی سطوح پایین شناختی
روایی محتوایی خوب	روایی پایین برای اهداف مهارتی و نگرشی
عینیت و پایایی بالا	روایی صوری نسبتاً پایین
امکان سنجش اطلاعات زیاد در زمان محدود	امکان طراحی سؤالات مبهم و پیچیده
امکان برگزاری آزمون مشابه برای تعداد زیاد دانشجو	امکان فراهم کردن سرنخ برای رسیدن به جواب
هزینه پایین	احتمال تشویق حدس زدن
امکان استفاده برای خودآموزی و یادگیری مستقل	احتمال تشویق یادگیری سطحی وکوتاه مدت
تصحیح آسان و سریع	طراحی نسبتاً دشوار

به طور خلاصه می‌توان گفت که سؤالات بسته‌پاسخ مزایای زیادی دارند و در اکثر مقاطع و رشته‌های علوم پزشکی به عنوان شایع‌ترین روش ارزیابی دانشجویان مطرح هستند. اما باید در نظر داشت که این سؤالات برای هر موقعیتی مناسب نیستند. در واقع، استفاده زیاد از آنها به قابلیت اجرای بالای آنها بر می‌گردد و نه لزوماً بی‌نقص بودنشان. مخصوصاً در مواردی که قرار است محتوا و حجم زیادی از مطالب ارزیابی شود، وقت آزمون محدود است یا قرار است تعداد زیادی دانشجو حتی در مکان‌های مختلف، در امتحان شرکت کنند، استفاده از سایر ابزارها از لحاظ اجرایی بسیار دشوار و در مواردی غیرممکن می‌شود. در چنین حالاتی استفاده از سؤال تستی غیرقابل اجتناب خواهد بود. هنگام استفاده از سؤالات چندگزینه‌ای مهم است که موارد زیر از لحاظ هدف ارزیابی مدنظر قرار گیرند:

- چنانچه هدف آموزشی این باشد که دانشجو بتواند از بین راه‌حل‌ها یا گزینه‌های مختلف، بهترین را انتخاب کند، سؤال چندگزینه‌ای، روش مناسبی است.
- گاهی یک هدف آموزشی وقتی بهتر اندازه‌گیری می‌شود که دانشجو خودش به پاسخ‌های محتمل فکر کند. در این صورت، سؤال کوته‌پاسخ یا تشریحی بهتر هستند.
- چنانچه تعداد زیادی گزینه همگن وجود دارد، می‌توان برای استفاده بهتر از وقت و گزینه‌ها، از سؤال جورکردنی استفاده کرد.
- گاهی اساساً هدف آموزشی این است که دانشجو کاری را عملاً انجام دهد. در این مواقع، بهتر است به جای اینکه در یک آزمون کتبی از دانشجو خواسته شود که روند یا شرایط کار را توضیح دهد، در یک موقعیت عملی مورد ارزیابی قرار گیرد.

سودمندی سؤالات بسته‌پاسخ

با توجه به نکات مثبت و منفی که برای آزمون‌های کتبی بسته‌پاسخ ذکر شد، این سؤال پیش می‌آید که آیا بالاخره باید از آنها استفاده کرد یا بهتر است به شیوه‌های دیگر ارزیابی روی آورد. این سؤال جواب مستقیم و کوتاهی ندارد. پاسخ

این است که ابتدا باید دید ارزیابی در چه موقعیتی و برای چه منظوری قرار است انجام شود. یک ابزار ارزیابی ویژگی‌های قطعی و ثابتی ندارد که بتوان در مورد مفید بودن آن حکم کلی داد بلکه در شرایط مختلف باید تصمیم گرفت که آیا یک ابزار سودمند است یا باید از روش‌های دیگر استفاده کرد.

همان‌طور که در بخش اول کتاب اشاره شد، ون‌درولوتن در سال ۱۹۹۶ برای ارزیابی سودمندی[1] یک ابزار ارزیابی فرمولی با ۵ معیار ارائه داد: روایی[2]، پایایی[3]، تاثیر آموزشی[4]، هزینه[5] و مقبولیت[6]. به صورت معمول، هزینه–اثربخشی روش و مقبولیت یک ابزار در کنار هم تحت عنوان قابلیت اجرا[7] بررسی می‌شوند.

ارزش و اهمیت این معیارها در تمام آزمون‌ها یکسان نیست. با توجه به اهدافی که هر امتحان دنبال می‌کند و سطح اهمیتی که دارد، ممکن است یکی از این معیارها بیشتر مورد توجه باشد. به عنوان مثال، آزمونی که برای ارائه مجوز کار و گواهینامه برگزار می‌شود، ضروری است که روایی و پایایی بالایی داشته باشد، اما لازم نیست اثر آموزشی چندانی داشته باشد و اگر هزینه بالایی داشته باشد، باز هم شاید مقرون به صرفه باشد. در حالی که در مورد ارزیابی مهارت‌های بالینی طی دوره لازم است که اثر آموزشی و قابلیت اجرای خوبی وجود داشته باشد تا پایایی بالا. به طور کلی، آنچه حائز اهمیت است، در نظر گرفتن تمام این موارد با یکدیگر و برقراری تعادل برای انتخاب آزمون مناسب است.

روایی سؤالات بسته‌پاسخ

روایی آزمون به این بر می‌گردد که تا چه حد آن چیزی که مدنظر ما بوده است، واقعاً با این آزمون قابل سنجش است. از آنجا که دانشجو می‌تواند در زمان محدودی به تعداد زیادی سؤال بسته‌پاسخ جواب دهد، می‌توان گفت که با این نوع سؤال پوشش خوبی از محتوا قابل دستیابی است و روایی محتوایی به صورت کلی خوب است. در عین حال، برای بررسی روایی محتوایی توجه به دو نکته زیر حائز اهمیت است:

- چنانچه هدف ارزیابی، سنجش حیطه دانشی و شناختی دانشجویان است، سؤالات بسته‌پاسخ مناسب هستند و روا محسوب می‌شوند. اما اگر اهداف مهارتی، عملی، بالینی و نگرشی نیز جزء اهداف دوره باشند، باید توجه داشت که سؤالات بسته‌پاسخ به تنهایی قادر به سنجش آنها نیستند و بنابراین برای این دسته از اهداف روایی ندارند. از آنجا که به ندرت پیش می‌آید در یک دوره آموزشی اهداف عملی و نگرشی وجود نداشته باشد، محدود کردن ارزیابی فراگیران تنها به یک آزمون کتبی اشتباه است. در همین راستا، استفاده از چندین نوع ابزار ارزیابی برای پوشش تمام اهداف دوره توصیه می‌شود.
- یکی از مسائلی که بر روایی محتوایی آزمون تاثیر می‌گذارد، نمونه‌گیری خوب از محتوا می‌باشد. معمولاً در یک آزمون امکان این که از تمام مباحث، تمام سؤالات ممکن پرسیده شود، وجود ندارد. به همین دلیل در یک آزمون، به اجبار تنها تعدادی از سؤالات انتخاب می‌شوند. انتظار می‌رود که این تعداد سؤال، نمونه خوب و مناسبی از کل اهداف آموزشی باشند به طوری که نمره حاصل از آنها، نمایانگر وضعیت واقعی دانشجو باشد و قابلیت تعمیم خوبی داشته باشد. در این صورت گفته می‌شود که آزمون روایی محتوایی مطلوبی دارد. از آنجا که دانشجو می‌تواند در واحد زمان به تعداد زیادی سؤال بسته‌پاسخ جواب دهد و امکان طرح تعداد زیاد سؤال وجود دارد، انتظار می‌رود در آزمونی که از سؤال بسته‌پاسخ تشکیل شده است، پوشش مناسبی از کل محتوا برقرار شود و روایی محتوایی خوبی به دست آید. تنها باید مراقب بود که توزیع سؤالات از مباحث مختلف به درستی صورت گیرد. یعنی به این شکل نباشد که از قسمت عمده سؤالات از یک یا دو مبحث طراحی شوند.

1. Utility
2. Validity
3. Reliability
4. Educational impact
5. Cost
6. Acceptability
7. Feasibility

آنچه روایی محتوایی را در مورد هر دو مسأله فوق تضمین می‌کند، تهیه بلوپرینت یا جدول مشخصات آزمون است که در بخش اول کتاب توضیح داده شد. البته این موضوع فقط به سؤال چندگزینه‌ای محدود نمی‌شود و در هر آزمونی، طرح سؤال باید بر اساس بلوپرینت انجام شود. در غیر این صورت، خطر کم پوشش دادن محتوا[1] وجود دارد.

خطر دیگری که روایی آزمون را تهدید می‌کند، واریانس بی‌ارتباط با سازه[2] است که ناشی از متغیرهایی است که به صورت نظام‌مند[3] بر نتایج آزمون اثر می‌گذارند و اجازه نمی‌دهند تفسیر معناداری از آنچه اندازه‌گیری شده صورت بگیرد. این متغیرها شامل خطاهای ساختاری سؤال[4]، تکنیک‌های تست‌زنی، تقلب و خستگی دانشجویان است. به دلیل تاثیر خطاهای ساختاری سؤال بر روایی آزمون، تلاش برای کاهش این خطاها از طریق طراحی سؤال بر اساس دستورالعمل‌های موجود و سپس مرور و ارزیابی آنها به کمک چک‌لیست‌ها همواره توصیه شده است.

همان‌طور که در اولین بخش کتاب اشاره شد، توصیه‌های اصلی برای بهبود روایی ارزیابی شامل موارد زیر است (اندرسون ۲۰۰۴، چاندراتیلکه و همکاران[5] ۲۰۱۱):

- تنوع ابزارهای ارزیابی
- استفاده از بلوپرینت برای طرح سؤال
- استفاده از دستورالعمل‌های طراحی سؤال
- طراحی سؤالات مناسب برای ارزیابی سطوح شناختی بالا

پایایی سؤالات بسته‌پاسخ

همان‌طور که در بخش اول کتاب بیان شد، پایایی در یک آزمون به صورت ساده به این معنا است که نمره دانشجو تا چه اندازه در تکرار اندازه‌گیری‌ها در شرایط مختلف یکسان باقی می‌ماند. به صورت کلی، انتظار می‌رود نمره دانشجو در شرایط مختلف ارزیابی یکسان باشد. اگر این طور نباشد، ناشی از خطای آزمون است که می‌تواند از منابع مختلفی ناشی شود. پایایی آزمون بیان می‌کند که نمره دانشجو در این امتحان تا چه حد قابل تعمیم به سایر شرایط است.

یکی از منابع خطای آزمون می‌تواند مصحح باشد. اگر برگه دانشجو توسط مصححان مختلف تصحیح شود، ممکن است نمره دانشجو تغییر کند. از آنجا که تصحیح سؤال چندگزینه‌ای به صورت صفر و یک است، امکان تفسیر پاسخ (مانند آنچه در سؤال تشریحی می‌بینیم) وجود ندارد. بنابراین، این نگرانی که مصححان مختلف با ذهنیت‌های متفاوت نمرات مختلفی به دانشجو بدهند، وجود ندارد.

منبع دیگر خطا، سؤالات آزمون هستند. به بیان دیگر این پرسش مطرح است که آیا اگر دانشجو در یک آزمون بسته‌پاسخ با نمونه سؤالات متفاوتی از همان دوره شرکت کند، نمره مشابه‌ای دریافت خواهد کرد؟ همان‌طور که در قسمت روایی بیان شد، بدیهی است که برای یک دوره، می‌توان سؤالات زیادی را در نظر گرفت اما معمولاً مقدور نیست که تمام آن سؤالات در آزمون پرسیده شود و باید تعدادی از سؤالات از بین این مجموعه انتخاب شوند. مهم است که سؤالات انتخاب شده، نماینده خوبی از محتویات کل دوره باشند. مثلاً توزیع مناسبی از نظر فصول کتاب یا جلسات کلاسی یا اهمیت مطلب داشته باشند. اگر برای انتخاب سؤال از بلوپرینت استفاده شود، نمونه بهتری از سؤالات ممکن به دست خواهد آمد و بیشتر می‌توان مطمئن بود که نتایج آزمون قابل تعمیم به آزمون‌های مشابه دیگری است که به صورت بالقوه امکان داشت برای همین دوره برگزار شوند.

عامل دیگر تاثیرگذار بر پایایی آزمون، تعداد سؤالات است. از آنجا که پاسخ به سؤالات بسته‌پاسخ نسبت به سایر انواع آزمون زمان

1. Content underrepresentation
2. Construct Irrelevant variance (CIV)
3. Systematic
4. Technical item flaws
5. Chandratilake et al.

کمتری می‌برد، تعداد سؤالات بیشتری در واحد زمان می‌توان طرح کرد که موجب بهبود پایایی آزمون می‌شود.

مسأله‌ای که در مورد پایایی آزمون‌های بسته‌پاسخ نگرانی ایجاد کرده است، تاثیر حدس زدن است. زیرا گفته می‌شود حدس، با وارد کردن یک متغیر تصادفی باعث کاهش روایی و پایایی آزمون می‌شود. در رابطه با حدس زدن، میزان آن، تاثیر آن بر ویژگی‌های روان‌سنجی آزمون‌های بسته‌پاسخ و همچنین راه‌کارهای کاهش آن در قسمت «سؤالات رایج» مفصلاً صحبت خواهد شد. اما به صورت خلاصه باید گفت اولاً پاسخ حدسی لزوماً و در همه موارد کار اشتباهی نیست و همچنین، راه‌کارهایی که برای مقابله با آن به کار رفته‌اند (مانند نمره منفی)، لزوماً پایایی آزمون را بهبود نبخشیده‌اند.

به طور کلی، سؤالات بسته‌پاسخ معمولاً در واحد زمان، پایایی بالاتری نسبت به سؤالات بازپاسخ دارند (شوورث و ون‌درولوتن ۲۰۰۴). (پایایی آزمون‌ها معمولاً در واحد یک ساعت آزمون بیان می‌شود. زیرا مدت زمان امتحان به علت خستگی دانشجویان و محدودیت منابع، یک فاکتور محدودکننده است). همان‌طور که گفته شد، یکی از دلایل این است که در یک مدت مشخص، تعداد سؤالات بسته‌پاسخی که می‌توان ارائه کرد بیشتر از سؤالات بازپاسخ است. همچنین، دخالت ذهنیت مصححان در تصحیح سؤالات بازپاسخ روی پایایی تاثیر می‌گذارد که در سؤالات بسته‌پاسخ صدق نمی‌کند.

همان‌طور که در بخش اول کتاب اشاره شد، برای بهبود پایایی ارزیابی، توصیه‌های اصلی شامل موارد زیر است:

- تنوع ابزارهای ارزیابی
- برگزاری آزمون‌های مکرر
- افزایش تعداد سؤالات
- همگون کردن سؤالات با اهداف دوره
- طراحی سؤالات از تمام اهداف و موضوعات

تاثیر آموزشی سؤالات بسته‌پاسخ

هر آزمونی جدا از اینکه میزان یادگیری دانشجویان را اندازه‌گیری می‌کند، بر یادگیری دانشجویان تاثیر هم می‌گذارد. این تاثیر آموزشی هم به رفتار دانشجو و هم به رفتار طراح سؤال برمی‌گردد.

به نظر می‌رسد دانشجویان آن چیزی را مطالعه می‌کنند و یاد می‌گیرند که بدانند یا حدس بزنند در امتحان مورد پرسش واقع می‌شود. به این ترتیب دانشجویان علاقه‌ای ندارند وقت خود را به خواندن مطالبی اختصاص دهند که در امتحان مورد تأکید نیست. بنابراین، احتمالاً نوع آزمون بر نحوه درس خواندن دانشجویان اثر می‌گذارد (اندرسون ۲۰۰۴). دانشجویان اگر بدانند آزمون چندگزینه‌ای دارند یا اگر به آنها گفته شود که سؤالات تشریحی هستند، متفاوت درس می‌خوانند. هرچند شواهد تجربی کافی وجود ندارد که این امر در نتیجه عملکرد آنها نیز تاثیری دارد یا نه.

اگر قرار باشد تمام سؤالات آزمون به صورت بسته‌پاسخ باشند، از آنجا که نمی‌توان از همه مطالب، سؤال بسته‌پاسخ خوبی طرح کرد، طراحان سؤال سراغ برخی از قسمت‌های کتاب نمی‌روند و پس از چند نوبت، دانشجویان احساس می‌کنند این بخش‌ها برای مدرسان بی‌اهمیت هستند. در حالی که اگر علاوه بر سؤال بسته‌پاسخ، انواع دیگر سؤالات مانند سؤالات بازپاسخ استفاده شود، این محدودیت کمتر می‌شود. همین موضوع در مورد مهارت‌های عملی دانشجویان نیز صدق می‌کند. به عنوان مثال، اگر ارزیابی دانشجویان در انتهای بخش داخلی صرفاً بر اساس آزمون کتبی باشد، دانشجویان توجه‌ای به تقویت مهارت‌های بالینی خود مانند معاینه و شرح حال و ... نمی‌کنند و ساعات خود را صرف مطالعه در کتابخانه می‌کنند. این موضوع نیز بیشتر از اینکه به اشکال استفاده از سؤال چندگزینه‌ای برگردد، ناشی از استفاده صرف از «یک» نوع آزمون است که احتمالاً با تنوع ابزار برطرف می‌گردد. البته باید توجه داشت که شواهد تجربی کافی دال بر این قضیه که آیا واقعا آمادگی برای امتحان‌های متفاوت باعث می‌شود که توانمندی‌های کسب شده و آموخته‌های دانشجویان متفاوت باشد یا نه، وجود ندارد (شوورث و ون‌درولوتن ۲۰۰۴).

به علاوه در مورد تاثیر آموزشی سؤالات بسته‌پاسخ می‌توان گفت که اگر سؤالات همیشه بر مبنای سطوح شناختی پایین باشند، دانشجویان بیشتر به سمت حفظ کردن مطالب کتاب می‌روند و مهارت‌های حل مسأله و تفکر در آنها کمرنگ می‌شود. در حالی که اگر شرکت کنندگان در آزمون از قبل بدانند که برای پاسخگویی به سؤالات نیازمند مطالعه عمیق‌تر و غور در مباحث هستند، احتمالاً از ابتدا طور دیگری درس می‌خوانند.

نگرانی دیگر که در مورد این دسته از آزمون‌ها مطرح می‌شود مسأله حدس زدن جواب درست است که موجب می‌شود دانشجویی که برای امتحان بسته‌پاسخ آماده می‌شود، قسمتی از موفقیت خود را به پاسخ‌های شانسی و بدون یادگرفتن مطالب واگذار کند.

با این حال، سؤالات بسته‌پاسخ امکان خودآموزی و یادگیری مستقل را برای دانشجو فراهم می‌کنند. دانشجو می‌تواند از این نوع سؤال برای سنجش میزان یادگیری خود استفاده کند و همچنین برآوردی از میزان آمادگی خود برای امتحان داشته باشد.

همچنین، با در اختیار گذاشتن کلید سؤالات بعد از آزمون می‌توان امیدوار بود که دانشجویان روی آموخته‌های خود تأمل کنند، به اشتباهات خود پی ببرند و یادگیری خود را اصلاح کنند. به علاوه مرور سؤالات بعد از آزمون توسط استاد و بحث در مورد پاسخ درست، تأثیر آموزشی روش ارزیابی را تقویت می‌کند.

در بخش اول کتاب توصیه‌هایی برای بهبود تأثیر آموزشی روش ارزیابی ارائه شد که مجدداً به آنها اشاره می‌شود:

- تنوع روش‌های ارزیابی
- طراحی سؤال با هدف سنجش سطوح بالای شناختی توجه به ارزیابی‌های تکوینی مکرر و همراه با بازخورد مؤثر در طول دوره در کنار آزمون‌های تراکمی
- در اختیار گذاشتن کلید سؤالات پس از برگزاری آزمون، ارائه بازخورد و بحث در مورد آنها

مقبولیت، هزینه و قابلیت اجرای سؤالات بسته‌پاسخ

بهترین ابزار ارزیابی اگر مورد قبول دانشجو و استاد نباشد، دوام نمی‌آورد و استفاده از آن مزیتی به دنبال نخواهد داشت. سؤالات بسته‌پاسخ به دلیل عینی بودن، سهولت استفاده و قابلیت اجرای بالا، مکرراً در مقاطع مختلف استفاده شده‌اند و از این رو هم بین دانشجویان و هم بین مدرسان پذیرفته شده‌اند، به طوری که شاید اکنون مشکل اصلی، متقاعد کردن گروه‌های آموزشی برای جایگزین کردن آنها با انواع آزمون‌های دیگر باشد.

هزینه آزمون شامل هزینه طراحی سؤال، هزینه اجرای آزمون و هزینه تصحیح برگه‌ها می‌شود. ممکن است که طرح سؤالات بسته‌پاسخ سخت‌تر از سؤالات بازپاسخ باشد. برای ساختن سؤالات بسته‌پاسخ مطلوب که محتوای غنی دارند، زمان زیادی باید صرف شود. در عوض، تهیه پاسخنامه و کلید و سپس تصحیح نرم‌افزاری برای سؤالات بسته‌پاسخ بسیار ساده‌تر از سؤالات بازپاسخ است. برای اینکه روند تصحیح سؤالات بازپاسخ به درستی انجام شود، نیاز به پاسخنامه‌ای وجود دارد که پاسخ‌های صحیح احتمالی با جزئیات در آن مشخص شده باشد که کار ساده‌ای نیست (شوورث و ون‌درولوتن ۲۰۰۴).

باورهای نادرست در مورد سؤالات بسته‌پاسخ

ارزیابی سطوح شناختی پایین

همیشه این انتقاد به سؤالات چندگزینه‌ای وارد بوده که فقط قادر به ارزیابی سطوح پایین شناختی هستند و برای سنجش مهارت‌های شناختی سطح بالا مانند نحوه استدلال و حل مسأله دانشجویان نمی‌توان از آنها استفاده کرد. به صورت رایجی تصور می‌شود که سؤالات بازپاسخ نسبت به سؤالات چندگزینه‌ای (یا کلاً سؤالات بسته‌پاسخ)، ارزیابی درست‌تر و عمیق‌تری از میزان آموخته‌های فراگیران ارائه می‌دهند.

درحالی که در واقعیت، آنچه عملاً تعیین می‌کند چه سطحی از اهداف مورد ارزیابی قرار گرفته‌اند، شکل و ساختار سؤال (به عبارت دیگر، اینکه سؤال چندگزینه‌ای است یا تشریحی) نیست بلکه محتوای سؤال و نحوه پرسش است (ون‌درولوتن ۱۹۹۶). به این ترتیب ممکن است یک سؤال تشریحی طراحی شود که فقط محفوظات را می‌سنجد و از طرفی با به کاربستن اصول و راهنماها، می‌توان یک سؤال چندگزینه‌ای طراحی کرد که دانشجو برای پاسخ به آن باید از مهارت‌های حل مسأله و استدلال خود استفاده کند. در واقع صرف اینکه در سؤال تشریحی، دانشجو کلمات و جملات را از خودش می‌سازد و می‌نویسد، دلیلی بر این نیست که همیشه همه سؤالات تشریحی قادر به ارزیابی توانایی استدلال، تفسیر، حل مسأله و توانایی‌هایی از این قبیل هستند.

با وجود اینکه مطالعات زیادی این ذهنیت‌ها را به صورت تجربی نیز مورد مطالعه قرار داده‌اند و آنها را تأیید نکرده‌اند، هنوز یکی از باورهای رایج در مورد انواع آزمون‌ها این است که سؤال چندگزینه‌ای، فقط یادگیری سطحی دانشجویان را مورد سنجش قرار می‌دهد و صرفاً بر محفوظات دانشجو تاکید دارد. اگر چه آزمون‌های چندگزینه‌ای امکان سنجش بالاترین سطوح حوزه شناختی در تاکسونومی بلوم (مانند سنتز داده‌ها، خلاقیت، فرضیه‌سازی) را ندارند ، اما این گونه هم نیست که فقط به سطح یادآوری[1] محدود شوند (شکل ۲-۵). در واقع، سطوح فهم و درک[2]، کاربرد[3]، تجزیه و تحلیل[4] و ارزشیابی[5] با این ابزار قابل سنجش هستند.

هر چقدر محتوای سؤال غنی‌تر باشد، احتمال اینکه سطوح بالاتر را بسنجد بیشتر است. این موضوع در مثال‌های زیر نشان داده می‌شود. سؤالات زیر در سطوح مختلف شناختی، حول یک موضوع واحد یعنی اختلالات آب و الکترولیت طراحی شده‌اند:

سطح شناختی: یادآوری

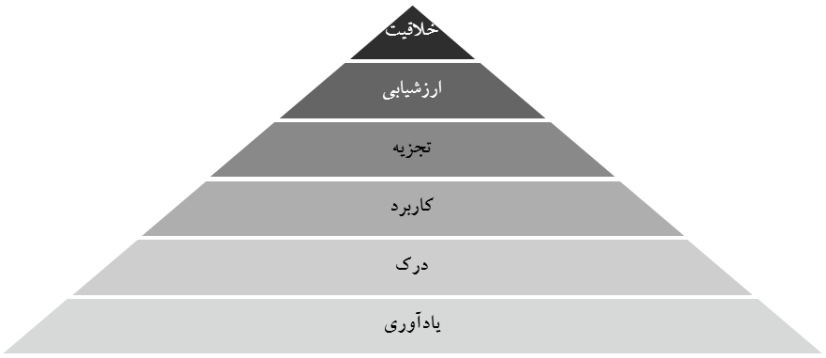
اسیدوز متابولیک و هایپرکالمی به عنوان عارضه جانبی کدام یک از دیورتیک‌های زیر است؟ پیش‌کارورزی اسفند ۷۷

الف) استازولامید ب) تیازید ج) فورزماید د) اسپیرونولاکتون

برای پاسخ به سؤال فوق لازم نیست که دانشجو درک عمیقی از اختلالات آب و الکترولیت داشته باشد یا در رابطه با یک سناریوی خاص وضعیت را تحلیل کند. سؤال در سطح یک یادآوری ساده از عوارض جانبی داروهاست که احتمالاً در کتاب فهرست شده‌اند.

در حالی که، برای پاسخ به سؤال زیر یادآوری ساده مطلب کافی نیست. دانشجو باید مطلب مربوط به تعادل آب و الکترولیت را درک کرده و فهمیده باشد. ولی در عین حال، سؤال ربطی به وضعیت بالینی این بیمار ندارد. بنابراین کاری که دانشجو باید انجام دهد، خیلی پیچیده نیست.

1. Recall
2. Comprehension
3. Application
4. Analysis
5. Evaluation



شکل ۲-۵: تاکسونومی بلوم

سطح شناختی: درک

در بیماری با آزمایشات زیر محتمل‌ترین تشخیص کدام است؟ دستیاری اسفند ۸۷

Glucose: 126 mg/dl	Cr: 1.2 mg/dl	HCO3: 14 meq/L	pH: 5.2
K: 3.3 meq/L	Na: 136 meq/L	Chloride: 113 meq/L	
Urine K: 10 meq/L	Urine Na: 32 meq/L	Urine Cl: 80 meq/L	

الف) اسهال ب) RTA ج) مصرف استازولامید

در سؤال زیر، دانشجو باید نوع اختلال را تشخیص دهد، فرمول مورد نظر برای اصلاح اختلال را بلد باشد و بتواند آن را در این موقعیت ویژه برای بیمار به کار برد:

سطح شناختی: کاربرد

سطح سرمی الکترولیت‌های بیمار به این صورت است:

Sodium 118 mmol/L Potasium 5.7 mmol/L Chloride 83 mmol/L

شما فکر می‌کنید که برای اصلاح تعادل الکترولیت‌ها، باید سدیم بیمار را به 130 mmol/L برسانید. با توجه به وزن نوزاد که یک کیلوگرم است، چه میزان سدیم لازم است که سطح سرمی سدیم به میزان مورد نظر برسد؟

الف) 12 mmol ب) 7/2 mmol ج)5/9 mmol د) 3/6 mmol

برای سنجش سطح ارزشیابی در تقسیم‌بندی بلوم، باید فرصتی برای دانشجو فراهم شود که مثلاً در آن برنامه تشخیصی یا درمانی خود را ارائه دهد. در مثال زیر، دانشجو باید رویکردهای درمانی محتمل را با یکدیگر مقایسه کند و بر اساس پیامدهای احتمالی هریک، مورد صحیح را انتخاب کند:

سطح شناختی: ارزشیابی

بیمار آقای ۶۰ ساله‌ای است که تنها زندگی می‌کند و با ضعف، بی‌حالی، بی‌اشتهایی و کاهش وزن به پزشک مراجعه کرده است. نتایج آزمایشات انجام شده به شرح زیر است:

WBC: 2000 RBC: 3.5 × 106 Hgb: 6 Plt: 80000 MCV: 108

بیمار تحت درمان با ویتامین B_{12} و اسید فولیک قرار می‌گیرد. ۴ روز پس از درمان دچار تشدید ضعف عضلانی شده است به گونه‌ای که نمی‌تواند راه برود. کدام یک از آزمایش‌های زیر اولویت دارد؟ دستیاری اسفند ۸۸

الف) CBC ب) اندازه‌گیری رتیکولوسیت‌ها

ج) اندازه‌گیری الکترولیت‌های سرم د) CT اسکن مغز

همان‌طور که مشاهده می‌شود، در سؤال با سطح شناختی بالا، اطلاعات به صورت خام و تفسیر نشده در اختیار دانشجو قرار می‌گیرد. خود دانشجو باید یافته‌های بالینی را کنار هم بگذارد، نتایج آزمایش‌ها را تفسیر کند و کار تجزیه و تحلیل اطلاعات را انجام دهد. به عنوان مثال، به جای اینکه ذکر شود: «بیمار، خانمی مسن است که به دلیل فشارخون بالا ارجاع شده است»، در پایه سؤال گفته می‌شود: «بیمار خانمی ۷۲ ساله با فشار خون ۱۸۰/۱۵۰ است». دانشجو باید تعریف و دامنه فشار خون طبیعی را بداند و نتیجه‌گیری کند که فشار خون بیمار بالا است.

سطح شناختی آخر در تقسیم‌بندی بلوم، خلاقیت است که نیازمند آن است که خود دانشجو با در کنار هم قرار دادن داده‌های مختلف، ایده جدیدی را مطرح کند. این سطح فقط با سؤال بازپاسخ سنجیده می‌شود نه سؤال بسته‌پاسخ که فهرست محدودی از گزینه‌ها را در اختیار دانشجو می‌گذارد.

نامطلوب بودن حدس زدن پاسخ

یکی از نگرانی‌هایی که همواره در مورد سؤالات چندگزینه‌ای وجود داشته است، حدسی جواب دادن است. به این معنا که دانشجو جواب درست را نمی‌داند و با حدس یک گزینه را انتخاب می‌کند. این مسأله از این نظر حائز اهمیت است که احتمال دارد دانشجویی که پاسخ صحیح را انتخاب کرده است، آن را حدس زده باشد بدون آنکه واقعاً دانش لازم را داشته باشد.

البته برخی معتقد هستند پاسخ حدسی را نمی‌توان به عنوان یک پیش‌فرض حتمی در امتحان چندگزینه‌ای در نظر گرفت. اینکه فکر کنیم سؤال چندگزینه‌ای بدون نمره منفی همیشه در معرض پاسخ حدسی است درست نیست. در یک مطالعه که آنالیز نمرات سه امتحان را توسط نظریه سؤال–پاسخ انجام داد، الگوی حدس سیستماتیک یافت نشد (اندره و ماگنانو[1] ۲۰۱۱).

در هر حال الگوی حدس زدن دانشجویان در برابر سؤال چندگزینه‌ای به یکی از این دو صورت است:

- در الگوی اول، فراگیر نسبت به هیچ یک از گزینه‌ها هیچ نظری ندارد و همه آنها برای وی علی‌السویه هستند. پس یکی از گزینه‌ها را به صورت تصادفی انتخاب می‌کند تا شاید جواب درست همان باشد و نمره به وی تعلق بگیرد. به این الگو، حدس زدن ناآگاهانه[2] یا تصادفی می‌گویند که در آن، تمام گزینه‌ها از احتمال برابر برای انتخاب شدن برخوردار هستند. به صورت نظری، در یک آزمون پنج گزینه‌ای، ۲۰ درصد احتمال دارد که به صورت حدسی گزینه درست انتخاب شود. بنابراین اگر دانشجو کل سؤالات را با حدس تصادفی جواب دهد، نمره او در نهایت ۴ از ۲۰ خواهد شد که باز هم به حدنصاب قبولی نمی‌رسد. به عبارت دیگر، باید در نظر گرفت که برای هر سؤالی که به صورت حدسی جواب داده می‌شود، ۲۰ درصد احتمال پاسخ صحیح وجود دارد و هم‌زمان ۸۰ درصد احتمال جواب نادرست وجود دارد. پس همان‌طور که در جدول ۳–۵ نشان داده شده است، برآیند سود احتمالی، در صورت وجود نمره منفی، برای یک سؤال که با حدس تصادفی پاسخ داده شده است، صفر خواهد شد.
- الگوی دوم هنگامی است که در نظر بگیریم حدس دانشجو همیشه به صورت کاملاً شانسی نیست بلکه دانشجو بر اساس دانش نسبی[3] خود برخی از گزینه‌ها را حذف یا انتخاب می‌کند. بنابراین توزیع پاسخ‌های درست به صورت تصادفی نخواهد بود (کرونباخ[4] ۱۹۸۴). در این حالت که حدس زدن آگاهانه[5] یا آموخته شده[6] نامیده می‌شود، فراگیر پاسخ صحیح را تا حدی می‌داند. به عبارت دیگر، با استفاده از استدلال منطقی یا با تکیه بر دانش نسبی خود در مورد موضوع سعی می‌کند با رد گزینه‌ها، جواب صحیح سؤال را حدس بزند و با درجاتی از اطمینان احتمال دهد که پاسخ

1. Andrà & Magnano
2. Blind guess, Wild guess
3. Partial knowledge
4. Cronbach
5. Informed guessing
6. Educated guessing

صحیح را انتخاب کرده است. گفتیم که همیشه حدس به منجر به افزایش نمره دانشجو نمی‌شود اما در حدس آگاهانه، آزمون شونده با در نظر گرفتن «احتمال» صحیح بودن یا نادرست بودن گزینه‌ها آنها را رد یا قبول می‌کند بنابراین سودی که عایدش می‌شود، بیشتر است. در جدول ۳-۵ میزان سودی که دانشجو با وجود نمره منفی از حدسی جواب دادن به سؤال پنج‌گزینه‌ای می‌برد، محاسبه شده است. همان‌طور که مشاهده می‌شود، مقدار افزایش نمره به میزان حدس آگاهانه دانشجو، یعنی تعداد گزینه‌هایی که می‌تواند با دانش نسبی خود حذف کند، بستگی دارد.

جدول ۳-۵: احتمال پاسخ به یک سؤال پنج گزینه‌ای بر اساس حدس آگاهانه و تصادفی، در صورت نمره منفی

میزان احتمال	حدس کاملاً تصادفی	حذف یک گزینه	حذف دو گزینه	حذف سه گزینه
درست حدس زدن	۲۰ درصد	۲۵ درصد	۳۳ درصد	۵۰ درصد
غلط حدس زدن	۸۰ درصد	۷۵ درصد	۶۷ درصد	۵۰ درصد
سود مورد انتظار	(۱×۲۰%) - (۲۵%×۸۰%) = ۰ نمره	(۱×۲۵%) - (۲۵%×۷۵%) = ۰/۰۶۲۵ نمره	(۱×۳۳%) - (۲۵%×۶۷%) = ۰/۱۶۶۷ نمره	(۱×۵۰%) - (۲۵%×۵۰%) = ۰/۳۷۵ نمره

تمام دانشجویان به یک میزان از حدس استفاده نمی‌کنند. در متون به عوامل متعددی اشاره شده است که بر حدس زدن موثر هستند:

- ویژگی‌های شخصیتی به خصوص میزان خطرپذیری افراد بر میزان حدس زدن موثر است. راه‌کارهایی مانند نمره منفی بیشتر از آن که تفاوت دانش و مهارت دانشجویان قوی و ضعیف را نشان دهند، نشانه تفاوت در میزان خطرپذیری آنها هستند. در واقع وقتی دانشجو به سؤالی جواب نداده است، مشخص نیست که آیا واقعاً پاسخ را بلد نبوده است یا از ترس تنبیه گزینه‌ای را انتخاب نکرده است. (موس[1] ۲۰۰۱، چاندراتیلکه و همکاران ۲۰۱۱).
- مشابه مورد قبلی، تاثیر جنسیت بر میزان حدس زدن مورد بحث است. گفته می‌شود خانم‌ها در ارائه پاسخ‌های حدسی نسبت به آقایان محتاطانه‌تر عمل می‌کنند (مک‌گیر[2] ۱۹۹۹) و به همین دلیل وجود نمره منفی موجب افزایش نمره آنها می‌شود. در مطالعه خطیبی و همکاران که از مدل ارزیابی اطمینان استفاده شده بود، میانگین نمرات دانشجویان مؤنث بیشتر بود. نویسندگان ذکر کردند این تفاوت می‌تواند ناشی از این باشد که آزمون مبتنی بر اطمینان میزان پاسخ حدسی را کاهش می‌دهد و دانشجویان مذکر به صورت کلی نسبت به دانشجویان مونث بیشتر از حدس استفاده می‌کنند (خطیبی و همکاران ۱۳۹۰). اما نتایج یک مطالعه تجربی دیگر تفاوت کوچکی بین نمرات دو جنس نشان داد. یعنی اگر چه تفاوت‌های جنسیتی در تمایل دانشجویان برای برای حدس زدن موثر است، اما احتمالاً تاثیر کمی روی نمره دارد (بن شاکر و سینایی ۱۹۹۱). مطالعه دیگری تفاوتی بین دو جنس نیافت (باند و همکاران[3] ۲۰۱۳).
- تفاوت‌های فرهنگی بر میزان حدس زدن موثر است. بر اساس نتایج مطالعه دیگری دانشجویان غیرانگلیسی‌زبان در مقایسه با دانشجویان انگلیسی‌زبان پاسخ‌های حدسی کمتری ارائه دادند (گلدیگ[4] ۲۰۰۸).
- سطح شناختی سؤال ممکن است بر میزان پاسخ‌های حدسی موثر باشد. بر اساس یافته‌های مطالعه‌ای که با استفاده از نظریه سؤال-پاسخ به آنالیز نمرات پرداخت، هنگامی که سطح شناختی سؤال بالا است، دانشجویان کمتر حدس می‌زنند (اندره و ماگنانو ۲۰۱۳).

1. Moss
2. McGuire
3. Bond et al.
4. Goldik

- موقعیت و شرایط آزمون و آزمون‌شوندگان نیز حدس زدن را تاثیر قرار می‌دهد؛ به عنوان مثال، استفاده از نمره منفی یا تعداد سؤالات آزمون بر میزان حدس زدن فراگیران تاثیر دارد. به همین دلیل است که گفته می‌شود باید از قبل به دانشجویان در مورد وجود یا عدم وجود نمره منفی اطلاع داده شود و دستورالعمل‌های شفاف در خصوص حدس زدن به آنها داده شود.

با توجه به سود کسب شده ناشی از حدس زدن، می‌توان گفت که یکی از دغدغه‌های اصلی متولیان آموزش و ارزیابی دانشجویان، پدیده حدس زدن در سؤالات چندگزینه‌ای است. پژوهشگران مختلف تلاش کرده‌اند تا با بررسی این پدیده توصیه‌هایی در مرد حدس زدن یا نزدن ارائه دهند اما در یک نگاه اجمالی به نظر می‌رسد که نتایج این مطالعات آن قدر پراکنده و متناقض هستند که در نهایت، تکلیف فراگیران و مدرسان را کاملاً روشن نمی‌کنند:

- برخی معتقد هستند ارزیابی‌های به عمل آمده از دانشجویان پزشکی نباید به گونه‌ای باشد که منجر به تشویق آنها در ارائه پاسخ های حدسی شود (هاردن[1] و همکاران ۱۹۷۶). چاپین[2] (۱۹۸۸) در مورد عواقب حدس زدن معتقد است که حدس، با وارد کردن یک متغیر تصادفی باعث کاهش روایی و پایایی می‌شود. همچنین، باعث افزایش نمره دانشجو می‌شود و از آنجا که میزان حدس زدن به میزان ریسک کردن دانشجویان وابسته است، در دو دانشجو با دانش مساوی، میزان حدس زدن متفاوت می‌شود و چون این موضوع ناعادلانه است، باید کلاً از آن جلوگیری کرد (پریهدا و همکاران[3] ۲۰۰۵).
- در نقطه مقابل، برخی پیشنهاد می‌کنند از آنجا که بین فراگیران مختلف از نظر میزان خطرپذیری تفاوت وجود دارد، بهترین راه برای حذف این تنوع، این دستورالعمل است که به هر سؤالی حتی با شک نیز پاسخ تا آزمون حالت تست شخصیت پیدا نکند و موضوعیت تست حفظ شود دهند (بودسکو و بارهیلل[4] ۱۹۹۳، دانینگ ۲۰۰۳).
- در عین حال، بر اساس نظر برخی از نویسندگان، توصیه قطعی مبنی بر دادن یا ندادن پاسخ‌های حدسی درست نیست و این امر بسته به شرایط آزمون می‌تواند متفاوت باشد. به عنوان مثال، تمیر[5] معتقد است هنگامی که سطح شناختی سؤال پایین است و حقایق و محفوظات با جواب‌های قطعی و حقیقی مطرح می‌شوند، باید از دانشجو خواست تا از حدس زدن خودداری کند. اما در ارزیابی سطوح شناختی بالا که معمولاً با عدم قطعیت همراه است و به استدلال و مقایسه و تجزیه و تحلیل نیاز است، حتی می‌توان فراگیران را به حدس زدن آگاهانه تشویق کرد (تمیر ۱۹۹۱).
- در همین راستا، هاموند و همکاران[6] با مطالعه‌ای که انجام دادند به این نتیجه رسیدند که به صورت کلی پاسخ به سؤال همیشه با دو فاکتور میزان دانش و میزان اطمینان از این دانش تعیین می‌شود که هر دو در فراگیران مختلف متفاوت است. بنابراین، نویسندگان پیشنهاد کردند توصیه کلی برای همه دانشجویان ارائه نشود و بهتر است هر فراگیر از قبل بهترین استراتژی را که برای خودش کاربرد دارد، شناسایی کند و در جلسه آزمون به کار گیرد (هاموند و همکاران ۱۹۹۸).

آنچه در این میان بر پیچیده بودن قضیه اضافه می‌کند و باعث می‌شود نتوان یک توصیه قطعی ارائه داد، این است که بسیاری از مطالعات ذکر شده در حیطه‌هایی غیر از پزشکی اجرا شده‌اند. در حوزه علوم پزشکی، جنبه دیگری نیز وجود دارد که نباید از آن غافل شد. این حقیقتی است که در پزشکی بسیاری از مسائل با دانش نسبی تصمیم‌گیری می‌شوند. در محیط واقعی طبابت هم پزشک هرگز جواب تمام سؤالات خود را در مورد بیمار به صورت قطعی نمی‌داند و در یک حالت عدم قطعیت تصمیم‌گیری می‌کند. در این حالت، کاهش حدس زدن نه لازم است و نه مطلوب؛ و حتی توصیه به حدس زدن

1. Harden
2. Choppin
3. Prihoda et al.
4. Budescu & Bar-Hillel
5. Tamir
6. Hammond et al.

آگاهانه می‌شود. راه‌کارهایی مانند نمره منفی، علاوه بر اینکه احتمال حدس زدن را کم می‌کند، احتمال حدس آگاهانه را نیز کم می‌کنند و اجازه نمی‌دهند که دانشجو بر اساس دانش نسبی و ناقص خود استدلال کند و جواب دهد. به اعتقاد اندرسون به همین دلیل است که نسل امروزی بسیار محتاط هستند و از ترس از دست دادن نمره خطر نمی‌کنند (اندرسون ۲۰۰۴).

نگرانی که در مورد حدس وجود دارد و متولیان ارزیابی را به سمت استفاده از راه‌کارهایی برای مقابله با آن سوق داده است، عمدتاً ناشی از تاثیری است که حدس بر افزایش نمرات دانشجویان و همچنین ویژگی‌های روان‌سنجی آزمون می‌گذارد. در قسمت «سؤالات رایج» به این موضوع خواهیم پرداخت که تاثیر به کارگیری راه‌کارهای مختلف مانند نمره منفی بر نمرات دانشجویان چطور است.

سؤالات رایج در مورد سؤالات بسته پاسخ

چه راه‌کارهایی برای مقابله با پاسخ حدسی به سؤال بسته‌پاسخ وجود دارد؟

از دیرباز برای کاهش میزان حدس زدن در سؤالات بسته‌پاسخ راه‌کارهایی مورد استفاده قرار گرفته است. هرچند با توجه به شرایط و ویژگی‌هایی که در قسمت پیشین در مورد حدس زدن گفته شد، در مورد استفاده یا عدم استفاده از آنها اختلاف نظر وجود دارد. در این قسمت، در مورد این راه‌کارها و اثرات آنها بحث می‌کنیم.

۱) استفاده از سؤالات چندپاسخی

یکی از روش‌هایی که سابقاً پیشنهاد شده بود، استفاده از سؤالات چندپاسخی بود. به این معنا که یک سؤال بیش از یک گزینه درست داشته باشد و برای کاهش حدس زدن، نمره سؤال بین گزینه‌های درست تقسیم شود و برای گزینه‌های نادرست، نمره منفی اعمال گردد. برای اطلاعات بیشتر در خصوص این نوع سؤال، به ابتدای همین فصل مراجعه کنید. امروزه با توجه به مشکلات این نوع سؤال امروزه دیگر مورد استفاده قرار نمی‌گیرد.

۲) کاهش خطاهای ساختاری سؤال

برخی معتقد هستند که بهترین راه برای به حداقل رساندن اثر حدس زدن، طراحی مناسب سؤالات است (دانینگ ۲۰۰۴). زیرا وجود خطاهای طراحی سؤال، به دانشجو برای یافتن گزینه صحیح سرنخ می‌دهد یا به او کمک می‌کند یک یا دو گزینه را حذف کند. به این ترتیب، مثلاً میزان شانس او از ۲۵ درصد به ۵۰ درصد می‌رسد و راحت‌تر می‌تواند از بین گزینه‌های باقی‌مانده به صورت تصادفی یک گزینه را انتخاب کند.

۳) افزودن تعداد گزینه‌ها

از آنجا که در آزمون سه‌گزینه‌ای، احتمال پاسخ درست بیشتر از آزمون چهارگزینه‌ای است، افزودن گزینه‌ها راه‌حل دیگری است که پیشنهاد شده است. کاراندیکار[1] معتقد است که ما نمی‌توانیم اثر حدس زدن تصادفی را حذف کنیم اما می‌توانیم با استفاده از استراتژی‌های مطلوب این اثر را به حداقل ممکن برسانیم. او در مقاله خود به افزایش تعداد گزینه‌های سؤال (مثلاً پنج گزینه) به عنوان یک استراتژی مطلوب اشاره می‌کند (کاراندیکار ۲۰۰۶). اما قطعاً طراحی چهار گزینه انحرافی مطلوب برای هر سؤال کار چندان آسانی نیست و اصولاً اینکه تعداد مطلوب گزینه‌ها چند تا باید باشد، چالش دیگری است که در ادامه همین فصل به آن خواهیم پرداخت.

1. Karandikar

۴) استفاده از نمره منفی

راه دیگر برای مقابله با حدس در نظر گرفتن نمره منفی برای پاسخ های نادرست است. در نمره‌دهی در حالت معمول، که به روش نمره‌دهی مثبت[1] یا نمره‌دهی به پاسخ‌های صحیح[2] معروف است، دانشجو به ازای هر پاسخ صحیح، یک واحد نمره دریافت می‌کند و به ازای سؤالاتی که جواب نداده یا غلط جواب داده است، صفر می‌گیرد. بنابراین، اگر دانشجو جواب سؤالی را نداند، می‌تواند به صورت تصادفی یکی از جواب‌ها را انتخاب کند. اگر جواب درست را انتخاب کند؛ یک نمره می‌گیرد ولی اگر پاسخ غلط را انتخاب کرده باشد، مشکلی پیش نمی‌آید. پس ترجیح می‌دهد حتی در صورت مطمئن نبودن از گزینه صحیح، جواب را حدس بزند.

در حالت دیگر که به نام نمره‌دهی منفی[3] شناخته می‌شود، اگر دانشجو به سؤالی درست جواب دهد، نمره می‌گیرد و اگر اشتباه جواب دهد، از او نمره کم می‌شود. به این ترتیب، انتخاب‌های نادرست جریمه می‌شوند و دانشجو به راحتی نمی‌تواند از حدس زدن استفاده کند. از آنجا در این مدل، نمره سؤالات بر اساس فرمول محاسبه می‌شوند، به آن نمره‌دهی با فرمول[4] هم می‌گویند.

این فرمول‌ها متنوع هستند و اینکه چه میزان نمره باید از دانشجو کم شود، همیشه محل بحث بوده است. برخی می‌گویند اگر واقعاً می‌خواهیم نمره منفی، تاثیرگذار باشد و دانشجویان ضعیف و قوی را از هم متمایز کنیم، حتما باید سهم نمره‌ای که برای آن در نظر گرفته می‌شود، به اندازه کافی بزرگ باشد تا به صورت جریمه عمل کند. یکی از انواع رایج آن که برای مقابله با حدس تصادفی به کار می‌رود، به این صورت است که اگر تعداد گزینه را M در نظر بگیریم، به ازای هر جواب غلط $\frac{۱}{M-۱}$ نمره کسر خواهد شد. به عنوان مثال، در آزمون چهارگزینه‌ای، نمره $\frac{۱}{۳}$ منفی و در آزمون پنج گزینه‌ای، $\frac{۱}{۴}$ نمره منفی خواهیم داشت که در این حالت، جمع نمره کل در صورت انتخاب تصادفی تمام سؤالات، صفر خواهد شد. فرمول‌های دیگری نیزی وجود دارد که میزان جریمه در آنها بیشتر است مثلاً: یک نمره مثبت (+۱) برای پاسخ صحیح، یک نمره منفی (−۱) برای پاسخ غلط و صفر (۰) برای موارد بدون پاسخ یا گزینه «نمی‌دانم». منطق این نحوه محاسبه به همان موضوع برمی‌گردد که میزان جریمه باید آنقدر بزرگ باشد تا واقعاً دانشجویان از انتخاب گزینه‌ها به صورت حدسی (حتی حدس آگاهانه) اجتناب کنند.

آزمون‌های مهم دنیا در خصوص استفاده از نمره منفی، رویکرد یکسانی در پیش نگرفته‌اند. مثلاً در آزمون گواهینامه پزشکی در امریکا[5] و آزمون گواهینامه نظام پزشکی کانادا[6] از نمره منفی استفاده نمی‌شود. در حالی که در آزمون CMS[7] هند، هر پاسخ اشتباه به سؤال چهار گزینه‌ای، به اندازه $\frac{۱}{۳}$ نمره منفی دارد. همچنین، در اسپانیا در آزمون MIR[8] برای هر سؤال پنج گزینه‌ای به اندازه $\frac{۱}{۴}$ نمره منفی منظور می‌گردد. حتی تصمیات داخل یک کشور نیز یکسان نیستند. مثلاً در دو آزمون SAT و ACT[9] که هر دو برای پذیرش در کالج‌های امریکا مورد استفاده قرار می‌گیرند، اولی نمره منفی دارد در حالی که دومی ندارد. در ایران نمره منفی در کنکور سراسری پذیرش دانشگاه و کنکور پذیرش دستیار تخصصی وجود دارد.

۵) استفاده از گزینه «نمی‌دانم»

روش دیگر، در نظر گرفتن گزینه «نمی‌دانم» در کنار سایر گزینه‌ها است تا دانشجو در صورتی که نتواند جواب درست را شناسایی کند، از آن استفاده کند. در این حالت، نمره کلی دانشجو تعداد پاسخ‌های صحیح منهای تعداد پاسخ‌های غلط است. یعنی

1. Positive marking
2. Right scoring
3. Negative marking
4. Formula scoring
5. United States Medical Licensing Examination (USMLE)
6. Medical Council of Canada Qualifying Examination (MCCQE)
7. Combined Medical Services
8. M´edico Interno Residente
9. American College Testing

به ازای هر پاسخ صحیح یک نمره به او تعلق می‌گیرد و به ازای هر پاسخ اشتباه یک نمره از او کم می‌شود. همان‌طور که مشخص است، به نوعی زیر مجموعه روش نمره منفی محسوب می‌شود که در آن مقدار جریمه بزرگ‌تر است.

در مدل نمره‌دهی مثبت که به صورت معمول استفاده می‌شود، چون دانشجو به ازای جواب غلط نمره منفی دریافت نمی‌کند، به تمام سؤالات پاسخ می‌دهد که ممکن است با دانش کامل باشد، یا با دانش نسبی (حدس آگاهانه) و یا بدون دانش (حدس کاملاً تصادفی). اما در امتحانی که گزینه «نمی‌دانم» دارد، دانشجو برخی از سؤالات را بی جواب باقی می‌گذارد. لرد[1] (۱۹۷۵) معتقد بود که گزینه «نمی‌دانم»، با دانش نسبی و ناکامل کاری ندارد و فقط قرار است حدس‌های تصادفی را کاهش دهد. بنابراین باید به دانشجو اعلام کرد تنها در صورتی از آن استفاده کند که هیچ اطلاعی در مورد جواب ندارد. اما از طرفی برخی معتقد هستند هنگامی که فراگیر می‌داند با گزینه «نمی‌دانم» مواجه می‌شود، باید برای هر سؤال حتی به صورت اجمالی به بررسی دانش خود بپردازد. این مسأله به او کمک می‌کند تا محدودیت دانش خود را بهتر شناسایی کند. در حالی که در یک آزمون معمولی، این توانمندی در او تقویت نمی‌شود. همچنین بررسی گزینه‌های «نمی‌دانم» به استادان نیز اطلاعات خوبی در خصوص کاستی‌ها و نقائص دوره می‌دهد. حتی به طراحان سؤال کمک می‌کند مطالبی را که تدریس نشده‌اند یا جزء محتوای اصلی دوره معرفی نشده بودند اما از آنها سؤال طراحی شده، تشخیص دهند (مویجنس و همکاران ۱۹۹۹).

یکی از مشکلات استفاده از این روش این است که اکثراً حتی دانشجویان خوب نیز تمایل ندارند که با قطعیت یک گزینه را مشخص کنند. بنابراین به سمت انتخاب گزینه «نمی‌دانم» جلب می‌شوند که این موضوع می‌تواند به کاهش نمره آنها منجر شود (گلدیک ۲۰۰۸).

۶) مدل ارزیابی اطمینان

در رویکرد دیگر که به مدل ارزیابی اطمینان[2] معروف است، هم به نوعی از نمره منفی استفاده می‌شود اما در این روش علاوه بر اطلاعات و دانش فرد، به میزان اطمینان وی از پاسخی که می‌دهد نیز توجه می‌شود. بنابراین در زیرمجموعه نمره‌دهی ارزش نسبی[3] قرار می‌گیرد (فراری[4] ۱۹۸۹). به این صورت که از دانشجو خواسته می‌شود همزمان با پاسخگویی به هر سؤال، اعلام کند که از جواب خود به چه میزان مطمئن بوده است. سپس برای پاسخ‌های غلطی که دانشجو از آنها مطمئن بوده است، نمره منفی و برای پاسخ‌های درستی که دانشجو از آنها مطمئن بوده است نمره مثبت در نظر گرفته می‌شود (گاردنر–مدوین و کورتون[5] ۱۹۹۶). یک نمونه از این نحوه نمره‌دهی در جدول ۴–۵ نشان داده شده است. در این روش، دانشجو برای پاسخ‌های اشتباهی که با اطمینان داده است، جریمه مضاعف دریافت می‌کند. زیرا این طور نبوده است که دانشجو جواب را بلد نباشد بلکه اساساً در آن زمینه اطلاعات نادرست و غلطی داشته است.

مزیت این روش نسبت به نمره منفی در این است که اولاً دانشجوی پزشکی باید بتواند اطلاعات خود را ارزیابی کند و نقص اطلاعات خود را بشناسد. حتی اگر سؤالی را با حدس جواب می‌دهد، باید به این قضیه آگاه باشد. این وضعیت در بالین بیمار احتمالا به این شکل قابل ترجمه خواهد بود که «آیا من از این موضوع (مثلاً عوارض یک داروی خاص) مطمئن هستم یا باید به کتاب مراجعه کنم؟». پزشکی که از اطلاعات اشتباه خود مطمئن است، به مراتب خطرناک‌تر از پزشکی است که مطلبی را نمی‌داند یا فراموش کرده است (بندر ۲۰۰۳). ثانیاً میزان مهارت دانشجو در حدس موفق را مشخص می‌کند. یعنی مشخص می‌شود که آیا فراگیر، حدس‌زننده خوبی هست یا خیر. پس از آزمون که دانشجو برگه خود را با کلید سؤالات مقایسه می‌کند، علاوه بر اینکه در می‌یابد از چه مطالبی نامطمئن بوده تا بتواند بیشتر بر آنها تمرکز کند، می‌تواند با مرور پاسخ‌های خود مهارت حدس آگاهانه را در خود تقویت کند.

1. Lord
2. Confidence assessment
3. Partial-credit scoring
4. Frary
5. Gardner-Medwin & Curton

جدول ٤-۵: روش ارزیابی اطمینان برای نمره‌دهی

درجه اطمینان از پاسخ	نمره برای پاسخ درست	نمره برای پاسخ غلط
نامطمئن	+۱	صفر
تقریبا مطمئن	+۲	-۲
اطمینان کامل	+۳	-۶

۷) حذف گزینه

در روش حذف گزینه[1] که زیر مجموعه نمره‌دهی ارزش نسبی است، از دانشجو خواسته می‌شود به جای تشخیص پاسخ صحیح تا جایی که می‌تواند گزینه‌های غلط را حذف کند. در واقع دانشجو در مواجهه با سؤالی که از پاسخ صحیح آن مطمئن نیست، با دانش نسبی خود به حذف گزینه اقدام می‌کند.

مدل‌های متفاوتی برای نمره‌دهی این روش در نظر گرفته شده است. یکی از انواع آن به این صورت است که اگر دانشجو هیچ گزینه‌ای را حذف نکرده باشد، صفر می‌گیرد (بدون دانش). به ازای حذف هر گزینه غلط یک نمره می‌گیرد (دانش نسبی). اگر همه گزینه‌های غلط را حذف کرده باشد، نمره کامل را می‌گیرد (دانش کامل). اگر گزینه صحیح بین گزینه‌هایی باشد که توسط دانشجو حذف شده است، نشانه اطلاعات کم و بیش غلط او است و اگر فقط یک گزینه را که همان گزینه صحیح است حذف کرده باشد، ۴ نمره از او کم می‌شود (دانش غلط). یعنی نمره دانشجو در یک سؤال بین ۴- تا ۳+ است (باند و همکاران ۲۰۱۳). نوع دیگری از نمره‌دهی است که بین ۳- تا ۴+ می‌تواند به دانشجو تعلق بگیرد (لاو و همکاران[2] ۲۰۱۱).

موافقان معتقدند این روش مشابه موقعیت‌های واقعی است که اکثراً در رابطه با مسائلی که قرار است تصمیم‌گیری کنیم، دانش کامل نداریم. این روش کمک می‌کند بین دانش کامل و نسبی دانشجو افتراق دهیم. در واقع، حدس تصادفی را از بین می‌برد اما به دانش نسبی پاداش می‌دهد (لاو و همکاران ۲۰۱۱، باند و همکاران ۲۰۱۳).

۸) انتخاب آزاد

روش انتخاب آزاد[3] زیر مجموعه نمره‌دهی ارزش نسبی است در این روش فراگیر مجاز است در صورت تمایل و مطمئن نبودن از پاسخ صحیح، بیش از یک گزینه را انتخاب کند (جنینگز و بوش[4] ۲۰۰۶). نمره‌دهی به این شکل است که به ازای انتخاب گزینه صحیح در یک سؤال چهار گزینه‌ای، ۱ نمره دریافت می‌کند و به ازای انتخاب هر گزینه انحرافی، یک سوم نمره از او کم می‌شود. گفته می‌شود که این روش هم جلوی حدس تصادفی را می‌گیرد و به تشخیص تفاوت بین دانش کامل و دانش نسبی دانشجو کمک می‌نماید (جنینگز و بوش ۲۰۰۶).

تاثیر راه‌کارهای فوق بر آزمون و میزان موفقیت آنها در کاهش حدس زدن در بخش‌های بعدی مورد بحث قرار می‌گیرد. اما توجه به این نکته ضروری است که استفاده از فرمول‌های نمره‌دهی منجر به اصلاح تاثیر حدس زدن بر آزمون نمی‌شوند، بلکه هدف اصلی استفاده از آنها، کاهش تشویق فراگیران به حدس زدن است. بنابراین زمانی تاثیرگذار هستند که از قبل به اطلاع دانشجویان رسیده باشند (برتون ۲۰۰۲).

1. Elimination Testing
2. Lau et al.
3. Liberal
4. Jennings & Bush

راهکارهای مقابله با حدس زدن چه تاثیری دارند؟

باور رایج این است که استفاده از راه‌کارهایی که عنوان شدند، از وقوع حدس زدن جلوگیری می‌کند و به همین دلیل به نظر بسیاری مطلوب است. اما به نظر می‌رسد که تاثیر آنها در کاهش میزان حدس زدن، به این سادگی هم نیست. پژوهشگران تلاش کرده‌اند تا از جنبه‌های مختلف این پدیده را بررسی کنند. باید دقت کرد که برخی از نظرات نشأت گرفته از مطالعات تجربی هستند و برخی دیگر به فرضیه‌های نظری اشاره دارند. حتی در تفسیر نتایج مطالعات تجربی نیز باید احتیاط کرد زیرا متدولوژی، زمینه و پیامدهای مورد بررسی آنها متفاوت بوده است. در اینجا اثرات راه‌کارهای مذکور را بر پیامدهای زیر بررسی می‌کنیم.

۱) پایایی آزمون

اعتقاد بر این است که حدس تصادفی پایایی آزمون را کاهش می‌دهد (برتون ۲۰۰۱). هر چند که مطالعات نظری اولیه موافق بودند که با استفاده از راه‌کارهای ذکر شده، روایی (لرد ۱۹۶۳) و پایایی (ماتسون[1] ۱۹۶۵) آزمون زیاد می‌شود اما مطالعات تجربی به نتایج متناقض رسیدند: برخی گزارش کردند که متعاقب استفاده از نمره منفی، پایایی آزمون کاهش داشته است و برخی دیگر ذکر کردند که افزایش پایایی مشاهده شده است، هرچند که گاهی مقدار افزایش‌یافته ناچیز بود (رولی و تراب[2] ۱۹۷۷، موجنس و همکاران ۱۹۹۹).

از جمله انتقادات وارده به نتایج برخی از این مطالعات تجربی این بود که آنالیزها با رویکرد گذشته‌نگر انجام شده بود و در آنها، دانشجویان قبل از آغاز پاسخگویی به سؤالات در مورد نحوه نمره‌دهی توجیه نشده بودند. در حالی که نتایج یک مطالعه نشان داده است اطلاع‌رسانی قبلی مشخصاً بر پایایی آزمون اثر می‌گذارد (دیاموند و ایوانز[3] ۱۹۷۳). بنابراین مطالعات دیگری با متدلوژی قوی‌تر انجام شد اما کماکان نشان ندادند که پایایی آزمون افزایش قابل توجه می‌یابد (کراس و فراری[4] ۱۹۷۷، بلیس ۱۹۸۰).

مطالعه‌ای مروری که روی داده‌های سه مطالعه تجربی انجام شد، به این نتیجه رسید که نمره‌دهی معمولی اثر بیشتری روی پایایی دارد تا فرمول نمره منفی که در آن دانشجویان محتاط‌تر عمل می‌کنند (برتون ۲۰۰۲). همچنین افزایش پایایی آزمون در استفاده از گزینه نمی‌دانم (موجنس و همکاران ۱۹۹۹) و روش حذف گزینه (لاو و همکاران ۲۰۱۱) گزارش شد.

مسأله دیگری که بر پایایی اثر دارد این است که میزان حدس زدن دانشجویان لزوماً در همه موارد یکسان نیست. در یک مطالعه یک گروه از دانشجویان مجاز به استفاده از حدس بودند اما به گروه دیگر گفته شد پاسخ را حدس نزنند. در گروه اول استفاده از نمره منفی، پایایی آزمون را زیاد کرد اما در گروه مقابل با استفاده از نمره منفی، پایایی تغییری نکرد (راچ و دگراف[5] ۱۹۲۶).

۲) روایی آزمون

در مورد روایی، همان‌طور که قبلا گفته شد چون میزان حدس زدن متاثر از میزان خطرپذیر بودن افراد است، با محاسبه نمره منفی مانند این است که نمرات دانشجویان علاوه بر دانش آنها از شخصیت آنها نیز تاثیر پذیرفته باشد. پس به طور نظری می‌توان گفت که با وارد شدن یک سازه دیگر و ایجاد «واریانس بی‌ارتباط به سازه»، روایی آزمون ممکن است مخدوش شود (داونینگ ۲۰۰۳).

در یک مطالعه تجربی، چون توافق بین آزمون تشریحی و تستی در حالت نمره منفی به طور معنادار بهتر از توافق آنها در صورت نمره‌دهی معمولی بود، نویسندگان نتیجه‌گیری کردند که روش احتساب نمره منفی روایی بهتری دارد. (پریهدا

1. Mattson
2. Rowley & Traub
3. Diamond & Evans
4. Cross & Frary
5. Ruch & Degraff

و همکاران ۲۰۰۶). البته این مطالعه گذشته‌نگر بوده و به دانشجویان در مورد وجود یا عدم وجود نمره منفی چیزی گفته نشده بود. در حالی که آگاهی دانشجو بر رفتار او هنگام حدس زدن تاثیر دارد (چمبرز[1] ۲۰۰۷).

پریهدا و همکاران ۲۰۰۵

از درس پاتولوژی ماگزیلوفاسیال در دانشجویان سال دوم رشته دندانپزشکی چهار امتحان دو ساعته گرفته شد. یک ساعت اول هر امتحان، ۲۵ سناریو بود که به ازای هر یک، دو سؤال کوته‌پاسخ عنوان شد (جمعا ۵۰ سؤال). پس از اتمام زمان آزمون برگه ها جمع شد. ساعت دوم آزمون شامل ۵۰ سؤال چندگزینه‌ای بود. نمره حاصل از دو قسمت تستی و تشریحی به صورت جداگانه محاسبه شد. نمرات آزمون تستی به دو صورت معمولی و با فرمول نمره منفی محاسبه شد (کسر ۱/۴ نمره به ازای هر جواب غلط). برای سنجش میزان توافق نمرات تستی و تشریحی از ضریب interclass correlation استفاده شد. توافق بین نمرات تشریحی با نمرات تستی همراه با احتساب نمره منفی به طور معناداری بیشتر از توافق بین نمرات تشریحی با نمرات تستی معمولی بود. هرچند نویسندگان ادعایی در این مورد نداشتند که سؤالات تشریحی لزوماً دانش فراگیران را بهتر می‌سنجند اما با این فرض که امتحانی که بتواند اثر حدس زدن را کاهش دهد، احتمالا شاخص مناسب‌تری برای سنجش دانش واقعی افراد است، نتیجه‌گیری کردند که روش احتساب نمره منفی روایی بهتری دارد.

۳) نمره دانشجویان

یکی از انتقاداتی که به پاسخ حدسی شده است، تاثیر آن در افزایش نمره دانشجو است. از دیدگاه نظری، دانشجو در مواجهه با سؤال چندگزینه‌ای به صورت حدسی به سؤالات پاسخ می‌دهد و نمرات بالاتری کسب خواهد کرد. بنابراین این باور عمومی وجود دارد که اگر از راه‌کارهایی مانند نمره منفی و ... استفاده شود، نمره کاهش می‌یابد یا حداقل به صورت کاذب زیاد نمی‌شود و هر دانشجو نمره واقعی خود را دریافت می‌کند. زیرا در صورت وجود نمره منفی دانشجو با احتیاط بیشتری رفتار می‌کند و دانش نسبی خود را کنار می‌گذارد. در نتیجه با بی‌جواب گذاشتن برخی از سؤالات که در صورت نمره‌دهی معمولی به آنها پاسخ می داد، نمره پایین‌تری کسب می‌کند. به این موضوع تاثیر افتراقی[2] می‌گویند. در حالی که عده دیگری معتقد هستند اثر این پدیده معادل اثر شانس است یعنی در صورت وجود یا عدم وجود نمره منفی، تغییری در میزان پاسخ دادن دانشجو ایجاد نمی‌شود. از این اثر تحت عنوان تاثیر تغییرناپذیری[3] یاد می‌شود (انگوف و اشرادر[4] ۱۹۸۴).

نتایج مطالعات تجربی در این باره یکسان نیست: نتایج برخی از مطالعات به نفع عدم تغییر بود یعنی استفاده از فرمول، تاثیر منفی بر رفتار دانشجویان ندارد (انگوف و اشرادر ۱۹۸۴). حتی در مطالعه تجربی دیگری مشخص شد که اگر افراد فقط بر اساس اطمینان کامل به سؤالات جواب می‌دادند، تنها قادر بودند نمره‌ای نزدیک به نمره قبولی دریافت کنند. در حالی که اگر سؤالاتی را که بی‌جواب گذاشته بودند، با حدس کاملاً تصادفی پاسخ می‌دادند، ازای هر ۱۰ سؤال، ۲ نمره افزایش پیدا می‌کردند و در صورت استفاده از حدس آگاهانه، مقدار افزایش به ازای هر ۱۰ سؤال، به ۵ نمره می‌رسید (هاموند و همکاران ۱۹۹۸).

هاموند و همکاران ۱۹۹۸

یک آزمون با ۳۰ سؤال ۵ گزینه‌ای در یک دوره فلوشیپ بیهوشی انتخاب شد که شرکت کنندگان برای هر گزینه یکی از این موارد را علامت می زدند: صحیح/غلط/نمی‌دانم. علاوه بر این، میزان اطمینان خود در هر سؤال را به این صورت مشخص می‌کردند: کاملاً مطمئن، حدس آگاهانه، کاملاً تصادفی.

دانشجویان به ۵۰ درصد کل سؤالات با اطمینان کامل جواب داده بودند که از این میزان پاسخ ۸۹/۲ درصد درست بود. حدس آگاهانه در ۷۴/۴ درصد موارد منجر به پاسخ‌های صحیح شده بود و در نهایت، ۸/۸ درصد موارد کاملاً به صورت حدسی پاسخ داده شده بودند که در ۶۵/۵ درصد موارد منجر به پاسخ درست شده بود.

پس اگر افراد فقط بر اساس اطمینان کامل به سؤالات جواب می‌دادند، فقط قادر بودند نمره‌ای نزدیک به نمره قبولی دریافت کنند. در حالی که اگر شرکت‌کنندگان سؤالاتی را که بی‌جواب گذاشته بودند (گزینه نمی‌دانم را انتخاب کرده بودند)، با حدس «آگاهانه» پاسخ می‌دادند، به ازای هر ۱۰ سؤال، ۵ نمره افزایش پیدا می‌کردند. در صورت استفاده از حدس کاملاً تصادفی، مقدار افزایش به ازای هر ۱۰ سؤال، به دو نمره می‌رسید.

1. Chambers
2. Differential effect
3. Invariance effect
4. Angoff and Schrader

با این حال، در چند مطالعه دیگر هنگامی که نمره منفی وجود داشت، نمره دانشجویان پایین‌تر بود و سؤالات بیشتری بدون جواب باقی مانده بودند (آلبنز ۱۹۸۸، بتس و همکاران[1] ۲۰۰۹، کیمیایی ۲۰۱۲). در یک مطالعه تجربی دیگر روی دو گروه دانشجو مشخص شد که اقدام به حدس زدن، علی‌رغم وجود نمره منفی، نمره دانشجویان را زیاد می‌کند (انگوف ۱۹۸۹).

انگوف و اشرادر ۱۹۸۱

پژوهشگران در این مطالعه دو آزمون انتخاب کردند: آزمون استعداد تحصیلی (SAT) با شرکت ۶۲۶۰ دانش‌آموز سال آخر دبیرستان (۱۷ ساله) و امتحان شیمی پیشرفته با حضور ۲۳۰۶ دانش‌آموز. هر یک از آزمون‌ها دو بار اجرا شد و در هر یک، نمره‌دهی با هر دو روش مثبت و منفی انجام شد تا مشخص شود آیا این روش‌ها تاثیری در عملکرد دانشجویان و نتیجه آزمون دارند یا خیر. یافته‌ها نشان داد که در آزمون پیشرفت تحصیلی که یک آزمون با محدودیت زمان بود، نمرات دانش‌آموزان با در نظر گرفتن نمره منفی پایین‌تر بود اما اختلاف معنادار نبود و لزوماً ترتیب دانش‌آموزان را تغییر نداد. اما دانش‌آموزان قوی در صورت وجود نمره منفی، بیشتر از دانشجویان ضعیف از حدس زدن سود برده بودند. در آزمون شیمی که از نظر زمانی محدودیت نداشت، تغییر روش تصحیح، تغییری در نمرات ایجاد نکرد.

در مورد استفاده از گزینه «نمی‌دانم» نیز برخی معتقد هستند استفاده از آن، همانند نمره منفی، ممکن است باعث شود که دانشجویان محتاط‌تر عمل کنند و حتی در صورت داشتن دانش نسبی، تلاشی برای انتخاب گزینه صحیح ننمایند. حتی دانشجویان خوب نیز اکثراً تمایل ندارند که با قطعیت یک گزینه را مشخص کنند. بنابراین به سمت انتخاب گزینه «نمی‌دانم» جلب می‌شوند که این موضوع می‌تواند به کاهش نمره آنها منجر شود (گلدیک ۲۰۰۸). مطالعاتی در این رابطه انجام شده است که از دانشجویان خواسته‌اند یک بار با وجود گزینه «نمی‌دانم» و بار دیگر با چشم‌پوشی کردن از آن به سؤالات پاسخ دهند. در یک مطالعه مشخص شد که در حالت وجود گزینه «نمی‌دانم»، نمرات دانشجویان به طور معنی‌داری افزایش یافته بود (هاردن و همکاران ۱۹۷۶) اما در مطالعه دیگری با استفاده از گزینه «نمی‌دانم»، نمره دانشجویان کاهش یافته بود و این تفاوت معنادار بود (موجنس و همکاران ۱۹۹۹).

موجنس و همکاران ۱۹۹۹

این مطالعه برای مقایسه مدل نمره‌دهی معمولی با مدلی که در آن گزینه «نمی‌دانم» وجود داشت، صورت گرفت. دو بلوک مولتی‌دیسیپلینری رشد و درد در انتهای سال دوم و سوم دوره پزشکی دانشگاه ماستریخ انتخاب شدند که به ترتیب ۱۶۱ و ۱۶۹ سؤال «درست-نادرست» داشتند. از دانشجویان خواسته شد تا یک بار هر سؤال را به صورت بله/خیر/نمی‌دانم جواب دهند و سپس یک بار دیگر برای تمام سؤالاتی که «نمی‌دانم» را انتخاب کرده بودند، یکی از گزینه‌های بله یا خیر را انتخاب کنند. در ۲۷/۱ درصد و ۲۲/۶ درصد موارد دو آزمون، گزینه «نمی‌دانم» انتخاب شده بود که در پاسخگویی دوم، وقتی تمام این موارد به بله یا خیر تبدیل شد، بیش از نیمی از موارد (به ترتیب ۴۵/۷ درصد و ۵۶/۷ درصد موارد) درست بود. این امر به این معناست که با استفاده از گزینه «نمی‌دانم»، نمره کلی دانشجویان در هر دو آزمون کاهش یافته بود. پایایی آزمون که با آلفای کرونباخ مورد سنجش قرار گرفت، در حالت بله/خیر/نمی‌دانم بیشتر بود (۰/۷۲ در برابر ۰/۶۶ و ۰/۷۴ در برابر ۰/۶۶). این فرضیه که در نظر گرفتن گزینه «نمی‌دانم» بیشتر دانشجویان قوی را تحت تاثیر قرار می‌دهد، تایید نشد.

در مطالعه‌ای که پژوهشگران روش «حذف گزینه» را با نمره منفی مقایسه کردند، به این نتیجه رسیدند که دانشجویان از روش حذف گزینه سود می‌برند. بر اساس نظرسنجی که از دانشجویان شد با روش حذف گزینه استرس کمتری داشتند، معتقد بودند با این روش تفکر نقاد در آنها تقویت می‌شود و به صورت کلی آن را ترجیح می دادند (باند و همکاران ۲۰۱۳).

در مورد «روش آزاد» باید گفت که مقایسه آن با روش نمره‌دهی معمولی در یک مطالعه نظری نشان داد که اگر دانشجو قادر باشد حداقل دو گزینه انحرافی را به درستی تشخیص دهد، در حالت آزاد نمره‌ای بالاتر از نمره‌دهی معمولی کسب می‌کند. روش آزاد می‌تواند با ترغیب دانشجو به استفاده از دانش نسبی، نمره وی را افزایش دهد اما در دانشجویی که اطلاعات ندارد و با حدس تصادفی پاسخ می‌دهد یا اطلاعات غلط دارد، باعث کسر نمره می‌شود (جنینگز و بوش ۲۰۰۶).

مساله دیگر، تاثیر اطلاع‌رسانی قبلی روی نمرات است. نتایج مطالعه فراری روی شش روش نمره‌دهی نشان داد که ارائه اطلاعات و دستورالعمل غلط هنگام امتحان در مورد حدس زدن و راه‌کارهای آن در مقایسه با مواردی که دستورالعمل‌ها

1. Betts et al.

نادیده گرفته می‌شوند، منجر به نمرات پایین‌تر می‌شود (فراری ۱۹۸۰). در مطالعه بتس آن دسته از دانشجویان که به آنها گفته شده بود نمره منفی حساب نمی‌شود، به طور معنادار عملکرد بهتری داشتند نسبت به دانشجویانی که به آنها گفته شده بود امتحان نمره منفی دارد (بتس و همکاران ۲۰۰۹).

۴) نمره دانشجویان قوی و ضعیف

در مورد این مسأله که با استفاده از راه‌کارهای کاهش حدس، دانشجویان قوی و ضعیف چقدر ضرر یا سود می‌کنند، نیز مطالعاتی انجام شده است. بر اساس نتایج یک مطالعه با محاسبه نمره منفی، دانشجویان قوی‌تر بیشتر از دانشجویان ضعیف ضرر می‌کنند. احتمالاً چون دانشجویان قوی به دستورالعمل آزمون مبنی بر حدس نزدن بیشتر توجه می‌کنند. (بلیس ۱۹۸۰). اما سود حاصل از حدس زدن در امتحانی که نمره منفی دارد، بیشتر نصیب دانشجویان قوی‌تر می‌شود و دانشجویان ضعیف با حدس زدن ضرر می‌کنند (انگوف ۱۹۸۹). احتمالاً به این علت که دانشجویان قوی‌تر حتی در مواردی که از جواب درست کاملاً مطمئن نیستند، باز هم درجه اطمینانشان از دانشجویان ضعیف بیشتر است.

شاید بتوان این‌گونه برداشت کرد که در دانشجویان با توانایی بالا، مانند دانشجویان علوم پزشکی، وجود نمره منفی مشکل زیادی ایجاد نمی‌کند. بر اساس یافته‌های مطالعه دیگری، با نمره منفی از حدس زدن افراد نامطمئن جلوگیری می‌شود؛ در حالی که کسی به ناحق ضرر نمی‌کند. نمره منفی بیشتر از اینکه بر عملکرد دانشجویان قوی اثر بگذارد و نمره آنها را تغییر دهد، بر رفتار افراد نامطمئن اثر می‌گذارد و آنها هستند که جریمه می‌شوند (موجنس و همکاران ۱۹۹۹).

۵) سایر پیامدها

به جز ویژگی‌های روان‌سنجی مانند روایی و پایایی و مقدار نمره، پیامدهای دیگری هم هستند که ممکن است در برخی از شرایط اهمیت پیدا کنند:

پمپلت و فارنیل[1] در مطالعه خود نشان دادند نمره منفی موجب دقیق‌تر شدن نتایج آزمون شد و در عین حال، سطح اضطراب دانشجویان را افزایش نداد (پمپلت و فارنیل ۱۹۹۵) اما برادی معتقد بود مقدار نمره منفی در نظر گرفته شده در این مطالعه خیلی بزرگ نبوده است و به همین دلیل استرس دانشجویان را افزایش نداده است. اگر سهم نمره منفی را بزرگ در نظر بگیریم تا به صورت موثر عمل کند، این کار به احتمال زیاد باعث استرس و اضطراب دانشجویان می‌شود و در این صورت نمره به دست آمده، نشان‌دهنده عملکرد واقعی آنها نیست (برادی[2] ۲۰۰۵).

پیامد دیگر اینکه نمره منفی زمان آزمون را طولانی می‌کند. زیرا فراگیران در امتحانی که نمره منفی دارد، کندتر عمل می‌کنند. این موضوع در آزمون‌های با محدودیت زمانی خود را بهتر نشان می‌دهد و در عملکرد افراد بیشتر تاثیر دارد (انگوف و اشرادر ۱۹۸۱).

مسأله آخر اینکه علی‌رغم این ادعا که نمره منفی حدس زدن دانشجویان را کاهش می‌دهد اما در واقع از جنس جریمه است. بر اساس نظریه‌های رفتارگرایی[3] پاداش می‌تواند موجب تقویت دائمی یا حداقل طولانی مدت رفتار شود اما جریمه و تنبیه اثری موقتی دارند. بنابراین نمره منفی لزوماً جلوی حدس زدن دانشجویان را نمی‌گیرد (چاندراتیلکه و همکاران ۲۰۱۱).

توصیه نهایی

در عمل به نظر می‌رسد که هر یک از دو روش نمره‌دهی معمولی و نمره منفی، مزایای خود را دارند و قابل استفاده هستند. اما توجه به چند نکته هنگام به کار بستن آنها ضروری است:

1. Pamplett & Farnill
2. Brady
3. Behaviorism

- هنگام استفاده از نمره منفی، یکی از مسائلی که باید مورد توجه قرار گیرد، هدف ارزیابی است. اگر امتحان تکوینی است و قرار است به یادگیری دانشجو کمک کند و به او نشان دهد که در چه قسمت‌هایی ضعف دارد تا عملکردش را بهبود بخشد، نیازی به نمره منفی نیست (برادی ۲۰۰۵). همچنین اگر هدف آزمون، رتبه‌بندی دانشجویان است، لازم نیست نمره منفی استفاده شود چون تغییری در رتبه افراد نسبت به هم نمی‌دهد (گلدیک ۲۰۰۸). اما اگر قرار است در مورد وضعیت دانشجویان تصمیم‌گیری شود و کسانی که عملکرد ضعیفی دارند، از سیستم خارج کنند، به نظر می‌رسد می‌توان از راه‌کارهای کاهش حدس زدن استفاده کرد.
- باید توجه داشت نمره منفی همانطور که احتمال حدس زدن را کم می‌کند، احتمال حدس آگاهانه را نیز کم می‌کند و اجازه نمی‌دهد که دانشجو بر اساس دانش نسبی و ناقص خود استدلال کند و جواب دهد. درست است که نمی‌خواهیم دانشجو حدس بزند اما نمی‌توان از مفید بودن حدس‌های آگاهانه مخصوصاً در حیطه پزشکی چشم‌پوشی کرد. نتیجه استفاده از آزمون‌هایی با نمره منفی این است که امروزه با دانشجویان محتاطی روبرو هستیم که از ترس از دست دادن نمره نمی‌خواهند خطر کنند (اندرسون ۲۰۰۴).
- نمره منفی در امتحانات فارغ‌التحصیلی و گواهینامه می‌تواند مفید باشد. زیرا فرد در شروع به کار، علاوه بر اینکه باید از دانش خود مطمئن باشد، باید محدودیت‌های دانش خود را نیز بشناسد و بتواند تشخیص دهد چه هنگام لازم است به کتب و منابع مراجعه کند. هنگامی که قرار است فرد با این آزمون فارغ‌التحصیل شود و صلاحیت تصمیم‌گیری او در مواجهه با بیماران تایید گردد، باید واجد درجاتی از اطمینان باشد و بر اطلاعات غلط با حدس و گمان تکیه نکند. (موس ۲۰۰۱)
- استفاده از فرمول‌های نمره‌دهی منجر به اصلاح تاثیر حدس زدن یا از بین بردن اثری که حدس روی نمرات گذاشته است، نمی‌شود. بلکه هدف اصلی استفاده از آنها منصرف کردن فراگیران از حدس زدن است (برتون ۲۰۰۲، داونینگ ۲۰۰۳)
- در هر حال به عنوان جمع‌بندی باید گفت آنچه مورد توافق است، این موضوع است که در همه آزمون‌ها فراگیران حتما از قبل باید در مورد اثر حدس زدن در آزمون توجیه شده باشند. همچنین نکته‌ای که بسیار مهم و تاثیرگذار است این است که نحوه نمره‌دهی به هر شکل که باشد، باید قبل از امتحان به صورت شفاف به دانشجویان گفته شود. به عبارت دیگر، چنانچه امتحان نمره منفی داشته باشد یا نداشته باشد، باید به اطلاع دانشجویان برسد و در دستورالعمل آزمون قید شود. به عبارت دیگر، دانشجویان باید از طریق دستورالعمل‌های واضح شفاهی و کتبی توجیه شوند که چه هنگام به سؤال جواب ندهند. آنها باید از نفعی که از حدس زدن تصادفی یا حتی حدس زدن آگاهانه می‌برند، اطلاع داشته باشند (دیاموند و ایوانز ۱۹۷۳، فراری ۱۹۸۰، بتس و همکاران ۲۰۰۵). زیرا عملکرد دانشجویان در آزمون چندگزینه‌ای متاثر از تجربه قبلی آنها و همچنین راهنمایی‌هایی است که قبل از آزمون به آنها می‌شود.

تعداد بهینه گزینه در سؤالات چندگزینه‌ای چند عدد است؟

یکی از سؤالاتی که سال‌ها ذهن طراحان سؤال و دست‌اندرکاران آموزش را به خود مشغول کرده است، این است که یک سؤال چندگزینه‌ای چند گزینه باید داشته باشد؟ در حال حاضر آنچه برای سؤالات چندگزینه‌ای در حوزه علوم پزشکی معمول است سه تا پنج گزینه است و به نظر می‌رسد که این باور عمومی وجود دارد که هر چه تعداد گزینه بیشتر باشد، بهتر است اما بررسی نتایج مطالعات لزوماً در این راستا نیست. علی‌رغم اینکه مطالعات زیادی، به صورت نظری و محاسبات ریاضی یا به صورت تجربی، در مورد این قضیه انجام شده است، این مسأله کماکان جواب قطعی ندارد و به عنوان یک چالش در حیطه سؤالات چندگزینه‌ای باقی مانده است.

پژوهشگران برای اینکه بررسی کنند تعداد گزینه‌های متفاوت چه تفاوتی با یکدیگر دارند، از روش‌های مختلفی استفاده کرده‌اند. یکی از آنها، محاسبه درصد جذب هر گزینه است. چنانچه در یک سؤال، تعداد کمی از دانشجویان (به صورت

معمول، کمتر از ۵ درصد)، یک گزینه را انتخاب کنند، به آن گزینه غیرعملکردی[1] می‌گویند. به این معنا که چنین گزینه‌ای به اندازه کافی جذاب نبوده و نتوانسته است عملکرد خوبی از خود نشان دهد و دانشجویان را از جواب درست منحرف کند. گاهی اصرار بر طراحی گزینه‌های زیاد منجر به ایجاد گزینه‌های غیرعملکردی می‌شود. در واقع، هنگامی که اساتید گزینه‌های مناسبی پیدا نمی‌کنند، به مواردی روی می‌آورند که گزینه‌های انحرافی ضعیفی هستند و دانشجو به سرعت نادرستی آنها را تشخیص می‌دهد. همچنین گاهی، چون گزینه خوبی به نظر طراح سؤال نمی‌رسد، از «همه موارد» یا «هیچ یک از موارد» استفاده می‌کند که در دستورالعمل‌ها به کرات توصیه شده که مورد استفاده قرار نگیرند. در تحلیل سؤالات آزمون توصیه می‌شود که گزینه‌های غیرعملکردی، حذف یا با موارد بهتر جایگزین شوند. مطالعات زیر یکی از پژوهش‌هایی است که برای یافتن میزان شیوع گزینه‌های غیرعملکردی انجام شد.

هالادینا و داونینگ ۱۹۸۹

مطالعه‌ای روی ۴ آزمون استانداردشده (مجموعا شامل ۴۷۷ گزینه) انجام شد. محققان به این نتیجه رسیدند که ۳۸ درصد گزینه‌ها کمتر از ۵ درصد دانشجویان را به خود جذب کرده بودند و غیرعملکردی بودند. دو سوم سؤالات مورد بررسی تنها یک یا دو گزینه عملکردی داشتند و درصد سؤالاتی که هر سه گزینه انحرافی آنها عملکردی بود، بسیار کم بود (از ۱ درصد تا ۸ درصد). نویسندگان ذکر کردند که تعداد گزینه‌های غیرعملکردی از حد انتظار بیشتر است

تارنت و همکاران ۲۰۰۹

در دانشگاهی در هنگ کنگ تمام آزمون های چندگزینه‌ای به عمل آمده در رشته پرستاری بین سال های ۲۰۰۱ تا ۲۰۰۵ جمع‌آوری شد (۱۲۱ آزمون). از بین آنها آزمون‌های دیسیپلینی با حداقل تعداد ۵۰ سؤال و پایایی بالای ۷۰ انتخاب شدند. نهایتاً ۷ آزمون شامل ۴۱۵ سؤال چهارگزینه‌ای که مجموعا ۶۵۰۲ گزینه داشتند، مورد بررسی قرار گرفتند. تمام این امتحانات بر اساس بلوپرینت و اهداف دوره طراحی شده بودند و قبل از اجرا توسط پانلی از استادان مرور شده بودند. در تحلیل آزمون، فراوانی گزینه‌های غیرعملکردی، و ضریب تمیز گزینه‌ها محاسبه شد. بر اساس نتایج به دست آمده، ۵۴۱ گزینه (۳۵/۱ درصد)، غیرعملکردی بودند. ۴۷۲ گزینه (۳۰/۶ درصد)، ضریب تمایز مثبت داشتند. ۱۰/۲ درصد گزینه‌ها آنقدر غیرجذاب بودند که توسط هیچ دانشجویی انتخاب نشدند. تنها در ۱۳/۸ درصد موارد، هر سه گزینه انحرافی به صورت قابل قبول عمل کرده بودند. به صورت متوسط هر سؤال بین ۱/۳۵ تا ۱/۷۴ گزینه (یعنی کمتر از دو گزینه) عملکردی داشت. به صورت کلی، سؤالاتی که گزینه عملکردی بیشتری داشتند، بهتر قادر به تمایز بین دانشجویان قوی و ضعیف بودند.

این سؤال مطرح است که اگر در واقعیت گزینه‌هایی که طراحی می‌شوند، قادر نیستند به خوبی وظیفه خود را انجام دهند، چرا باید اصرار داشته باشیم که سؤال حتماً چهار یا پنج گزینه داشته باشد. طراحی گزینه‌های انحرافی، یکی از دشوارترین مراحل طراحی سؤال چندگزینه‌ای است. گزینه‌های انحرافی در عین حال که درست نیستند، باید بتوانند دانشجویان با اطلاعات ناکافی را به خود جذب کنند یعنی اشتباه بودن آنها نباید خیلی واضح باشد. طراحی گزینه‌هایی با این خصوصیات که همچنین عاری از خطا باشند و بتوانند به خوبی دانشجوی قوی و ضعیف را از یکدیگر متمایز کنند، کار دشواری است و نیاز به تمرین و صرف وقت دارد. هر چه تعداد گزینه بالاتر باشد، طراحی دشوارتر است و اغلب احتمال خطاهای طراحی سؤال را افزایش می‌دهد. بنابراین مطلوب است که با تعداد گزینه کمتر بتوان به اهداف ارزیابی دست یافت. سؤال این است که آیا بهتر نیست هیأت علمی به جای اینکه وقت خود را برای طراحی گزینه‌های انحرافی با تعداد زیاد، اما با کیفیت نه چندان خوب، صرف کنند، همان وقت را به طراحی گزینه‌های کمتر اما با با کیفیت خوب اختصاص دهند؟ پیش‌فرض این پرسش این است که اگر بتوان تعداد گزینه‌ها را کاهش داد، فشار از روی طراحان سؤال برداشته می‌شود و خطای طراحی سؤال کم می‌شود.

از طرف دیگر، تعداد گزینه‌ها، یکی از مواردی است که بر وقت آزمون اثر می‌گذارد زیرا خواندن تک‌تک آنها طول می‌کشد. اگر بتوان تعداد گزینه‌ها را کم کرد، دانشجویان در زمان کمتری به سؤالات جواب می‌دهند. به این ترتیب در همان بازه زمانی می‌توان تعداد سؤالات را افزایش داد و از این طریق، روایی و پایایی آزمون را بهبود بخشید.

1. Non-functioning

با توجه به موارد فوق، یکی از سؤالات اساسی این است که تعداد بهینه گزینه‌ها برای آزمون تستی، چند گزینه است. به عبارت دیگر، حداقل چه تعداد گزینه لازم داریم تا در عین حال که کیفیت آزمون حفظ می‌شود، نکات مورد بحث نیز لحاظ شوند. برای بررسی تعداد گزینه‌های مناسب آزمون چندگزینه‌ای، چند مطالعه مروری انجام شده است که آخرین آنها درسال ۲۰۰۸ به چاپ رسیده است:

رودریگوئز ۲۰۰۵

نویسندگان در قالب یک متاآنالیز، پژوهش‌های موجود در خصوص تعداد مناسب گزینه‌های سؤال را از سال ۱۹۲۰ تا سال ۲۰۰۰ بررسی کردند و تأثیر تعداد گزینه‌های انحرافی را بر ضریب دشواری، ضریب تمایز، پایایی و روایی آزمون سنجیدند.
این پژوهش با بررسی ۲۷ مطالعه به این نتیجه رسید که کاهش تعداد گزینه‌ها موجب افزایش ضریب دشواری شده است. تغییرات ایجادشده در صورت کاهش تعداد گزینه‌ها از ۴ به ۳، کم و از ۳ به ۲، زیادتر بود. همچنین، کاهش تعداد گزینه‌ها، باعث کاهش ضریب تمایز و کاهش پایایی شد مگر در حالت کاهش گزینه‌ها از چهار به سه، که آنها را به مقدار جزیی زیاد می‌کند. کاهش تعداد گزینه‌ها تغییری در روایی ایجاد نکرد، هرچند که نویسندگان اظهار کردند مطالعات برای اظهار نظر در زمینه روایی کافی نبود.

ویاس و سوپی ۲۰۰۸

هدف نویسندگان یافتن مقالات منتشر شده در مورد تعداد گزینه‌های مناسب در آزمون‌های چندگزینه‌ای در چارچوب مطالعه مروری نظام‌مند بود. نویسندگان یک جستجوی سیستماتیک در سه پایگاه ERIC ،Ovid و Pubmed انجام دادند. ۲۳ مقاله به دست آمد که از بین آنها برخی جنبه نظری و تحلیلی داشتند و برخی مطالعه تجربی بودند. پژوهشگران، تاثیر استفاده از سؤالات سه گزینه‌ای را از جنبه‌های مختلف شامل پایایی، روایی، ضریب تمایز، ضریب دشواری، کارایی آزمون و حدس زدن بررسی کردند.
این مطالعه استفاده از سؤالات سه‌گزینه‌ای را برای استفاده در امتحانات پزشکی توصیه کرده است زیرا به این نتیجه رسیده است که آزمون سه‌گزینه‌ای نسبت به چهار و پنج گزینه‌ای کارایی بیشتری دارد و راحت‌تر اجرا می‌شود. از لحاظ زمانی، وقت کمتری برای نوشتن گزینه‌ها، خواندن گزینه‌ها و اجرای آزمون صرف می‌شود.

پس از این مطالعات مروری، مقالات دیگری نیز در این حوزه به چاپ رسیدند:

تارنت و ویر ۲۰۱۰

آزمونی شامل ۵۰ سؤال چهارگزینه‌ای از ۳۶ دانشجوی مقطع کارشناسی پرستاری گرفته شد. پس از تحلیل آزمون، گزینه غیرعملکردی هر سؤال حذف شد. به این ترتیب آزمونی سه گزینه‌ای ایجاد شد که ۴۱ سؤال از آن انتخاب شد و برای گروه دیگری از دانشجویان در سال بعد مورد استفاده قرار گرفت.
با مقایسه دو آزمون پژوهشگران به این نتیجه رسیدند که میانگین و دامنه نمرات در دو سال، تقریبا مشابه بود اما درصد قبولی در سال دوم کمی کمتر بود. آزمون اول ، ضریب دشواری بیشتری داشت و آسان‌تر بود (۷۳ درصد در برابر ۷۰ درصد) اما این افزایش معنادار نبود. از نظر تمایز، سؤالات سال دوم ضریب تمیز بزرگتری داشتند اما معنادار نبود (۲۶ در برابر ۲۵). همچنین، آزمون سه گزینه‌ای نسبت به آزمون چهارگزینه‌ای از نظر پایایی وضع بهتری داشت (۰/۷۱ در برابر ۰/۶۵). در آزمون سال دوم ۷۴ درصد گزینه ها عملکردی بودند. این میزان در آزمون سال اول تنها ۲۱ درصد بود.

بررسی شواهد مرتبط با تعداد گزینه‌ها نشان می‌دهد که پژوهشگران در پی پاسخ این سؤال بوده‌اند که با کاهش تعداد گزینه‌ها، چه تفاوتی در میزان روایی، پایایی، ضریب دشواری و ضریب تمیز اتفاق خواهد افتاد. در زیر اشاره کوتاهی به هر یک از آنها می‌شود و جمع‌بندی نتایج در جدول شماره ۵-۵ نشان داده شده است:

۱) پایایی

- گریر[1] (۱۹۷۵) با استفاده از فرمول‌های ریاضی ثابت کرد که پایایی آزمون سه‌گزینه‌ای بیشتر است به این شرط که تعداد کل گزینه‌ها در دو آزمون یکسان باشد. در واقع، چون در این حالت تعداد سؤالات آزمون سه‌گزینه‌ای بیشتر است، پایایی جبران می‌شود (ویاس و سوپی ۲۰۰۸).
- یک مقاله مروری با استفاده از نتایج ده مطالعه تجربی نشان داد پایایی با افزایش گزینه‌ها افزایش می‌یابد؛ البته

1. Grier

میزان افزایش پایایی هنگامی که از بیش از سه گزینه استفاده شده بود، بسیار اندک بود (هالادینا و داونینگ ۲۰۰۲).

- متاآنالیز دیگری نشان داد کاهش تعداد گزینه‌ها به هر صورتی (از پنج به چهار، از پنج به دو و از چهار به دو) باعث کاهش پایایی آزمون می‌شود اما کاهش گزینه‌ها از چهار به سه موجب افزایش این ضریب به مقدار جزیی می‌شود (رودریگوئز ۲۰۰۵).
- مطالعات دیگر که به بررسی مقدار پایایی در صورت استفاده از تعداد گزینه‌های مختلف پرداخته بودند یا میزان پایایی را بعد از حذف گزینه غیرعملکردی سنجیده بودند، تغییر اندکی در مقدار پایایی گزارش کردند. در یک مقاله، پایایی آزمون بعد از حذف گزینه غیرعملکردی، مختصراً و به صورت غیرمعنادار افزایش یافت (تارنت و ویر ۲۰۱۰).

۲) روایی

- یک مطالعه در سال ۱۹۸۷ به این نتیجه رسید که سؤال سه‌گزینه‌ای روایی را بهبود می‌بخشد. بر اساس نتایج این مطالعه، نشان داده شد که در یک آزمون ۱۰۰ سؤالی کاهش تعداد گزینه‌ها از چهار به سه، زمان آزمون را به میزان ۱۷ درصد کاهش می‌دهد. نویسندگان پیشنهاد کرده‌اند که با توجه به اینکه در این زمان می‌توان تعداد سؤالات آزمون را به میزان هشت تا نه سؤال افزایش داد، احتمالاً با افزایش نمونه‌گیری، روایی محتوایی افزایش می‌یابد (اوئن و فورمان[1] ۱۹۸۷).
- مطالعه دیگر یک در سال ۱۹۹۴ به این نتیجه رسید که سؤال سه‌گزینه‌ای روایی را بهبود می‌بخشد (ترویزان و همکاران[2]).
- در متاآنالیز رودریگوئز کاهش تعداد گزینه‌ها (از پنج به سه و از پنج به چهار به سه) تغییری در روایی ایجاد نکرد (رودریگوئز ۲۰۰۵).

۳) ضریب تمیز

- تورسکی[3] به صورت ریاضی ثابت کرد که اگر زمان آزمون نسبتی از تعداد کل گزینه‌های به کاررفته در آزمون باشد که معمولاً همین‌طور است، آزمون سه‌گزینه‌ای ضریب تمیز بیشتری در واحد زمان دارد است (تورسکی ۱۹۶۴).
- کوستین[4] با بررسی تجربی سؤالات به این نتیجه رسید که ضریب تمیز در آزمون سه‌گزینه‌ای بیشتر است (کوستین ۱۹۷۰).
- متاآنالیز رودریگوئز (۲۰۰۵) نشان داد کاهش تعداد گزینه‌ها با هر ترکیبی باعث کاهش ضریب تمیز می‌شود به جز حالت کاهش گزینه‌ها از چهار به سه که موجب افزایش این ضریب به مقدار کم می‌شود (رودریگوئز ۲۰۰۵).
- بر اساس مطالعه تارنت و ویر، با حذف گزینه‌های غیرعملکردی، ضریب تمیز به مقدار کم و غیرمعنادار افزایش پیدا کرد (تارنت و ویر ۲۰۱۰).

۴) ضریب دشواری

- طبق نوشته مطالعه مروری ویاس و سوپی، برخی از مقالات به این نتیجه رسیده بودند که بین سؤال سه و پنج‌گزینه‌ای یا بین سؤال سه و چهارگزینه‌ای از نظر ضریب دشواری تفاوتی نبوده است (اوئن و فورمان ۱۹۸۷، کرهان و همکاران[5] ۱۹۹۳).
- بر اساس مقاله مروری هالادینا و داونینگ، تعداد بهینه گزینه از نظر ضریب دشواری، سه یا چهار عدد بود. در واقع، مطالعات تجربی تفاوت بارزی در ضریب دشواری آزمون سه یا چهارگزینه‌ای نیافتند (هالادینا و داونینگ ۲۰۰۲).

---

1. Owen and Froman
2. Trevisan et al.
3. Tversky
4. Costin
5. Crehan et al.

- متاآنالیز رودریگوئز نشان داد که وقتی تعداد گزینه‌ها از چهار به سه کاهش یافت، ضریب دشواری به مقدار کمی افزایش یافت (آزمون آسان‌تر شد) و وقتی که تعداد گزینه‌ها به دو رسید، میزان افزایش ضریب دشواری زیاد بود (آزمون خیلی آسان شد) (رودریگوئز ۲۰۰۵).

جدول۵-۵: نتایج مطالعات در مورد تاثیر تعداد گزینه روی روایی، پایایی و ضرایب تمیز و دشواری آزمون

نام نویسنده مطالعه	پایایی	روایی	ضریب تمیز	ضریب دشواری
هالادینا و داونینگ ۲۰۰۲ (مرور نظام‌مند)	با افزایش گزینه‌ها پایایی به میزان اندک افزایش می‌یابد.			تفاوت بارزی در ضریب دشواری آزمون ۳ یا ۴ گزینه‌ای یافت نشد.
رودریگوئز ۲۰۰۵ (متاآنالیز)	کاهش تعداد گزینه‌ها (از ۵ به ۴، از ۵ به ۲ و از ۴ به ۲) باعث کاهش پایایی آزمون می‌شود ولی کاهش گزینه‌ها از ۴ به ۳ موجب افزایش آن به مقدار کم می‌شود.	کاهش تعداد گزینه‌ها تغییری در روایی ایجاد نمی‌کند.	کاهش تعداد گزینه‌ها باعث کاهش ضریب تمیز می‌شود ولی کاهش گزینه‌ها از ۴ به ۳ موجب افزایش آن به مقدار کم می‌شود	با کاهش گزینه‌ها (از ۴ به ۳) آزمون کمی آسان‌تر شد. با کاهش گزینه‌ها به ۲، آزمون خیلی آسان شد.
مقالات مورد استناد در مطالعه ویاس و سوپی ۲۰۰۸ (مرور).	پایایی آزمون سه گزینه‌ای بیشتر است به شرط تساوی تعداد کل گزینه‌ها.	آزمون سه گزینه‌ای روایی را بهبود می‌بخشد.	آزمون سه گزینه‌ای ضریب تمیز بیشتری دارد.	بین سه و ۵ گزینه یا بین ۳ و ۴ گزینه از نظر ضریب دشواری آزمون تفاوتی نبوده است
مطالعه تارنت و ویر (۲۰۱۰)	با حذف گزینه غیرعملکردی، پایایی مختصراً و به صورت غیرمعنادار افزایش می‌یابد.		با حذف گزینه‌های غیر عملکر د ی ، ضر یب تمیز به مقدار کم و غیرمعنادار افزایش می‌یابد.	

به صورت خلاصه و در کمال تعجب، مطالعات نظری و تجربی متعدد، نتایجی به نفع استفاده از سؤال سه‌گزینه‌ای ارائه داده‌اند. بر اساس یافته‌های این مطالعات، ویژگی‌های روان‌سنجی سؤالات سه‌گزینه‌ای با سؤالات چهار یا پنج‌گزینه‌ای مشابه بوده است و با کاهش تعداد گزینه‌ها، روایی و پایایی آزمون یا ضریب‌های دشواری و تمایز تغییر چندانی نکرده است. گذشته از موضوع ثابت ماندن خصوصیات آزمون، به نظر می‌رسد کاهش تعداد سؤالات، زمان طراحی سؤال و زمان اجرای امتحان را کاهش می‌دهد و از این طریق، وقت و انرژی هیأت علمی و دانشجو را نیز حفظ می‌کند. نتیجه‌گیری اکثر مطالعات این بوده است که اگر واقعاً با کاهش تعداد گزینه‌ها، تغییری در ویژگی‌های روان‌سنجی آزمون رخ نمی‌دهد، استفاده از سؤال سه‌گزینه‌ای به دلیل صرف انرژی و وقت کمتر، مقرون به صرفه است.

با وجود این، آنچه کماکان در آزمون‌های مهم حوزه علوم پزشکی در سراسر دنیا رایج است، استفاده از سؤالات چهارگزینه‌ای و پنج‌گزینه‌ای است. همچنین اکثر راهنماها و دستورالعمل‌های موجود طراحی سؤال، بر پایه سؤالات چهار و پنج‌گزینه‌ای تدوین شده‌اند. علت این امر نامشخص است. یکی از دلایل ذکر شده، ناآگاهی دست اندرکاران امر آموزش و امتحان از نتایج مطالعات مرتبط است. همچنین، ذکر شده است که شاید یکی از دلایل عدم استفاده از سؤالات سه‌گزینه‌ای، نگرانی از افزایش احتمال حدسی پاسخ دادن دانشجویان است زیرا به صورت معمول، مدرسان فکر می‌کنند در سؤال سه‌گزینه‌ای احتمال اینکه دانشجو به صورت حدسی جواب درست را انتخاب کند، به شدت افزایش پیدا می‌کند. البته مطالعات

تجربی لزوماً به نفع این فرضیه نیستند کمااینکه مقایسه آزمون سه و چهارگزینه‌ای در مطالعه راجرز[1] (۱۹۹۹) نشان داد که میزان حدس زدن و استفاده از تکنیک‌های تست‌زنی در سؤال سه‌گزینه‌ای کمتر بوده است (تارنت و ویر ۲۰۱۰).

برخی از پژوهشگران با این استدلال که دلیلی ندارد در یک آزمون، تعداد گزینه در همه سؤالات یکسان باشد، پیشنهاد کرده‌اند که می‌توان از طراحان خواست سؤال سه‌گزینه‌ای طرح کنند و فشاری برای طراحی سؤال چهارگزینه‌ای به آنها وارد نکرد اما در عین حال، این اختیار را برای آنها در نظر گرفت که در صورت تمایل و هر جا که قابلیت اجرا وجود داشت، گزینه اضافه طراحی کنند.

در نهایت، به نظر می‌رسد که مطالعات جدیدتر تلاش دارند تا از طریق نظریه کلاسیک و مدل‌های مختلف نظریه سؤال–پاسخ به بررسی این موضع بپردازند. باید منتظر ماند و دید که آیا نهایتاً در مورد تعداد گزینه‌ای مناسب سؤال، یک پاسخ مشخص و قطعی که مورد توافق پژوهشگران باشد و در عمل به کار برده شود، به دست خواهد آمد یا خیر.

آیا دانشجو از تغییر جواب خود متضرر می‌شود و آیا می‌توان تغییر جواب در جلسه آزمون را به دانشجویان توصیه کرد؟

گاهی پیش می‌آید که دانشجو در جلسه آزمون بعد از اینکه جواب سؤالات را انتخاب می‌کند، در مرور دوباره آنها، پاسخ برخی از سؤالات را عوض می‌کند. باور عمومی بر این است که باید با این وسوسه مقابله کرد. به کرات شنیده می‌شود که مراقبان به دانشجویان تذکر می‌دهند که اگر آزمون را تمام کرده‌اند، پاسخ‌های خود را پاک نکنند و سریع‌تر برگه‌ها را تحویل دهند. این تصور به صورت گسترده وجود دارد که تغییر پاسخ سؤالات موجب کاهش نمره می‌شود. دانشجویان این ذهنیت را دارند که جواب اولیه‌ای که به ذهنشان خطور کرده است، درست است و چنانچه احتمال دهند گزینه دیگری جواب صحیح است، همچنان گزینه اول را حفظ می‌کنند و پاسخ خود را تغییر نمی‌دهند. این باور از طرف استادان و مسؤولان آموزش نیز تقویت می شود.

در واقع، هنگامی که دانشجو جواب سؤال را قبلاً انتخاب کرده است و پس از مرور مجدد، آن را عوض می‌کند، سه حالت امکان دارد رخ دهد: پاسخ قبلاً اشتباه بوده و به درست تبدیل می‌شود (نادرست–به–درست)، پاسخ قبلاً اشتباه بوده و مجددا گزینه اشتباه انتخاب می‌شود (نادرست–به– نادرست) و در نهایت، مواردی که در ابتدا درست انتخاب شده بوده است و سپس نادرست می‌شود (درست–به– نادرست).

مقاله مروری بنجامین و همکاران[2] (۱۹۸۴) بر روی ۳۳ مطالعه که بین سال‌های ۱۹۲۸ تا ۱۹۸۳ انجام شده بودند، صورت گرفت و به این نتیجه رسید که دانشجویان جواب‌های خود را در دو تا نه درصد موارد تغییر می‌دهند. که از این بین ۵۷/۸ درصد نادرست–به–درست و ۲۰/۲ درصد درست–به–نادرست است.

در مطالعه فریدمن–اریکسون[3] (۱۹۹۴) گزارش شد که ۵۶ درصد تغییرات از جواب نادرست به درست و ۲۴ درصد تغییرات از درست به نادرست بوده است. دو مطالعه در نتایج خود به رابطه بین فراوانی تغییر جواب با افزایش نمره دانشجویان اشاره داشتند (فیشر و همکاران۲۰۰۵ [4]).

1. Rogers
2. Benjamin et al
3. Friedman-Erickson
4. Fischer et al.

در اینجا به یکی از مطالعات جدیدتر اشاره می‌شود:

باستروک و یونیورسیتزی ۲۰۱۱

این مطالعه بر روی اطلاعات حاصل از یک آزمون در رشته آموزش انجام شد. ۱۷۱ مورد تغییر جواب (۴/۱ درصد) ثبت شد که ۵۹ درصد نادرست-به-درست، ۲۵ درصد درست-به- نادرست و ۱۶ درصد نادرست -به- نادرست بود. ۴۹ درصد آقایان و ۵۳ درصد خانم‌ها جواب خود را تغییر داده بودند. در این مطالعه اختلاف معناداری در رابطه با تغییر جواب بین دو جنس دیده نشد. به طور متوسط دانشجویان از تغییر جواب سود کردند: نمره کل آقایان از ۱۴/۶۶ به ۱۵/۱۳ رسید که اختلاف معنادار بود. نمرات خانم‌ها از ۱۵/۱۷ به ۱۵/۹۴ رسید که این تفاوت نیز معنادار بود. با محاسبه ضریب شواری سؤالات و رابطه آن با میزان تغییر دادن در جواب‌ها به این نتیجه رسیدند که بین این دو موضوع رابطه وجود دارد. به عبارت دیگر، هرچه سؤال سخت‌تر باشد، میزان تغییر جواب در آن بالاتر است. در این مطالعه رابطه‌ای بین میزان تغییر در جواب با ضریب تمایز دیده نشد.

به صورت کلی، مطالعات تجربی نشان می‌دهند اولاً پدیده تغییر جواب در بین دانشجویان شایع نیست و اکثراً دوست دارند بر سر همان پاسخ اولیه بمانند. ثانیاً در همان موارد محدود تغییر جواب، دانشجویان از این موضوع متضرر نمی‌شوند. بنابراین به نظر می‌رسد این باور که دانشجو باید به جواب اولیه خود اعتماد کند و نباید آن را تغییر دهد، چندان درست نیست. شواهد نشان می‌دهند اگرچه اکثر فراگیران تصور می‌کنند تغییر پاسخ به ضررشان است و آن را کار نادرستی می‌دانند، در واقعیت از آن سود می‌برند. شاید مشکل اینجاست که تغییر موارد درست به پاسخ‌های نادرست دردناک‌تر است و بر اساس آنچه به اثر فون رسترف[1] معروف است، بیشتر در حافظه افراد ماندگار می‌شود.

پس مشخص شد که شواهد نشان می‌دهند تغییر جواب، افزایش نمره به دنبال دارد. حال سؤال این است که آیا آگاه کردن دانشجو از نتیجه این شواهد، به نفع اوست؟

هرچند در سال ۱۹۸۲ ساتن[2] گزارش داد که ارائه آموزش سیستماتیک به دانشجویان در این خصوص باعث افزایش معنادار تعداد موارد تغییریافته و همچنین افزایش نمره دانشجویان شد، مطالعات فوت و همکاران[3] (۱۹۷۲) و پرینسل و همکاران[4] (۱۹۹۴) این یافته را تایید نکردند و نشان دادند که قبل و بعد از چنین آموزشی، تفاوتی در فراوانی جواب‌های تغییریافته و یا نمرات دانشجویان مشاهده نشد (بائر و همکاران[5] ۲۰۰۷). اما بائر و همکاران در مقاله خود به این نتیجه رسیدند که عملکرد دانشجویان با دریافت توصیه‌ای مبنی بر تغییر دادن جواب سؤالاتی که از آنها مطمئن نیستند، بهبود می‌یابد و به نظر می‌رسد که این بهبود بر اساس بازاندیشی و تامل دقیق دانشجویان است.

بائر و همکاران ۲۰۰۷

۷۹ دانشجوی پزشکی سال سوم از دانشگاه مونیخ در این مطالعه شرکت کردند. دانشجویان به صورت تصادفی به دو گروه تقسیم شدند. به یکی از گروه‌ها (۴۱ نفر) توصیه شد که در مواردی که به جواب خود شک دارند، (تا یک بار) جواب خود را تغییر دهند. ولی گروه کنترل (۳۸ نفر) اطلاعی از این موضوع نداشت. از هر دو گروه خواسته شد که موارد تغییر جواب را در دفترچه خود مشخص کنند. آزمون شامل ۷۸ سؤال یک نمره‌ای بود که از ۴ تا ۱۱ گزینه داشت. در کل ۶۱۶۲ سؤال از مجموع دانشجویان بررسی شد.

نتایج نشان داد که ۷۲ دانشجو در کل حداقل یک بار جواب خود را تغییر داده بودند. در مجموع، جواب ۳۲۳ سؤال یک بار تغییر کرده بود (۵/۲ درصد). ۴۸/۲ درصد موارد غلط-به-درست، ۲۱/۶ درصد موارد درست-به-غلط و ۳۰/۲ درصد غلط-به-غلط تغییر کرده بود. این ۷۲ دانشجو از تغییر جواب به طور میانگین ۱/۴ نمره سود برده بودند. تغییر جواب بیش از یک بار، توسط ۲۰ دانشجو و در ۲۷ سؤال دیده شد که به طور میانگین موجب افزایش ۰/۱۱ نمره شد. از این موارد، ۱۸/۴ درصد تبدیل غلط-به-درست بود و ۳/۷ درصد تبدیل درست-به-غلط.

از نظر مقایسه بین دو گروه، یافته ها حاکی از این بود که گروه مداخله به طور معنادار بیشتر جواب خود را تغییر داده بود. علی رغم اینکه دانشجویان گروه مداخله، به خاطر تغییر جواب، از نظر نمره سودی بیشتر از گروه کنترل دریافت کردند (۱/۸ نمره در برابر ۰/۹۱ نمره)، این افزایش از نظر آماری معنادار نبود. پژوهشگران معتقد بودند این مقدار افزایش نمره هر چند از لحاظ آماری معنادار نشده است اما با توجه به نمره حد نصاب قبولی (۶۰ از ۱۰۰)، می‌توانسته تاثیرگذار باشد و بنابراین از نظر آموزشی مهم و چشمگیر است.

1. Von Restorff effect
2. Sutton
3. Foote et al.
4. Prinsell et al.
5. Bauer et al.

در حوزه پزشکی مطالعه دیگری نیز انجام شده است که از این جهت حائز اهمیت است که توصیه روشن‌تری برای دانشجویان دارد.

فیشر و همکاران ۲۰۰۵

در یک امتحان بورد کشوری در آلمان که آزمونی با ۵۸۰ سؤال پنج گزینه‌ای و بدون نمره منفی بود، ۳۶ دانشجو از چند دانشگاه برای شرکت در مطالعه داوطلب شدند. از آنها خواسته شد تا سؤالاتی که جواب آنها را تغییر می‌دهند، در دفترچه آزمون مشخص کنند.
نتایج نشان داد دفترچه‌های هر یک از این دانشجویان حداقل یک جواب تغییریافته داشت. روی هم رفته از بین ۲۰۸۸۰ سؤال مربوط به ۳۶ دانشجو، جواب ۸۱۹ سؤال تغییر پیدا کرده بود. از آنجا که دانشجویان برخی از جواب ها را بیش از یک بار تغییر داده بودند، کلاً ۸۴۶ (۴/۳ درصد) بار تغییر ثبت شد. ۵۵ درصد غلط-به-درست، ۲۵ درصد درست-به-غلط و ۲۰ درصد غلط-به غلط بود. یافته‌ها نشان داد که تغییر جواب باعث افزایش نمره شد اما در کل، دانشجویان با نمرات بالا کمتر جواب خود را تغییر داده بودند. در مواردی که شرکت‌کنندگان بیش از یک بار جواب خود را عوض کرده بودند، جواب‌ها بیشتر نادرست شده بود.

بر اساس نتایج مطالعه فیشر و همکاران، هر چند دانشجویان از تغییر جواب سود می‌برند اما باید به خاطر داشت که تغییر جواب بیش از یک بار نشان دهنده این است که سؤال واقعا برای دانشجو سخت بوده است. بنابراین، حدسی که از ابتدا زده، چندان خوب نبوده و در نهایت با تغییر جواب احتمالاً نتیجه خوبی نمی‌گیرد. پس باید به دانشجویان یاد داد که در مواردی که از جواب خود مطمئن نیستند و تمایل دارند جواب را عوض کنند، به شک خود اعتماد کنند و مجدداً در مورد سؤال و پاسخ آن فکر کنند. در این صورت احتمالاً از تغییر جواب سود می برند، مگر در مواردی که شکشان زیاد باشد به طوری که مجدداً بخواهند جواب خود را تغییر دهند که در این صورت، احتمال انتخاب گزینه نادرست زیاد می‌شود (فیشر و همکاران ۲۰۰۵).

منابع

1. Albanese MA. The projected impact of the correction for guessing on individual scores. Journal of Educational Measurement 1988;25(2):149-157.
2. Albanese MA. Type K and other complex multiple-choice items: An analysis of research and item properties. Educ Meas: Issues and practice 1993; 12:28–33.
3. Anderson J. Multiple choice questions revisited. Med Teach 2004; 26(2):110-3.
4. Angoff WH, Schrader WB. A Study of Alternative Methods for Equating Rights Scores to Formula Scores. ETS Research Report 1981.
5. Angoff WH, Schrader WB. A study of hypotheses basic to the use of rights and formula scores. Journal of Educational Measurement1)21;1984):1-17.
6. Angoff WH. Does Guessing Really Help? Journal of Educational Measurement336–26:323;1989
7. Andrà C, Magnano G. Multiple-Choice Math Tests: Should We Worry About Guessing? Quaderni Di Ricerca In Didattica 2011;21:235-243
8. Baştürk R, Üniversitesi P. Impact of Answer-Switching Behavior on Multiple-Choice Test Scores in Higher Education. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi 2011; 2(1):114-120.
9. Bauer D, Kopp V, Fischer MR. Answer changing in multiple choice assessment change that answer when in doubt – and spread the word! BMC Medical Education 2007, 7:28
10. Bender DA. MCQ, EMSQ or multiple true/false questions? Bioscience Education e-journal 2003; 2. available at: http://bio.ltsn.ac.uk/journal/vol2/beej-2-L1.htm
11. Ben-Shakhar G, Sinai Y. Gender Differences in Multiple-Choice Tests: The Role Of Differential Guessing Tendencies. Journal of Educational Measurement 1991;28(1):23-35
12. Betts LR, Elder TJ, Hartley J, Trueman M. Does Correction For Guessing Reduce Students' Performance On Multiple Choice Examinations? Yes? No? Sometimes? Assessment & Evaluation In Higher Education 2009;34(1): 1-15
13. Bliss LB. A test of Lord›s assumption regarding examinee guessing behavior on multiple-choice tests using elementary school students. Journal of Educational measurement 1980;17(2):147-153
14. Bond AE, Bodger O, Skibinski DO, et al. Negatively-Marked MCQ Assessments That Reward Partial Knowledge Do Not Introduce Gender Bias Yet Increase Student Performance and Satisfaction and Reduce Anxiety. PloS one 2013;8(2):e55956.
15. Brady AM. Assessment of learning with multiple-choice questions. Nurse Education in Practice 2005;5:238-242
16. Budescu D, Bar-Hillel M. To Guess Or Not To Guess: A Decision-Theoretic View Of Formula Scoring. Journal of Educational Measurement 1993;30(4):277–291
17. Burton SJ, Sudweeks RR, Merrill PF, Wood B. How to Prepare Better Multiple Choice Test Items:

Guidelines for University Faculty, 1991

18. Burton RF. Misinformation, partial knowledge and guessing in true/false tests. Med Educ 2002;36:805–11
19. Burton R. Quantifying The Effects Of Chance In Multiple Choice And True/False Tests: Question Selection And Guessing Of Answers. Assessment & Evaluation in Higher Education 2001;26(1):41-50
20. Burton RF. Guessing in selected response tests. Med Educ38;2003 : 112
21. Campbell DE. How to write good multiple-choice questions. J Paediatr Child Health 2011; 47(6):322-5.
22. Case SM, Swanson DB. Constructing Written Test Questions for the Basic and Clinical Sciences, Philadelphia, National Board of Medical Examiners; 2002
23. Case SM, Swanson DB, Becker DB. Verbosity, window dressing, and red herrings: do they make a better test item? Academic Medicine 1996;71(10): S28–S30.
24. Chambers DW. Correcting for guessing on multiple-choice exams. J Dent Educ 2007; 71(2): 193-4.
25. Chandratilake M, Davis M, Ponnamperuma G. Assessment of medical knowledge: the pros and cons of using true/false multiple choice questions. Natl Med J India 2011; 24(4):225-8.
26. Collins J. Writing Multiple-Choice Questions for Continuing Medical Education Activities and Self-Assessment Modules. RadioGraphics 2006; 26: 543-551.
27. Costin F. The optimal number of alternatives in multiple-choice achievement tests: Some empirical evidence for a mathematical proof. Educ Psychol Meas 1970;30:353–8.
28. Cronbach LJ. Essentials of psychological testing (4th edn), New York: Harper Row1984 .
29. Cross LH, Frary RB. An empirical test of Lord›s theoretical results regarding formula-scoring of multiple-choice tests. Journal of Educational Measurement 1977,14(4)313-321:
30. Diamond J, Evans W. The correction for guessing. Review of Educational Research 1973;43:181-191
31. Downing SM. Guessing on selected- response examinations Med Educ 2003;37: 670–1.
32. Downing SM. On guessing corrections. Med Educ 2004;38: 113
33. Downing SM. True-false, alternate-choice, and multiple-choice items. Educ Meas: Issues and practice 1992;11:27–30
34. Espinosa MP, Gardeazabal J. Optimal Correction for Guessing in Multiple-Choice Tests. Technical report, Department of Foundations of Economic Analysis II, University of the Basque Country 2007
35. Fischer MR, Herrmann S, Kopp V. Answering multiple-choice questions in high-stakes medical examinations. Med Educ 2005; 39(9): 890-4.
36. Frary R. The Effect Of Misinformation, Partial Information, And Guessing On Expected Multiple-Choice Test Item Scores. Applied Psychological Measurement January 1980;4(1):79-90
37. Frary RB. Partial-credit scoring methods for multiple-choice tests. Applied Measurement in Educa-

tion 1989;2(1):79-96

38. Gardner-Medwin AR, Curton NA. Confidence Assessment in the Teaching of Physiology. Journal of Physiology 1996;494:74P.
39. Jennings S, Bush M. A comparison of conventional and liberal (free-choice) multiple-choice tests. Practical Assessment, Research & Evaluation 2006;11(8):1-5.
40. Godlik Z. Abandoning negative marking. European Journal of Anaesthesiology 2008;25: 349-351
41. Haladyna TM, Downing SM. Validity of a taxonomy of multiple-choice item-writing rules. Applied Measurement in Education 1989; 2(1): 51-78.
42. Hammond EJ, McIndoe AK, Sansome AJ, Spargo PM. Multiple-choice examinations: adopting an evidence-based approach to exam technique. Anaesthesia 1998; 53(11): 1105-8.
43. Harden RM, Brown RA, Biran LA, Dallas Ross WP. Wakeford RE. Multiple choice questions: to guess or not to guess, Medical Education 1976; 10:27.
44. Karandikar RL. Multiple-Choice Tests, Negative Marks and an Alternative. Indian Statistical Institute 2006;86-93
45. Kimiai K. The Effects of Application of Correction-For-Guessing Formula on The Validity of The M-Ctests Of English Grammar Of EFL Students. Frontiers of Language and Teaching 2012;3:228-238
46. Lord FM. Formula scoring and validity. Educational Psychology Measurement 1963;23:663-72
47. Mattson D. The effects of guessing on the standard error of measurement and the reliability of test scores. Educational and Psychological Measurement 1965;25:727-730
48. Lau PNK, Lau SH, Hong KS, Usop H. Guessing, Partial Knowledge, and Misconceptions In Multiple-Choice Tests. Educational Technology & Society 2011;14(4):99-110.
49. McGuire B. Multiple choice examinations. Anaesthesia1999 ; 54(7):720
50. Moss E. Multiple choice questions: their value as an assessment tool. Curr Opin Anaesthesiol 2001;14(6):661-6.
51. Muijtjens AM, Mameren HV, Hoogenboom RJ, Evers JL, van der Vleuten CP. The effect of a ‹don›t know› option on test scores: number-right and formula scoring compared. Med Educ 1999;33(4):267-75.
52. Nnodim JO. Multiple-choice testing in anatomy. Med Educ 1992;.9–26:301
53. Owen SV, Froman RD. What’s wrong with three-option multiple choice items? Educ Psychol Meas 1987; 47: 513–522.
54. Pamplett R, Farnill D. Effect of anxiety on performance in multiple-choice examination. Medical Education 1995;29:298–302
55. Prihoda TJ, Pinckard RN, McMahan CA, Jones AC. Correcting for guessing increases validity in multiple-choice examinations in an oral and maxillofacial pathology course. J Dent Educ 2006; 70(4): 378-86.
56. Rodriguez MC. Three options are optimal for multiple-choice items: A metaanalysis of 80 years of

research. Educ Meas: Issues Pract 2005;24:3–13.

57. Rowley GL, Traub RE. Formula Scoring, Number-Right Scoring, and Test-Taking Strategy. Journal of Educational Measurement 1977;14(1):15-22
58. Ruch GM, Degraff MH. Corrections for Chance and «Guess» Vs. «Do Not Guess» Instructions in Multiple Response Tests. Journal of Educational Psychology 1926;17(6):368-375
59. Schuwirth LWT, van der Vleuten CPM. ABC of learning and teaching in medicine: Written assessment. BMJ 2003; 326(7390):643-645.
60. Schuwirth LW, van der Vleuten CP. Different written assessment methods: what can be said about their strengths and weaknesses? Med Educ 2004; 38(9): 974-9.
61. Tamir P. Multiple Choice Items: How to Gain the Most Out of Them. Biochemical Education 1991; 19(4): 188-192
62. Tarrant M, Ware J, Mohammed AM. An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. BMC Med Educ 2009;9-40.
63. Tarrant M, Ware J. A comparison of the psychometric properties of three- and four-option multiple-choice questions in nursing assessments. Nurse Educ Today 2010; 30(6):539-43.
64. Tractenberg RE, Gushta MM, Mulroney SE, Weissinger P. Multiple choice questions can be designed or revised to challenge learners' critical thinking. Advances in Health Sciences Education 2013; 18(5):945-961
65. Trevisan MS, Sax G, Michael WB. The effects of the number of options per item and student ability on test validity and reliability. Educ Psychol Meas 1991;51:829–37.
66. Tversky A. On the optimal number of alternatives at a choice point. J Math Psychol 1964;1:386–91.
67. Vyas R, Supe A. Multiple choice questions: a literature review on the optimal number of options. Natl Med J India 2008; 21(3):130-3.

۶۸. خطیبی ر، قادرمرزی م، یزدانی ش، زارع‌زاده ی. رابطه جنسیت با نمره دانشجویان در انتخاب گزینه‌های اطمینان در آزمون های مبتنی بر اطمینان. مجله ایرانی آموزش در علوم پزشکی ۱۳۹۰؛ ۱۱ (۷): ۹۳۲–۹۲۶.

۶۹. مرتاض هجری س، خباز مافی نژاد م، جلیلی م. پاسخ حدسی به سؤالات چند گزینه ای: چالش ها و راه کارها. مجله ایرانی آموزش در علوم پزشکی ۱۳۹۰؛ ۱۴ (۷): ۶۰۴–۵۹۴

فصل ۶

سؤال «چندگزینه‌ای با بهترین پاسخ»

ساختار سؤال «چندگزینه‌ای با بهترین پاسخ»

منظور از «سؤال چندگزینه‌ای با بهترین پاسخ» همان سؤال نوع A است که در فصل اول این بخش اجمالاً توضیح داده شد و در این فصل به صورت تفصیلی مورد بحث قرار می‌گیرد. این نوع سؤال که به اختصار «سؤال چندگزینه‌ای» نامیده می‌شود، شامل یک پایه[1]، یک سؤال هدایت‌کننده[2] و تعدادی گزینه[3] است. پایه سؤال در واقع معرف یک وضعیت است و می‌تواند شامل یک مورد بالینی شامل شکایت اصلی به همراه علایم و نشانه‌های مرتبط، اطلاعات آزمایشگاهی و نظیر آن باشد. سؤال هدایت‌کننده، همان سؤالی است که از داوطلب خواسته می‌شود به آن جواب دهد. جزء آخر، گزینه‌های پاسخ است. یکی از گزینه‌ها به عنوان صحیح‌ترین گزینه[4] انتخاب می‌شود و بقیه گزینه‌ها، گزینه‌های انحرافی خوانده می‌شوند. انتظار می‌رود پایه سؤال نسبتاً مفصل و طول گزینه‌های آن کوتاه باشد. نمای ظاهری یک سؤال چندجوابی مطلوب به صورت زیر است:

مثالی از ساختار مطلوب سؤال «چندگزینه‌ای با بهترین پاسخ»

یک مرد ۶۰ ساله با شکایت ضعف پیش‌رونده دست‌ها و پاها مراجعه می‌کند. او می‌گوید که هنگام شانه زدن موی سر و بالا رفتن از پله مشکل دارد. همچنین دچار اشکال در بلع نیز شده اما اختلال بینایی ندارد. در معاینه، شما متوجه ضایعات پوستی ماکولوپاپولر بر روی پلک‌ها، بینی و گونه‌های بیمار می‌شوید. معاینه مفاصل بیمار طبیعی است.
محتمل‌ترین تشخیص کدام است؟

الف) درماتومیوزیت　　ب) میاستنی گراو　　ج) پلی‌میالژی روماتیکا　　د) آرتریت روماتوئید

خطاهای ساختاری سؤال «چندگزینه‌ای با بهترین پاسخ»

دانشجویان در مقاطع مختلف و در هر ترم امتحانات گوناگونی را پشت سر می‌گذرانند که برای بسیاری از آنها استرس و فشار زیادی تحمل می‌کنند. اما متأسفانه کیفیت این امتحانات همیشه تضمین شده نیست. در واقع خطاهای طراحی سؤال مواردی هستند که با اصول طراحی سؤال مطابقت ندارند و از دستورالعمل‌هایی که در این زمینه وجود دارد، تخطی کرده‌اند. این خطاها می‌توانند در نتیجه آزمونی که از دانشجویان گرفته می‌شود، تاثیرگذار باشند زیرا در بسیاری از موارد چیزی را می‌سنجند که مربوط به هدف و محتوای سؤال نیست.

1. Stem
2. Lead-in
3. Alternative
4. Correct answer

توجه به این مسأله ضروری است که علاوه بر خطاهای ساختاری سؤال، مورد دیگری وجود دارد که کیفیت سؤالات چندگزینه‌ای را تحت تاثیر قرار می‌دهد و آن سطح شناختی پایین سؤال است. مخصوصاً در علوم پزشکی انتظار می‌رود فراگیران بتوانند اطلاعات زیاد و پیچیده را تحلیل کنند و برای تصمیم‌گیری به کار ببرند، چنانچه آزمون این توانایی را نسنجد، نمی‌توان مطمئن بود که آیا این فرد قادر است در صورت لزوم از مهارت‌های شناختی سطح بالای خود استفاده کند یا خیر. علاوه بر این، نشان داده شده است که میزان شیوع خطاهای ساختاری در سؤالاتی که سطح شناختی پایین دارند، به مراتب بیشتر است. ضمن در نظر گرفتن اهمیت سطح شناختی سؤال که در فصل قبلی به آن پرداخته شد، در اینجا در خصوص خطاهای ساختاری سؤال صحبت می‌کنیم. خطاهای ساختاری سؤال را می‌توان به دو دسته کلی تقسیم کرد: خطاهایی که منجر به تکنیک‌های تست‌زنی می‌شوند و خطاهایی که موجب دشواری بی‌مورد سؤال می‌گردند.

خطاهای مربوط به تست‌زنی

این خطاها انواعی هستند که منجر به سرنخ دادن به آزمودنی می‌شوند و موجب می‌گردند فراگیر بدون داشتن دانش مربوطه و صرفاً از روی حدس و گمان، گزینه درست را انتخاب کند. در حالت معمول، انتظار می‌رود پاسخ صحیح دانشجو ناشی از تجربه و دانش او در حوزه مورد بررسی باشد نه حاصل میزان تجربه و مهارتی که بر اساس تست زدن زیاد به دست آورده است. این خطاها به دانشجو کمک می‌کنند تا با تکیه بر مهارت‌های تست‌زنی و بدون آنکه از محتوای مورد پرسش چیز زیادی بداند، جواب درست را پیدا کند. در صورت وجود چنین خطایی در سؤال، هرچقدر میزان تجربه و آشنایی دانشجو با سؤالات چندگزینه‌ای بیشتر باشد، امکان استفاده او از تکنیک‌های تست‌زنی بیشتر می‌شود. مواردی از خطا که باعث می‌شوند دانشجو از راه مهارت تست‌زنی به پاسخ صحیح برسد، عبارت هستند از:

- سرنخ‌های گرامری یا منطقی: این خطا معمولاً ناشی از عدم هماهنگی گزینه‌های انحرافی با سؤال هدایت‌کننده یا پایه سؤال ایجاد می‌شود. زیرا تمام حواس طراح سؤال معطوف به گزینه صحیح است که باعث می‌شود گزینه‌های انحرافی از نظر گرامری یا منطقی با آنچه در سؤال هدایت‌کننده یا پایه سؤال نوشته شده است، همخوانی نداشته باشند. در مثال زیر، به علت وجود کلمه تجویز در پایه سؤال دانشجو مطمئن است که گزینه‌های الف و د جواب نیستند:

مثالی از خطای گرامری یا منطقی

آقای ۶۰ ساله که به صورت بیهوش در پیاده رو پیدا شده است، توسط پلیس به اورژانس آورده شده است. پس از اینکه از باز بودن راه هوایی مطمئن شدید، به عنوان اولین اقدام کدام یک از موارد زیر را تجویز می‌کنید؟

الف) آزمایش مایع مغزی-نخاعی
ب) گلوکز و ویتامین B_1
ج) فنی توئین
د) CT scan

- استفاده از کلمات مطلق: دانشجو می‌داند که وقتی عبارتی به صورت قاطع بیان می‌شود (کلماتی مانند همیشه و هرگز)، به احتمال زیاد پاسخ صحیح نیست. چنانچه کلمات مطلق در پایه سؤال باشند، مشکلی ایجاد نمی‌شود اما هنگامی که طراحی سؤال به جای اینکه افعال را در پایه یا سؤال هدایت‌کننده قرار دهد، در گزینه‌ها می‌گنجاند، معمولاً این خطا رخ می‌دهد.

مثالی از خطای کلمه مطلق

آمبولی پولمونر:

الف) همیشه همراه با تب است.
ب) هیچ گاه در غیر سیگاری ها دیده نمی‌شود.
ج) همیشه با پنومونی اشتباه می‌شود.
د) با تزریق هپارین درمان می‌شود.

- گزینه صحیح طولانی: گاهی طراح سؤال به علت توجه زیادی که به گزینه صحیح دارد، آن را به صورت طولانی، با جزئیات بیشتر و اختصاصی‌تر می‌نویسد. این مسأله باعث می‌شود که دانشجو از روی ظاهر گزینه‌ها تشخیص دهد پاسخ صحیح کدام است.

مثالی از خطای گزینه طولانی

مرد ۴۰ ساله‌ای مبتلا به سندرم متابولیک است. موثرترین اقدام جهت کاهش بروز دیابت کدام است؟

الف) تجویز مت فورین
ب) تجویز ویتامین E
ج) تجویز تیازولیدین دیون
د) توصیه به کاهش وزن، محدودیت مصرف چربی، افزایش فعالیت فیزیکی

- تکرار کلمات: گاهی کلمه‌ای که در پایه سؤال آمده است، عیناً یا مشابه یا معادل آن در یکی از گزینه‌ها هم تکرار شده است. در مثال زیر، گزینه ج جواب درست است و در واقع کلمه real معادل واژه «واقعی» در پایه و گزینه تکرار شده است.

مثالی از خطای تکرار کلمات

آقای ۵۸ ساله با سابقه مصرف مواد و بستری در بیمارستان روانی، در حالی که گیج و آژیته است، به بیمارستان منتقل شده است. او می‌گوید که احساس می‌کند دنیا غیرواقعی است. این نشانه چه نام دارد؟

الف) Depersonalization ب) Derailment ج) Derealization د) Anxiety

- استراتژی همگرایی[1]: هنگامی که یک کلمه در گزینه‌ها تکرار می‌شود، دانشجو متوجه می‌شود که احتمالاً قسمتی از پاسخ صحیح را در خود دارد. در مثال زیر، کلمه ششم دو بار (از چهار حالت ممکن) و کلمه چپ سه بار (از چهار حالت ممکن) تکرار شده‌اند. بنابراین دانشجو حدس می‌زند جواب درست قوس ششم چپ باشد.

مثالی از خطای همگرایی

منشا مجرای شریانی چیست؟

الف) قوس ششم آئورتی راست
ب) قوس سوم آئورتی چپ
ج) قوس چهارم آئورتی چپ
د) قوس ششم آئورتی چپ

خطاهای مربوط به دشواری بی‌مورد

این خطاها با ایجاد ابهام و پیچیدگی، سؤال را به صورت بی‌جا و بی‌مورد دشوار می‌کنند و باعث می‌شوند دانشجویی که دانش مرتبط را دارد، به علت اِشکال سؤال منظور را به درستی متوجه نشود و نمره سؤال را کسب نکند. مثال‌هایی از خطاهای منجر به دشواری بی‌مورد عبارت هستند از:

- ابهام و پیچیدگی جملات: هنگامی که جمله‌بندی گزینه‌ها به صورت مبهم و پیچیده و طولانی است، دانشجو مجبور است چندبار آنها را بخواند. در واقع، آنچه در اینجا مورد سنجش قرار می‌گیرد، دانش و اطلاعات فراگیر نیست بلکه ممکن است مثلاً میزان سرعت او در خواندن گزینه‌های طولانی و یا تشخیص پیچیدگی‌های سؤال باشد. البته این خطا ربطی به پایه طولانی ندارد. اینکه آیا باید در پایه سؤال اطلاعات زیاد داده شود تا دانشجو از بین آنها اطلاعات

1. Convergence strategy

مرتبط را شناسایی کند و تصمیم بگیرد، یا اینکه باید پایه متمرکز و هدفمند و بدون اطلاعات غیرضروری طراحی شود، مسأله‌ای است که به هدف آزمون بستگی دارد و لزوماً منجر به ابهام نمی‌شود.

- طرح پایه سؤال به صورت ناقص: اطلاعاتی که برای پاسخ دادن به سؤال مورد نیاز است باید به صورت شفاف در متن سؤال بیاید. طرح پایه سؤال به صورت ناقص باعث می‌شود جواب صحیح دانشجویان کاهش پیدا کند. یک مطالعه میزان این کاهش را ۱۰ تا ۱۵ درصد برآورد کرده است (کولینز ۲۰۰۶).
- استفاده از قیود مبهم: مطالعات نشان داده‌اند که کلمات مبهم مانند «به ندرت، معمولاً، گاهی، اغلب و ...» تعریف مشخص و واحدی ندارند و افراد مختلف، تفاسیر و برداشت‌های متفاوتی در مورد میزان آنها دارند. به‌کاربردن این قیود در سؤال باعث پیچیدگی و ابهام سؤال می‌شود.
- نحوه نگارش داده‌های عددی: هنگامی که گزینه‌ها به صورت عدد ارائه می‌شوند، باید به ترتیب کاهشی یا افزایشی مرتب شوند و همه در یک فرمت باشند مثلاً همه به صورت رقم یا حروف نوشته شوند. اگر گزینه‌ها نامرتب ردیف شوند، باعث سردرگمی دانشجو شده و سختی بی‌مورد ایجاد می‌کنند. در مثال زیر در گزینه ب، درصد با علامت نوشته شده است. در حالی که در بقیه موارد با حروف نوشته شده است. همچنین گزینه ب، به صورت دامنه اعداد بیان شده است.

مثالی از خطا در نگارش اعداد

متعاقب عفونت تناسلی ثانویه، چقدر احتمال دارد که بیمار نابارور شود؟

الف) کمتر از ۰۲ درصد
ب) ۰۲% تا ۰۳%
ج) ۰۹ درصد
د) بیشتر از ۵۷ درصد

- وجود گزینه «هیچ یک از موارد فوق»[1] یا «همه موارد فوق[2]»: هنگامی که طراح سؤال برای تهیه گزینه‌های انحرافی مناسب با مشکل مواجه می‌شود، به ناچار از این گزینه‌ها استفاده می‌کند. این گزینه‌ها مشکلات زیادی ایجاد می‌کنند. اول اینکه معمولاً به دانشجو سرنخ می‌دهند. به عنوان مثال، اگر پاسخ صحیح سؤال، گزینه همه موارد فوق» باشد و دانشجو بتواند صرفاً درستی دو گزینه را تشخیص دهد، به جواب می‌رسد. در واقع اگر «هیچ یک از موارد فوق»، جواب درست باشد، بیشتر از اینکه توانایی دانشجو را در رسیدن به پاسخ صحیح اندازه بگیرد، توانایی او را در شناسایی گزینه اشتباه می‌سنجد. اگر این عبارت، گزینه انحرافی باشد، به نظر برخی از دانشجویان قابل‌قبول نمی‌رسد و مشکل دارد. عبارت «همه موارد فوق» نیز جواب درست باشد، دانشجو صرفاً با تشخیص صحت دو گزینه به جواب می‌رسد و اگر گزینه انحرافی باشد، دانشجو صرفاً با تشخیص نادرستی یک گزینه به جواب می‌رسد. همچنین، با استفاده از این دو گزینه، سؤال از حالت چندگزینه‌ای به سؤال «درست–نادرست» تبدیل می‌شود. پس سایر گزینه‌ها باید کاملاً درست یا کاملاً نادرست نوشته شوند. اما واقعیت این است که چون بیشتر توجه طراح سؤال به گزینه صحیح معطوف می‌شود و در نوشتن سایر گزینه‌ها دقت زیادی نمی‌کند، معمولاً سایر گزینه‌ها کاملاً غلط یا کاملاً درست نیستند. بنابراین، دانشجویان سردرگم می‌شوند و مدام باید بین گزینه‌ها بالا و پایین می‌روند. مثلاً اگر دو گزینه کاملاً نادرست باشد و در مورد گزینه سوم نتوان با اطمینان گفت کاملاً نادرست است و به درجاتی صحیح باشد، دانشجو نمی‌داند باید گزینه کاملاً صحیح را انتخاب کند یا گزینه صحیح‌ترین را. در حالت اول، باید سه گزینه را نادرست در نظر بگیرد و گزینه چهارم یعنی «هیچ یک از موارد فوق» را انتخاب کند. در حالی که در حالت دوم باید گزینه سوم را که نسبتاً صحیح است، انتخاب کند. در خصوص پیامدهای استفاده از این عبارات، هر چند که شواهد در مورد استفاده از

1. None of the above
2. All of the above

«همه موارد فوق» به جمع‌بندی نرسیده‌اند اما در مورد «هیچ یک از موارد فوق» اتفاق نظر دارند که قدرت تمیز و پایایی آزمون را کاهش می‌دهد (برتون و همکاران ۱۹۹۱). یکی از راه‌کارهای مشکل فوق این است که طراح سؤال، به جای «هیچ یک از موارد»، گزینه را به شکل دقیق‌تر و متمرکزتری بنویسد. مثلاً اگر سؤال هدایت‌کننده این است که: «کدام دارو را تجویز می‌کنید؟»، می‌توان نوشت: «هیچ دارویی تجویز نمی‌کنیم». به این ترتیب ابهام موجود در سؤال کم می‌شود. البته مطالعات در این زمینه متعدد هستند و بعضاً یافته‌های متناقضی داشته‌اند. حتی یک مطالعه به این نتیجه رسیده است که در صورت استفاده مناسب و صحیح از این عبارت، چون احتمال حدس زدن دانشجو کم می شود، بهتر می‌توان دانشجوی توانمند را از ضعیف افتراق داد (بولند و همکاران[1] ۲۰۱۰).

- سؤال منفی: سؤال منفی یعنی در پایه سؤال از دانشجو بخواهیم گزینه غلط را تشخیص دهد. مثلاً بپرسیم «کدام گزینه درست نیست؟» یا «همه گزینه‌ها درست هستند، «به جز». توجه به این نکته ضروری است که شکل دوم که با «به جز» همراه شده است، هرچند از نظر ظاهری منفی نیست اما از نظر مفهومی منفی محسوب می‌شود و نباید استفاده شود.

علی‌رغم اینکه تأکید زیادی شده است که سؤال به صورت منفی طراحی نشود، طراحان همچنان از این نوع سؤالات طرح می‌کنند زیرا ساختن آنها آسان‌تر است. در این حالت طراحان به جای اینکه دنبال سه گزینه انحرافی باشند، سه گزینه درست در نظر می‌گیرند و تنها لازم است یک گزینه غلط بنویسند. در خصوص پیامدهای استفاده از این سؤال، بعضی از مطالعات ذکر کرده‌اند که استفاده از نوع مثبت یا منفی سؤال، تأثیری در عملکرد دانشجویان نداشته است اما برخی از تحقیقات آنها را موجب سردرگمی دانشجو و نشانه سختی بی‌مورد دانسته‌اند. البته مطالعاتی هم به این نتیجه رسیده‌اند که سؤال منفی، توانایی دانشجویان را بیش از حد واقعی برآورد می‌کند زیرا طراح با این تصور که سؤال پیچیده شده است، گزینه درست را طوری در نظر می‌گیرد که به راحتی قابل تشخیص باشد (بولند و همکاران ۲۰۱۰). در این رابطه شایان ذکر است که استفاده از سؤال منفی یک استثنا دارد که در فصل اول این بخش بیان شد.

شیوع و اهمیت وجود خطاهای ساختاری سؤالات چندگزینه‌ای

با توجه به کثرت انواع خطاهایی که نام برده شد، این سؤال پیش می‌آید که امتحاناتی که به صورت معمول برگزار می‌شوند، تا چه حد در معرض این خطاها هستند. مطالعات مختلفی در این زمینه انجام شده است که بر اساس یافته‌های آنها می‌توان گفت در مجموع شیوع انواع خطاها در آزمون‌ها واقعاً چشمگیر می‌باشد:

جوزفوویچ و همکاران ۲۰۰۲

نه امتحان در مقطع علوم پایه از سه دانشکده پزشکی در امریکا انتخاب شدند و سه نفر که چندین سال تجربه و تخصص مرور سؤالات در آزمون‌های USMLE و NBME داشتند، ۵۵۵ سؤال این امتحانات را به صورت مستقل از یکدیگر و بر اساس یک لیکرت ۵ تایی مورد ارزیابی قرار دادند، بدون اینکه از نام و مشخصات دانشکده آگاه باشند. به سؤالی نمره ۵ تعلق می‌گرفت که اولاً دارای سناریوی بالینی یا آزمایشگاهی مطلوب بود، ثانیاً واقعاً سؤال «چندگزینه‌ای با بهترین پاسخ» بود (و نه درست-نادرست)، و ثالثاً حاوی خطا نبود.
بر اساس نتایج مطالعه، میانگین نمره سؤالات در کل ۲/۳۹ (از ۵) به دست آمد که این میزان در سه دانشکده متفاوت بود. به طوری که دانشکده‌ای که استادان آن قبلا در دوره‌های NBME تحت آموزش قرار گرفته بودند، میانگین نمره بالاتری کسب کرده بود. در واقع، میانگین نمره ۹۲ سؤالی که توسط افراد آموزش دیده طراحی شده بود، ۴/۲۴ (در مقایسه با میانگین ۲/۰۳ بقیه سؤالات) بود و اختلاف بین آنها معنادار بود ($p<0.001$).

همچنین این مسأله که شیوع خطا در سؤالاتی که سطح شناختی پایینی دارند، به مراتب بیشتر است، بر اساس مطالعات تجربی نشان داده شده است:

1. Boland et al.

تارنت و همکاران ۲۰۰۶

طی سال‌های ۲۰۰۱ تا ۲۰۰۵ تمام آزمون‌های یک گروه پرستاری شامل ۲۲۷۰ سؤال جمع‌آوری شد و چهار نفر بر اساس یک چک‌لیست ۱۹ ایتمی به ارزیابی سؤالات پرداختند. ۱۲۸۰ سؤال یعنی ۴۶/۲ درصد، حاوی حداقل یک خطا تشخیص داده شدند و ۲۱ درصد بیشتر از یک خطا داشتند. بیشترین خطا مربوط به پایه مبهم و غیرشفاف (۷/۵ درصد) و استفاده از کلمات منفی (۶/۹ درصد) بود. از نظر سطوح شناختی، ۹۱/۱ درصد سؤالات در سطح یادآوری بودند که حدود نیمی از آنها حاوی خطا بودند. در حالی که میزان خطا در سؤالات با سطح شناختی بالا به صورت معناداری پایین تر بود (۱۸/۶ درصد).

وفامهر و دادگسترنیا ۱۳۸۹

در این مطالعه که در دانشگاه علوم پزشکی اصفهان انجام شد، آزمون‌های یک دوره چهار ساله دوره مقدمات پزشکی بالینی، شامل ۳۹۷۳ سؤال چهارگزینه‌ای جمع‌آوری شد. این سؤالات از نظر ۱۰ آیتم کیفی مربوط به وجود خطا در ساختار سؤال، سطح شناختی سؤال و ضریب دشواری و تمیز مورد بررسی قرار گرفتند. نتایج گویای آن بود که ۶۶/۹ درصد از سؤالات دارای حداقل یک خطا در ساختار خود بودند. ۷۹/۱ درصد سؤالات در سطوح پایین شناختی طراحی شده بودند. سؤالاتی که دارای سطح پایین شناختی بودند به طور معناداری دارای خطاهای ساختاری بیشتری بودند. ۴۷/۳ درصد سؤالات دارای ضریب تمیز مناسب و ۲۸/۸ درصد سؤالات دارای ضریب دشواری مناسب بودند. در حالی که ارتباطی بین ضریب دشواری و سطح شناختی سؤالات به دست نیامد، سؤالات با سطح شناختی بالاتر با ضریب تمیز مناسب رابطه معناداری داشتند.

اهمیت خطادار بودن سؤالات هنگامی پررنگ می‌شود که دیده می‌شود سؤالات خطادار در نمره دانشجویان و عملکرد آنها تاثیر دارند. هر چند دو مطالعه‌ای که در پی می آیند، در مورد جهت این تأثیر و میزان آن متفق‌القول نیستند:

داونینگ ۲۰۰۵

از بین سؤالات چهار آزمون مربوط به دانشکده‌های پزشکی امریکا که مورد بررسی قرار گرفتند، به طور متوسط ۴۶ درصد حاوی خطا بودند و میزان قبولی دانشجویان ۴۷ درصد به دست آمد. سپس سؤالات خطادار حذف شدند و بار دیگر نمرات دانشجویان محاسبه شد. میزان قبولی در حالت دوم، ۵۳ درصد محاسبه شد.
سؤالات خطادار بر پایایی آزمون تاثیر اندکی داشتند (احتمالاً به این دلیل که سؤالات خطادار منجر به خطای سیستماتیک می‌شوند و نه خطای تصادفی). حذف سؤالات خطادار صرفاً بر میزان قبولی اثر گذاشته بود و تاثیری بر عملکرد دانشجویان قوی نداشت. از نظر نویسنده خطاهای طراحی سؤال، خطای سیستماتیکی از نوع «واریانس بی ارتباط با سازه» ایجاد می‌کند و روایی آزمون را تحت تاثیر قرار می‌دهد.

تارنت و ویر ۲۰۱۰

بررسی ده امتحان مهم و سطح بالای دانشجویان پرستاری در انگلیس نشان داد که در صورت حذف سؤالات خطادار، میزان قبولی کاهش پیدا می‌کند (۹۰/۶ درصد در برابر ۴۹/۳ درصد). همچنین دانشجویان بیشتری نمره بالای ۸۰ کسب می‌کردند (۲۰/۹ درصد در برابر ۱۴/۵ درصد). محققین نتیجه گیری کردند که سؤالات خطادار به دانشجویان مرزی کمک کرده بود نمره بهتری کسب کنند و در عوض، به ضرر دانشجویان قوی بوده‌اند.

دلایل بروز خطاهای ساختاری سؤالات چندگزینه‌ای

موارد متعددی به عنوان دلیل بروز خطاهای ساختاری سؤال ذکر شده‌اند که در اینجا به صورت مختصر به آنها اشاره می‌شود:

- یکی از این دلایل، مهارت طراحان سؤال است. با وجود اینکه اغلب استادان علوم پزشکی فکر می‌کنند ذاتاً این توانایی را دارند که سؤالات خوبی طراحی کنند، تعداد کمی از آنها واقعاً در این زمینه آموزش رسمی دیده‌اند و با اصول طراحی صحیح سؤالات آشنا هستند. این در حالی است که طراحی سؤال مهارتی است که با تمرین زیاد و دریافت بازخورد مستمر توسط متخصصان امر شکل می‌گیرد (رودریگوئز ۱۹۹۷). این مهارت ذاتی نیست و چنانچه آموزش مناسب داده نشود، اکثر طراحان تازه‌کار سؤالاتی طراحی می‌کنند که خطا دارد، از کیفیت پایینی برخوردار است، موارد کم اهمیت را می‌سنجد و سطوح شناختی پایین را مورد ارزیابی قرار می‌دهد (داونینگ و هالادینا ۲۰۰۶).
- دلیل دیگر بروز خطاهای طراحی سؤال کمبود برنامه‌های آموزشی مناسب و اثربخش در زمینه طراحی سؤال است. در بسیاری

از موارد حتی دستورالعمل‌های طراحی سؤال در اختیار استادان قرار نمی‌گیرد. یعنی حتی اگر استادان به دنبال افزایش مهارت خود در این زمینه باشند، برنامه آموزشی که هدف‌مند، جذاب و متناسب با نیاز ایشان باشد، به ندرت در دسترس است. مهم است که اثربخشی دوره‌های آموزشی طراحی سؤال برای مدرسان به طور مداوم بررسی شود و تغییرات لازم در آنها ایجاد شود.

- مسأله دیگر نبود فرایند مرور[1] (توسط همکار یا یک کمیته) است که می‌تواند ناشی از کمبود وقت استادان یا عدم تمایل آنها برای انتقاد از همکارانشان باشد. اختصاص وقت اندک برای طراحی آزمون معضل بزرگی است. متاسفانه اغلب امتحانات داخلی در دانشکده‌ها در دقایق آخر و با سر هم کردن تعدادی سؤال طراحی می‌شوند، به طوری که حتی فرصتی برای مرور سؤالات نیست (جوزفوویچ و همکاران ۱۹۹۸). اهمیت این مسأله هنگامی بیشتر نمایان می‌شود که استادان یک درس متعدد هستند و هر کس تنها از بخش مربوط به خود سؤال طرح می‌کند. در این حالت، نیاز به ارزیابی سؤالات برای حذف سؤالات تکراری، هماهنگی سؤالات با یکدیگر، یکدست کردن آنها و اطمینان از کیفیت کلی دفترچه آزمون بیشتر احساس می‌شود اما عملاً چنین کاری به ندرت صورت می‌پذیرد.

مجموع عوامل فوق منجر می‌شود که تعدادی سؤال ناهمگون شامل انواع «چندگزینه‌ای با بهترین پاسخ»، «درست–نادرست متعدد» و «جای خالی»، با کیفیت‌های مختلف در دفترچه ردیف می‌شوند. این موضوع از این جهت قابل تأمل است که طراحان سؤال معمولاً کسانی نیستند که به آموزش بی‌اعتنا باشند. آنها واقعاً برای کلاس درس خود اهمیت زیادی قائل می‌شوند و ساعت‌های متمادی برای آماده کردن سخنرانی خود وقت صرف می‌کنند اما به نظر می‌رسد آن جدیت و تعهد را در هنگام ارزیابی دانشجویان از خود نشان نمی‌دهند. در نتیجه استادان خیلی خوب ممکن است امتحاناتی بگیرند که از لحاظ کیفیت چندان مورد قبول نیست (جوزفوویچ و همکاران ۱۹۹۸). آزمونی که از سؤالات بدون خطا تشکیل شده است، به دانشجویان نشان می‌دهد که مسؤولان آموزشی و استادان به تمام جنبه‌های یادگیری و آموزش آنها توجه می‌کنند. طراحان سؤال باید به این موضوع توجه کنند که آزمون جهت‌دهنده مسیر یادگیری دانشجویان است و هر امتحان تاثیری بر یادگیری دانشجویان دارد. دانشجویان آنچه را که در امتحان مورد سؤال قرار می‌گیرد، مهم می‌پندارند و با تأکید بیشتر آن را مطالعه می‌کنند.

پیشنهادها برای کاهش خطای سؤالات چندگزینه‌ای

بر اساس مشاهدات و مطالعات صورت گرفته به منظور ارتقاء کیفیت سؤالات چندگزینه‌ای می‌توان اقدامات زیر را انجام داد:

- برگزاری دوره‌های توانمندسازی استادان در حوزه ارزیابی دانشجو: برگزاری کارگاه‌ها و دوره‌های آموزشی برای طراحان در افزایش مهارت طراحی سؤال آنها می‌تواند کمک‌کننده باشد. مهم است که محتویات کارگاه به صورت هدف‌مند، کاربردی و متناسب با نیاز اعضای هیات علمی در نظر گرفته شود و همچنین شیوه ارائه مطالب به صورت جذاب و متناسب با علائق ایشان باشد. شواهد نشان می‌دهند که برای افزایش اثربخشی آموزش‌ها بهتر است برنامه‌های آموزشی به صورت هفته‌ای یک‌بار و طولانی‌مدت در نظر گرفته شوند و محدود به دوره‌های کوتاه‌مدت و مجزا نباشند (دارلینگ-هاموند[2] ۱۹۹۷ و ۱۹۹۹، هیل[3] ۲۰۰۷).
- ارزشیابی دوره‌های توانمندسازی استادان در حوزه ارزیابی دانشجو: دانشکده‌های مختلف ممکن است به صورت مکرر دوره‌هایی برای توانمندسازی استادان تشکیل دهند. اما آنچه حائز اهمیت است این است که نظرات و بازخوردهای شرکت‌کنندگان به صورت مداوم دریافت شود تا دوره‌های با کیفیت‌تری برگزار شوند. ارزشیابی دوره‌ها از منظر اثربخشی آنها نیز باید جدی گرفته شود. به این معنا که بررسی شود آیا با گذراندن این دوره‌ها واقعاً کیفیت سؤالات ارتقاء پیدا کرده است یا خیر. در بررسی متون و شواهد می‌بینیم که با وجود برگزاری دوره‌های متعدد توانمندسازی استادان، مطالعات اندکی به بررسی اثربخشی آنها پرداخته‌اند.

1. Review
2. Darling-Hammond
3. Hill

نعیم و همکاران ۲۰۱۱

در این مطالعه، میزان اثربخشی یک دوره آموزشی یک هفته‌ای در زمینه طراحی سؤال در دانشگاه آقاخان پاکستان بررسی شد. ۵۱ استاد در دوره شرکت کردند. از هر یک از آنها خواسته شد تا قبل از شروع دوره، در اواسط دوره و در انتهای آن یک سؤال که به نظر خودشان بهترین است، ارائه دهند. این سؤالات توسط فرد متخصص و بر اساس چک‌لیستی که قبلا تهیه شده بود مورد ارزیابی قرار گرفت. نتایج نشان داد که نمرات افراد در این سه نوبت به طور معنادار افزایش داشته است.

معیاری و همکاران ۱۳۹۱

هدف از انجام این مطالعه، تعیین تاثیر کارگاه آموزشی در مورد آزمون های چندگزینه‌ای، بر بهبود کیفیت طراحی سؤالات آزمون ارتقای دستیاری دانشکده دندانپزشکی همدان بود. کلیه سؤالات آزمون های ارتقای دستیاری دانشکده دندانپزشکی همدان در سال ۸۷ از نظر ساختار طراحی و سطح تاکسونومی مورد بررسی قرار گرفت. بعد از برگزاری کارگاه آموزشی جهت طراحان آزمون در هر دو سال، سؤالات آزمون ارتقای سال ۸۸ نیز مورد بررسی قرار گرفت. داده‌های قبل و بعد از مداخله مقایسه گردید.
از نظر ساختار کلی، از کل ۱۲۳۹ سؤال، به ترتیب ۶۳/۱ درصد و ۷۶/۳ درصد سؤالات مربوط به سال های ۸۷ و ۸۸ بدون اشکال طراحی شده بود که این تفاوت از نظر آماری معنادار بود. درصد سؤالات طراحی شده با تاکسونومی بالا نیز در سال ۸۸ نسبت به سال ۸۷ به طور معناداری افزایش یافته بود. نویسندگان نتیجه‌گیری کردند که با توجه به نتایج مطالعه، طراحی و برگزاری کارگاه آموزشی، جهت طراحان سؤال موثر بود و برگزاری این کارگاه‌ها می تواند راه‌گشایی برای بهبود آزمون‌های مشابه باشد.

خوشرنگ و همکاران ۱۳۹۲

هدف از انجام این مطالعه، تعیین تاثیر کارگاه آموزشی طراحی سؤالات چندگزینه‌ای بر بهبود کیفیت طراحی سؤالات آزمون ارتقای دستیاری دانشکده پزشکی دانشگاه علوم پزشکی گیلان بود. در این مطالعه کلیه سؤالات مربوط به ۱۲ رشته تخصصی در سال ۸۹ از نظر ساختار طراحی و سطح تاکسونومی بررسی شد. بعد از برگزاری ۱۲ دوره کارگاه آموزشی یک روزه، سؤالات آزمون ارتقای سال ۹۰ نیز مورد بررسی قرار گرفت. در آزمون کتبی سال اول، ۳۱ سؤال (۱/۷ درصد) و در سال دوم، ۱۹۲ سؤال (۱۰/۶ درصد) تاکسونومی سه داشتند که تفاوت آنها معنادار بود. نسبت سؤالات بدون اشکال در آزمون سال اول و دوم به ترتیب، ۶۶/۷۷ درصد و ۷۵/۲۲ درصد بود که تفاوت معنادار بین آنها وجود داشت. نویسندگان نتیجه‌گیری کردند که پیشرفت معنادار در کاربرد سؤالات مربوط به سطوح عمیق دانش در آزمون ارتقای، بیانگر تاثیر مثبت مداخله آموزشی انجام شده بود

- آماده‌سازی دستورالعمل‌های طراحی سؤال: جدا از برنامه‌های کلی توانمندسازی، هر آزمونی که در دانشکده برگزار می‌شود، فرصتی برای ارتقاء مهارت طراحی سؤال استادان است. به این ترتیب، آماده کردن دستورالعمل‌ها و چک‌لیست‌های ارزیابی سؤال و ارائه آنها به استادانی که قرار است سؤال طراحی کنند، می‌تواند کمک‌کننده باشد.
- تشکیل کمیته مرور سؤالات: در ادامه پیشنهاد فوق، می‌توان برای آزمون‌ها، مخصوصاً آزمون‌های مهم و سطح بالای دانشکده، کمیته مرور سؤال تشکیل داد که در آن کارشناسان آموزش دیده در کنار خود طراحان سؤال به مرور سؤالات بپردازند. در کمیته، تک‌تک سؤالات از نظر محتوا، سطح شناختی، خطاهای ساختاری، نکات نگارشی و ... بررسی می‌شوند تا در صورت نیاز، بازبینی یا جایگزینی با سؤالات بهتر صورت گیرد (والاش و همکاران[1] ۲۰۰۶).

والاش و همکاران ۲۰۰۶

در این مطالعه، ۵۲۰ سؤال به صورت تصادفی از آزمون‌های دو سال تحصیلی انتخاب شدند. ۲۵۰ سؤال که مربوط به سال تحصیلی ۲۰۰۱-۲۰۰۰ بود، مربوط به آزمون‌های مبتنی بر دیسیپلین بودند که توسط گروه‌های آموزشی به صورت جداگانه برگزار می‌شدند. ۲۷۰ سؤال دیگر مربوط به سال ۲۰۰۲-۲۰۰۱ (بعد از بازنگری) بود که محتوای آموزشی به صورت ادغام‌یافته در قالب کورس‌های مبتنی بر ارگان-سیستم توسط همان گروه‌های آموزشی قبلی ارائه می‌شد. در این سال، قبل از طراحی سؤال، گایدلاین‌ها در اختیار استادان قرار گرفت و همچنین سؤالات، درون کمیته بین‌رشته‌ای با صدای بلند خوانده و مرور شد، اشتباهات گرامری و نگارشی ویرایش شد و مواردی که از نظر اصول طراحی سؤال، واجد خطا بود یا نامرتبط تشخیص داده شد، همراه با بازخورد و پیشنهادهای اصلاحی برای ویرایش در اختیار طراح اصلی قرار گرفت.
سپس سؤالات به صورت تصادفی و کور در اختیار متخصصان NBME قرار گرفت و هر یک از آنها به صورت انفرادی سؤالات را بر اساس یک چک لیست مشخص بررسی کرد و طبق یک لیکرت ۵ تایی نمره داد. میانگین نمرات سؤالات مربوط به آزمون‌های قبل از بازنگری، ۲/۵۱ با انحراف معیار ۱/۲۷ بود. این شاخص برای سؤالات سال تحصیلی بعدی، ۳/۱۶ با انحراف معیار ۱/۳۳ بود که اختلاف بین آنها معنادار بود.

1. Wallach et al.

- تغییر استراتژی در زمان‌بندی آزمون: تأکید این راهبرد، تغییر برنامه زمان‌بندی طراحی سؤالات آزمون است. با این روش از تهیه سؤالات در دقایق پایانی اجتناب می‌شود و به این ترتیب وقت کافی برای مرور و بازبینی سؤالات نیز فراهم می‌شود. آن گونه که در مطالعه جوزفوویچ پیشنهاد شده است، مناسب است آغاز روند تهیه سؤالات، ۲ تا ۳ هفته زودتر از موعد امتحان در نظر گرفته شود (جوزفوویچ و همکاران ۱۹۹۸). در این حالت، می‌توان زمان کافی برای مرور و بررسی سؤالات اختصاص داد (جدول ۱-۶).

جدول ۱-۶: جدول زمان‌بندی پیشنهادی برای یک امتحان

	عنوان	موعد
۱	طرح سؤال توسط مدرسان	سه هفته مانده به امتحان
۲	جمع‌آوری سؤالات و تایپ آنها	دو هفته مانده به امتحان
۳	برگزاری جلسات مرور سؤالات	یک هفته مانده به امتحان
۴	نهایی کردن دفترچه	سه روز مانده به امتحان
۵	تکثیر دفترچه	یک روز مانده به امتحان
۶	امتحان	

گام‌های طراحی سؤال «چندگزینه‌ای با بهترین پاسخ»

همه کسانی که به امر تدریس اشتغال دارند، به کرات سؤال چندگزینه‌ای طراحی کرده‌اند اما همان‌طور که گفته شد، رعایت اصول و قواعدی که برای طراحی نوع مطلوب این سؤال توصیه شده است، به راحتی میسر نمی‌شود. در واقع این مهم در عمل با تمرین فراوان، مرور سؤالات و دریافت بازخورد به دست می‌آید. انواع راهنماها و دستورالعمل‌ها برای طراحی سؤال چندگزینه‌ای تدوین شده‌اند (هالدینا و همکاران[1] ۲۰۰۲، کولینز ۲۰۰۶، بولند و همکاران ۲۰۱۰)

در عین حال که رعایت این اصول توصیه شده است، باید توجه داشت که این دستورالعمل‌ها، قوانین اثبات شده نیستند بلکه تلاشی به منظور جمع‌بندی نظرات و تجربیات متخصصان امر در راستای طراحی بهتر سؤال هستند. در خیلی از موارد، مطالعات برای مشاهده نحوه اثربخشی آنها و تاثیر به کارگیری آنها در عملکرد دانشجویان صورت نگرفته است. البته مطالعاتی نیز وجود دارد که در این زمینه تلاش کرده‌اند اما در این موارد نیز اگرچه در برخی زمینه‌ها شواهد، دستورالعمل‌ها را تایید می‌کنند (مثلاً «متجانس بودن گزینه‌ها» هم بر روایی آزمون تاثیر دارد و هم دشواری و تمیز را افزایش می‌دهد)، به نظر نمی‌رسد شواهد تجربی کافی در حمایت از «همه» این اصول وجود داشته باشد.

به صورت کلی، با وجود کثرت راهنماهای تدوین شده در زمینه طراحی سؤال چندگزینه‌ای و همچنین، وجود مواردی متناقص در دستورالعمل‌های مختلف، موارد اشتراک موجود در آنها بسیار زیاد است. در اینجا، کل فرایند طراحی سؤال چندگزینه‌ای در شش مرحله با ذکر اصول و مثال‌های مرتبط توضیح داده می‌شود. در جدول ۲-۶ خلاصه‌ای از مراحل طراحی سؤالات چندگزینه‌ای نشان داده شده است.

1. Haladyna et al.

جدول ۲-۶: خلاصه مراحل طراحی سؤالات چندگزینه‌ای

ردیف	عنوان مرحله	توضیح
۱	انتخاب یک موضوع مناسب و مهم	موضوعات قسمتی از اطلاعات و دانش داوطلب هستند که قرار است سؤال آن را بسنجد. این موضوعات در ارتباط با اهداف دوره هستند که قرار است مورد ارزیابی قرار گیرد.
۲	انتخاب محور مناسب برای سؤال	محور موقعیتی است که تعیین می‌کند چه اطلاعاتی (تفسیر داده‌ها، تشخیص، مراقبت از بیمار) باید در سؤال ارایه شود.
۳	نگارش پایه سؤال	پایه سؤال بهتر است یک مورد بالینی باشد. علاوه بر این باید حاوی تمام اطلاعاتی باشد که برای پاسخ به سؤال ضروری است.
۴	نگارش سؤال هدایت‌کننده	سؤال هدایت‌کننده، رابط بین گزینه‌ها و پایه سؤال است و مشخص می‌کند که دانشجو چه کاری باید انجام دهد.
۵	طراحی گزینه‌ها	یک گزینه کاملاً صحیح است و گزینه‌های دیگر که انحرافی هستند، باید مشخصاً غلط بوده اما در عین حال برای داوطلبان ضعیف جذاب باشند.
۶	مرور و ارزیابی سؤال	سؤال توسط خود طراح، همکاران و کمیته آزمون توسط چک لیست ارزیابی می‌شود.

انتخاب یک موضوع مناسب و مهم برای طرح سؤال

برای انتخاب موضوعاتی که قرار است از آنها سؤال طرح شود، داشتن بلوپرینت دوره یا جدول مشخصات آزمون به منظور تضمین روایی آزمون ضروری است. می‌توان فهرستی از مشکلاتی که به طور معمول دانشجو قرار است در فعالیت حرفه‌ای خود با آن مواجه شود، تهیه کرد. این فهرست راهنمای خوبی برای انتخاب موضوع سؤالات خواهد بود. در برخی از امتحانات، مخصوصاً امتحانات جامع و کشوری، موضوعات توسط مسؤول کمیته امتحانات از میان بلوپرینت آزمون انتخاب و به طراحان سؤال ارائه می‌شود. در این صورت، هر سؤال باید مرتبط با محتوای مشخصی از بلوپرینت آزمون باشد. برای اطلاعات بیشتر در خصوص بلوپرینت به اولین بخش کتاب مراجعه کنید.

رعایت این نکته بسیار توصیه شده است که محتوای هر سؤال باید مستقل از محتوای سؤالات دیگر باشد. پاسخ دادن به یک سؤال نباید در پاسخ دادن به سؤال بعدی به دانشجو کمک کند یا پیش‌شرط پاسخ به سؤال دیگر باشد چرا که پاسخ غلط به سؤال اول، پاسخ غلط به سؤال دوم را در پی دارد و این از انصاف به دور است.

در انتخاب موضوع، سطح دشواری سؤال باید مدنظر قرار بگیرد. دشواری سؤال باید متناسب با سطح شرکت‌کنندگان در آزمون باشد. سنجیدن سطح علمی دانشجویان الزاماً به معنای طرح سؤالات بسیار دشوار نیست. اگر سؤال برای سنجیدن اطلاعات علمی داوطلب در مورد مسایلی است که برای کار حرفه‌ای او الزامی است، ممکن است بسیار ساده باشد.

موضوع سؤال باید روی یک موضوع «مهم» تمرکز داشته باشد. بهتر است از طرح سؤال در مورد موضوعات کم اهمیت یا موارد ناشایع صرف نظر شود و سؤالاتی طراحی گردند که از لحاظ کاربردی مهم و مرتبط با آینده حرفه‌ای فرد باشند.

بهتر است از سؤالاتی که صرفاً اطلاعات نظری و محفوظات را می‌سنجند، پرهیز شود و از موضوعاتی استفاده شود که سطوح بالاتر یادگیری را می‌سنجند. سؤالات تستی چهارجوابی که خوب طراحی شده‌اند، به جای آن که صرفاً یادآوری اطلاعات را مورد ارزیابی قرار دهند، به بررسی میزان به کارگیری این اطلاعات توسط داوطلب می‌پردازند. در مورد سطح شناختی در فصل اول همین بخش توضیحاتی ارائه شد و مجدداً در قسمت نگارش پایه سؤال نیز به آن اشاره می‌شود.

علاوه بر موارد فوق، هر سؤال باید بر اساس «یک» محتوای مهم طرح شود. این مسأله به این معناست که نباید بیش از یک جنبه از یک موضوع در سؤال مورد پرسش قرار گیرد. به مثال صفحه بعد توجه کنید:

سؤال ضعیف			
در مورد مننژیت باکتریال کدام نادرست است؟			
الف) با افزایش پروتئین، سلول و قند مایع CSF همراه است.		ب) ارزش رنگ‌آمیزی گرم در تشخیص عامل آن اندک است.	
ج) در نوزادان شایع‌ترین عامل آن هموفیلوس آنفلونزا است.		د) محل اگزودای چرکی به عامل اتیولوژیک بستگی دارد.	
سؤال بهتر			
شایع‌ترین علت مننژیت باکتریال در نوزادان چیست؟			
الف) هموفیلوس آنفلونزا	ب) نیسریا مننژیتیس	ج) منونوکلئوز عفونی	د) پنوموکوکوس

در سؤال اول، هر یک از گزینه‌ها به یک موضوع جداگانه می‌پردازد و پراکنده است. هرچند برخی از استادان این گونه استدلال می‌کنند که به این روش می‌توان چند مطلب مختلف را در آن واحد سنجید و از این نوع سؤال استقبال می‌کنند، طراحی این نوع سؤال به علت سردرگمی و ابهامی که برای دانشجو ایجاد می‌کند، توصیه نمی‌شود. در مثال زیر نشان داده شده است که چگونه می‌توان با رفع این اشکال، سؤال بهتری طرح کرد:

سؤال ضعیف			
دور سر نوزاد:			
الف) با قد وی ارتباط بیشتری دارد.		ب) باید با وزن او مقایسه شود.	
ج) و دور سینه او رشد مشابهی دارند.		د) و دور بازوی وی به صورت همراه بررسی می‌شوند.	
سؤال بهتر			
دور سر نوزاد با کدام یک از شاخص‌های رشدی زیر بیشتر مرتبط است؟			
الف) قد	ب) وزن	ج) دور سینه	د) دور بازو

انتخاب محور مناسب برای سؤال

محور سؤال، بیان‌کننده موقعیتی است که موضوع را مورد ارزیابی قرار خواهد داد. محور در یک سؤال بسیار مهم است زیرا اولاً تعیین می‌کند چه اطلاعاتی باید در پایه سؤال گنجانده شود و دوم این که مشخص می‌کند گزینه‌ها چه باید باشند. به دو مثال زیر توجه کنید:

موضوع: سندرم ترنر و محور: معاینه بالینی			
یک دختر ۱۸ ساله با آمنوره اولیه مراجعه کرده است. کدام یک از نشانه‌های زیر بیشتر به نفع تشخیص سندرم ترنر می‌باشد؟			
الف) پرفشاری خون	ب) هیرسوتیسم	ج) کوتاهی قد	د) چین‌های اپی کانتال

موضوع: سندرم ترنر و محور: تشخیص	
خانمی ۱۸ ساله به علت آمنوره اولیه مراجعه کرده است. در معاینه متوجه می شوید که قد بیمار ۱۴۸ سانتی‌متر است. علاوه بر این متوجه می‌شوید که اندام تناسلی خارجی بیمار نابالغ بوده و تکامل پستان ایجاد نشده است. محتمل ترین تشخیص کدام است؟	
الف) سندرم ترنر	ب) دیس ژنزی گنادی مرکب
ج) دیس ژنزی گنادی خالص	د) سندرم نونان

همان‌گونه که مشاهده می‌کنید، هر دو تست یک موضوع یعنی سندرم ترنر را می‌سنجند. ولی محور دو تست متفاوت است. همین تغییر باعث شده است که اطلاعاتی که در دو سؤال ارایه شده است، کاملاً تفاوت داشته باشد. محور سؤال اول، معاینه بالینی است. به همین دلیل پایه و گزینه‌ها نیز حاوی اطلاعاتی است که انتظار می‌رود در جریان یک معاینه به دست آید. در حالی که در تست دوم،

محور سؤال، تشخیص است بنابراین پایه سؤال حاوی علایم و نشانه‌هاست و گزینه‌ها، حاوی تشخیص‌های احتمالی هستند.

محورهایی که برای طراحی سؤالات چند جوابی در مباحث بالینی مورد استفاده قرار می‌گیرد، می‌تواند به صورت زیر باشد:

- عوامل خطر[1]
- شرح‌حال
- معاینه
- تست آزمایشگاهی
- تفسیر اطلاعات
- تشخیص و تشخیص افتراقی
- مراقبت‌های اولیه
- عوارض جانبی دارو و موارد منع مصرف
- مشاوره
- ملاحظات اخلاقی
- و ...

نگارش پایه سؤال

مرحله بعدی نگارش پایه سؤال است. ایده اصلی سؤال باید در پایه سؤال مطرح شود؛ نه در گزینه‌ها. به همین دلیل پایه سؤال خوب معمولاً طولانی است در حالی که گزینه‌ها کوتاه هستند.

سؤال ضعیف
بیشترین میزان نیروی تولید شده در انقباض: الف) زمانی حاصل می شود که طول سارکومر بیش از ۲/۲ میکرومتر باشد. ب) زمانی حاصل می شود که طول سارکومر بین ۲ تا ۲/۲ میکرومتر باشد. ج) زمانی حاصل می شود که طول سارکومر کمتر از ۲ میکرومتر باشد. د) زمانی حاصل می شود که طول سارکومر بیش از ۳ میکرومتر باشد.
سؤال بهتر
بیشترین میزان نیروی تولید شده در انقباض زمانی حاصل می‌شود که طول سارکومر چند میکرومتر باشد؟ الف) کمتر از ۲ — ب) بین ۲ تا ۲/۲ ج) بیش از ۲/۲ — د) بیش از ۳

تمامی اطلاعاتی که یک دانشجوی توانمند برای پاسخ‌گویی به سؤال لازم دارد باید در پایه سؤال آورده شود. برای اطمینان از این موضوع، می‌توان از آزمون پوشاندن گزینه‌ها[2] استفاده کرد. اگر بتوان بدون مطالعه گزینه‌ها به سؤال پاسخ داد، یعنی سؤال خوب طراحی شده است. برای درک بهتر این موضوع به مثال زیر توجه کنید:

سؤال ضعیف
در مورد درمان سندرم گیلن‌باره کدام درست است؟ الف) گاماگلوبولین داخل وریدی — ب) پلاسمافرز ج) کورتیکواستروئیدها — د) درمان حمایتی

1. Risk Factors
2. Cover the options test

سؤال بهتر

درمان انتخابی در سندرم گیلن‌باره کدام است؟

الف) گاماگلوبولین داخل وریدی
ب) پلاسمافرز
ج) کورتیکواستروئیدها
د) اقدامات حمایتی

در واقع، سؤال اول یک سؤال «درست–نادرست متعدد» است که در ظاهر شبیه سؤال «چندگزینه‌ای با بهترین پاسخ» است اما باید از آن اجتناب شود.

هنگام نگارش پایه سؤال، این نکته باید مدنظر قرار گیرد که خواندن پایه طولانی برای دانشجو وقت‌گیر است پس برای محاسبه زمان کلی آزمون، علاوه بر تعداد سؤالات باید حجم هر سؤال نیز لحاظ شود. به طور معمول، دانشجو طی یک دقیقه، به یک تا دو تست جواب می‌دهد (کولینز ۲۰۰۶). اگر به صورت محسوسی غیر از این باشد و دانشجویان وقت کم بیاورند، بهتراست سؤال بررسی شود و علت سردرگمی دانشجویان مشخص شود.

علاوه بر موارد فوق، از طرح نکته انحرافی[1] در پایه سؤال باید خودداری کرد. توجه به این نکته ضروری است که زمان آموزش دادن گذشته است و وقت دانشجو در جلسه آزمون نباید برای یاد گرفتن مطالب جدید تلف شود.

این مسأله نیز در متون مورد اشاره قرار گرفته است که پایه سؤال بهتر است به صورت جای خالی نوشته نشود و اصولاً نباید شروع پایه با جای خالی باشد. مانند نمونه زیر:

مثالی از جای خالی در آغاز پایه سؤال

..... زمانی است که دور سر نوزادان نارس به حد نوزادان ترم هم سن می‌رسد.

الف) سه ماهگی
ب) شش ماهگی
ج) نه ماهگی
د) دوازده ماهگی

سایر نکاتی که در خصوص پایه سؤال ذکر شده است، شامل این موارد است: باید اطمینان یافت که اطلاعات مندرج در پایه سؤال، واضح، روان و بدون ابهام است. زبان مورد استفاده باید متناسب با گروه داوطلبان باشد. بدنه سؤال از نظر لغوی و دستوری نباید پیچیده باشد و از استفاده از مخففها[2] خودداری گردد. از کلمات و عباراتی که در سطح کشور و در میان جامعه علمی مورد قبول هستند، استفاده شود و واحدهای یافته‌های آزمایشگاهی قید شود.

سؤال مخدوش

اگر مغز کودک مبتلا به بیماری شربت افرا (MSUD) را در آورده و آن را ببینیم مانند کدام بیماری زیر است؟

الف) Phenylketonuria (PKU)
ب) Glycogen Storage Disease type I (GSD type I)
ج) GSD type II
د) Thyrosinemia

سؤال بهتر

تغییرات پاتوبیولوژیک مغزی در بیماران Maple Syrup Urine Disease به کدام‌یک از بیماری‌های زیر شبیه‌تر است؟

الف) Phenylketonuia (PKU)
ب) Glycogen Storage Disease type I (GSD type I)
ج) GSD type II
د) Thyrosinemia

1. Red herring
2. Acronym

پایه سؤال باید به صورت جمله مثبت نوشته شود و از به کار بردن جملاتی مانند «کدام گزینه درست نیست» یا «همه گزینه‌ها درست هستند، به جز» پرهیز شود. همان‌طور که قبلاً ذکر شد، به صورت کلی سؤالات منفی توصیه نمی‌شوند زیرا هنگامی که دانشجو به درستی تشخیص می‌دهد گزینه‌ای نادرست است، نمی‌توان دریافت که جواب صحیح را می‌داند یا خیر و همچنین، اکثر اهداف آموزشی وقتی بهتر مورد سنجش قرار می‌گیرند که از دانشجو خواسته شود آنچه می‌داند درست است، انتخاب کند تا اینکه بگوید چه مواردی نادرست هستند. البته این نیز اشاره شد که گاهی با رعایت شرایط می‌توان از سؤال منفی استفاده کرد. به عنوان مثال، پزشکی که قرار است با توجه به وضعیت بیمار برای او تصمیم‌گیری کند، گاهی لازم است حتماً کاری را انجام ندهد. در حقیقت در این موارد دانستن این موضوع که چه اقداماتی نباید انجام شود، ضروری است و اولویت دارد و می‌توان با احتیاط از سؤال منفی استفاده کرد. البته سؤال منفی باید به گونه‌ای باشد که باعث سردرگمی دانشجو نشود. در این صورت، اولاً پایه سؤال باید به صورت منفی طراحی شود نه گزینه‌های آن و دوم این که، قسمت منفی باید حتماً در متن با قلم برجسته[1] یا خط‌کشی زیر آن مشخص و متمایز شود تا توجه داوطلبان به آن جلب شود.

از طراحی سؤالات منفی مضاعف حتماً باید پرهیز شود. منظور از منفی مضاعف به کارگیری توأم جملات منفی در پایه سؤال و گزینه‌ها یا به کار بردن دو علامت منفی‌ساز در پایه سؤال است؛ چنین حالاتی آزمون شونده را دچار سردرگمی می‌کند. در مثال زیر، از آنجا که سؤال اول واجد چندین خطا از جمله به کار بردن منفی مضاعف است، به هیچ‌وجه پذیرفته‌شده نیست و حتماً باید ویرایش شود. در سؤال دوم، هر چند که خطاهای دیگر همچنان وجود دارند و سؤال مطلوبی نیست، از حالت اول کمی بهتر است.

سؤال ضعیف

در مورد فلج مغزی (Cerebral palsy) نمی‌توان گفت که:

الف) شایعترین علت آن آسفیکسی زایمانی (Encephalopathy Hypoxic Ischemic) نیست.
ب) در اغلب موارد همراه با عقب ماندگی ذهنی نیست.
ج) اشکال حرکتی مرکزی و غیر پیش‌رونده بوده که معمولاً تا قبل از یک سالگی پدید نمی‌آید.
د) تشنج در این بیماران به صورت همراه دیده نمی‌شود.

سؤال بهتر

کدام یک از عبارات زیر در مورد فلج مغزی (Cerebral palsy) صدق نمی‌کند؟

الف) شایعترین علت آن آسفیکسی زایمانی (Hypoxic Ischemic Encephalopathy) است.
ب) در اغلب موارد همراه با عقب ماندگی ذهنی است.
ج) اشکال حرکتی مرکزی و غیر پیش‌رونده معمولاً تا یک سالگی پدید می‌آید.
د) تشنج در این بیماران به صورت همراه دیده می‌شود.

به این موضوع بارها اشاره شد که یک سؤال چندگزینه‌ای مطلوب بیش از آنکه بر حفظیات تمرکز کند، باید توانایی داوطلب در به کار بستن دانش پزشکی را بسنجد. یکی از مواردی که برای سنجش کاربرد دانش توصیه می‌شود، استفاده از یک سناریوی بالینی در پایه سؤال است. با این رویکرد، علاوه بر اینکه تاکسونومی سؤال افزایش می‌یابد و سطوح شناختی بالاتر مورد ارزیابی قرار می‌گیرد، سؤال برای داوطلب از نظر بالینی مرتبط‌تر با محیط کار واقعی به نظر می‌رسد و دارای روایی صوری بیشتر است. موردهای بالینی معمولاً با مطرح کردن یک مشکل شروع می‌شوند و در ادامه علایم و نشانه‌های مرتبط، نتایج آزمایشات، اقدامات درمانی اولیه و یافته‌های متعاقب این اقدامات ذکر می‌شود.

شاید تصور شود استفاده از سناریوی بالینی در طراحی سؤال دروس علوم پایه جایگاهی ندارد اما در واقع، بسیاری از

1. bold

مطالب علوم پایه را نیز می‌توان در قالب سناریوی بالینی بیان کرد. اتفاقاً با ارائه سؤال به این شکل، دانشجوی دوره علوم پایه متوجه می‌شود که آنچه در این مقطع مطالعه می‌کند، در حرفه آینده او کاربرد دارد. البته هنگام استفاده از اصطلاحات و واژه‌ها در پایه سؤال باید مراقب بود که دانشجو در این مقطع هنوز اطلاعات بالینی چندانی ندارد و شاید لازم باشد اطلاعات بیشتری در متن سؤال ارائه شود. طراحی چنین سؤالی نیازمند تجربه و تمرین است. به مثال‌های زیر که نحوه استفاده از مشکل بالینی را برای سنجش دانش علوم پایه نشان می‌دهند، توجه کنید:

سؤال آناتومی با موضوع مشابه در دو سطح شناختی

کدام یک از موارد زیر عمل عضله گلوتئوس ماکزیموس است؟

الف) فلکسیون پا ب) اکستانسیون زانو ج) اکستانسیون هیپ د) فلکسیون هیپ

سؤال بهتر

مرد ۵۶ ساله ای در بلند شدن از حالت نشسته و راست کردن تنه خود مشکل دارد ولی با خم کردن پای خود مشکلی ندارد. کدام یک از عضلات زیر به احتمال بیشتری درگیر است؟

الف) عضله گلوتئوس ماکزیموس ب) گلوتئوس مینیموس ج) هامسترینگ د) ایلیوپسواس

سؤال تغذیه با موضوع مشابه در دو سطح شناختی

کدام یک از موارد زیر هنگام سوختن کالری بیشتری را تولید می‌کنند؟

الف) لیپیدها ب) پروتئین‌ها ج) کربوهیدرات‌ها د) رزین‌ها

شکل بهتر

خانم کارمندی به علت افزایش وزن به شما مراجعه کرده است. او می‌خواهد بداند حذف کدام یک از مواد غذایی زیر از رژیم روزانه باعث کاهش بیشتری در کالری دریافتی‌اش خواهد شد؟

الف) یک قاشق کره ب) یک قاشق سفیده تخم مرغ ج) یک قاشق شکر د) یک کاسه کاهو

بنابراین، از سناریوی بالینی می‌توان در طراحی سؤال دروس علوم پایه نیز استفاده کرد. چارچوب‌های مشخصی برای نوشتن پایه سؤالات این چنینی پیشنهاد شده است. یکی از شایع‌ترین این چارچوب‌ها، که مخصوصاً برای درس آناتومی به خوبی استفاده می‌شود؛ به صورت زیر است:

مواردی که در توضیح بیمار می‌توان استفاده کرد:

- سن و جنس (آقای ۴۵ ساله)
- محل مراجعه (به درمانگاه مراجعه کرده است)
- شکایت اصلی[1] (مبتلا به سردرد)
- مدت بیماری (از سه روز قبل)
- سابقه بیمار (با سابقه خانوادگی فشار خون)
- یافته‌های معاینه
- نتیجه تست‌های آزمایشگاهی

سایر چارچوب‌هایی که برای طراحی پایه سؤال در دروس مربوط به علوم پایه (مانند آناتومی، فیزیولوژی، تغذیه، آسیب‌شناسی، فارماکولوژی، باکتری‌شناسی و ...) توصیه شده‌اند، به صورت زیر است:

1. Chief complaint

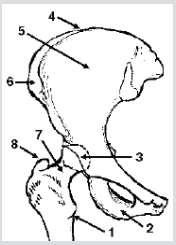
چارچوب‌های توصیه شده برای طراحی پایه سؤال در علوم پایه

یک [بیمار]، مبتلا به [آسیب و محل آن] می‌باشد. احتمالا کدام ساختار آسیب دیده است؟
یک [بیمار]، با [معاینه وشرح حال]، تحت درمان است. کدام یک از داروهای زیر موجب [یافته معاینه یا شرح حال] در او شده است؟
یک [بیمار] با [یافته‌های معاینه، شرح حال، تست آزمایشگاهی] دارید. کدام یک از یافته‌های زیر به نفع [تشخیص ۱] است تا [تشخیص ۲]؟
یک [بیمار] با [علایم ونشانه‌ها] مراجعه کرده است. یافته‌ها حاکی از آن است که [کاهش/افزایش] در کدام یک از [آنزیم‌ها/هورمون‌ها] وجود دارد؟
برای یک [بیمار] که تحت [رژیم] است، کدام یک از شرایط زیر محتمل است؟
یک [بیمار] با [یافته فیزیکی] تحت درمان با [دارو] است. این دارو از طریق مهار چه [مکانیسمی] اثر می‌کند؟
یک [بیمار] با [یافته فیزیکی] مراجعه کرده است. کدام [یافته آزمایشگاهی] مورد انتظار است؟
یک [بیمار]، بعد از گذشت [زمان] از [مسافرت/خوردن غذا/مهمانی] مریض شده است. کدام یک از ارگانسیم‌ها احتمالاً مسوول بیماری اوست؟
متعاقب یک [پروسیجر]، بیمار دچار [یافته‌های فیزیکی وآزمایشگاهی] شده است. علت اصلی کدام مورد است؟
یک [بیمار] بر اثر [بیماری] می‌میرد. کدام یافته در اتوپسی مورد انتظار است؟
یک [بیمار] دارای [یافته‌های فیزیکی غیرطبیعی] است.اما در عین حال، [یافته‌های فیزیکی طبیعی] دارد. کدام مورد توضیح بهتری است؟

به طور خلاصه، استفاده از سناریوی بالینی کمک می‌کند سطوح شناختی بالا مورد ارزیابی قرار گیرند زیرا از دانشجو خواسته می‌شود تا معلومات و محفوظات خود را در بستر یک موقعیت نزدیک به واقعی به کار برد. بنابراین، در طراحی سؤالات چندگزینه‌ای توصیه می‌شود که از سناریوهای بالینی استفاده شود. با وجود اینکه سناریوی بالینی، بستر مناسبی برای افزایش سطح شناختی ایجاد می‌کند، تنها راه برای ارتقاء تاکسونومی سؤال نیست. یک راه دیگر، طرح سؤال بر اساس یک شکل، نمودار، جدول یا عکس است.

نمونه سؤال با استفاده از شکل

شماره ۴ در شکل چه چیزی را نشان می‌دهد؟



الف) کرست ایلیاک
ب) سر فمور
ج) تروکانتر بزرگ
د) ایسکیوم

راه دیگر برای افزایش سطح شناختی سؤال، فکر کردن به این موضوع است که این مطلب واقعاً چه کاربردی برای فراگیر دارد. به مثال زیر توجه کنید:

سؤال آمار با موضوع مشابه در دو سطح شناختی

در مورد مقایسه نسبت ها کدام صحیح است؟

الف) یک عدد: مک نمار ب) دو عدد مستقل: کای دو ج) دو عدد وابسته: کوکران د) سه عدد وابسته: اسپیرمن

شکل بهتر

در یک کارآزمایی بالینی، دو گروه ۰۰۱ نفره از بیماران انتخاب شدند و به هر یک از گروه‌ها یکی از دو داروی A و B تجویز شد. ۸۰ نفر با داروی A و ۰۶ نفر با داروی B بهبود یافتند. کدام آزمون آماری را برای مقایسه اثربخشی داروها مناسب می‌دانید؟

الف) مک نمار ب) کای دو
ج) کوکران د) اسپیرمن

نکته جالب توجه این است که از دید متخصصان موضوعی ممکن است سطوح مختلف تفاوت چندانی نداشته باشند. به عبارت دیگر این که سؤال به کاربرد و حل مسأله پرداخته است یا صرفاً محفوظات را می‌سنجد، بیشتر برای دانشجویان تازه‌کار و کسانی ملموس است که از نظر محتوایی صاحب‌نظر نیستند. اساساً به همین دلیل ارزیابی سطوح شناختی بالا در سؤالات چندگزینه‌ای چالش‌برانگیز است. زیرا برای طراحی چنین سؤالی، یک متخصص باید از دید یک دانشجوی تازه‌کار به مسائل نگاه کند. در بسیاری از مواقع، تخصص در محتوای مربوطه از طراحی سؤال تاکسونومی بالا جلوگیری می‌کند و همین امر کار را دشوار می‌سازد. بررسی و بازبینی تاکسونومی سؤالات توسط هیات علمی آموزش‌دیده از نظر سطوح شناختی اما غیرمتخصص از نظر محتوا می‌تواند منجر به بهبود سؤال شود (ترکتنبرگ و همکاران[1] ۲۰۱۳).

در تلاش برای افزایش سطح شناختی سؤالات، توجه به نکات زیر اهمیت زیادی دارد:

- تا حد امکان نباید در سؤال از جملات و مثال‌های متن کتاب و منبع امتحانی استفاده کرد. این کار باعث می‌شود سطح شناختی سؤال پایین بیاید. زیرا به جای اینکه قدرت تحلیل دانشجو را در شرایط جدید بسنجد، او را به حفظ کردن ترغیب می‌کند.
- استفاده از سناریو نباید این مسأله را در ذهن ایجاد کند که هر سؤال حاوی سناریوی بالینی، لزوماً به ارزیابی سطوح شناختی بالا می‌پردازد. تکرار این موضوع خالی از لطف نیست که که شکل و ساختار آزمون نیست که مشخص می‌کند چه نوع اهدافی قرار است اندازه‌گیری شوند. آنچه درون سؤال قرار می‌گیرد و پرسیده می‌شود، تعیین کننده سطح شناختی سؤال است. گاهی به نظر می‌رسد سؤال حاوی یک مورد بالینی است. اما در نهایت سؤالی که مطرح می‌شود، مستقل از سناریو قابل پاسخگویی است. در واقع، دانشجو برای جواب دادن به سؤال، نیازی به مطالعه سناریو و تفسیر یافته‌های آن ندارد. بنابراین، وجود سناریو لزوماً منجر به ارتقاء تاکسونومی سؤال نمی‌شود. به مثال زیر توجه کنید:

سوال حاوی سناریوی منفصل از پرسش اصلی

بیماری به دلیل تورم و قرمزی پوست بینی خارجی به پزشک مراجعه کرده و برای وی تشخیص سلولیت داده شده است. یکی از اعصاب حسی بینی خارجی، عصب Infraorbital می باشد که به دلیل این عارضه حساس شده است. این عصب شاخه کدام یک از اعصاب زیر است؟

الف) Olfactory nerve ب) Ophthalmic nerve ج) Pharyngeal plexus د) Maxillary nerve

- قبلاً گفته شد که در سطح شناختی بالا، اطلاعات به صورت خام و تفسیر نشده در اختیار دانشجو قرار می‌گیرد و خود دانشجو باید با کنار هم گذاشتن یافته‌های بالینی و تفسیر نتایج آزمایش‌ها، داده‌ها را تحلیل کند. اما تلاش برای ارتقا سطح شناختی سؤال از طریق سناریوهای بلند و ارائه اطلاعات به صورت خام و تفسیر نشده، نباید منجر به طولانی شدن بی‌دلیل پایه سؤال شود. پایه سؤال باید حاوی اطلاعات ضروری باشد و تا حد ممکن کوتاه نوشته شود. استفاده از سناریوی بالینی نباید این تصور را به وجود آورد که می‌توان اطلاعات زیادی که برای جواب لازم نیستند، در متن سؤال گنجاند. این موضوع که به ظاهرسازی[2] نیز معروف است، توصیه نمی‌شود. درست است که استفاده از سناریو کمک می‌کند تا به جای سنجش محفوظات، بر سطوح بالاتر و کاربرد دانش متمرکز شویم اما این موضوع نباید منجر به گزافه‌گویی[3] شود (کیس و سوانسون ۱۹۹۶، کولینز ۲۰۰۶). چنین وضعیتی منجر به سردرگمی و ابهام دانشجویان می‌شود و پایایی و روایی نمرات را کاهش می‌دهد (برتون و همکاران ۱۹۹۱).
- هنگامی که گفته می‌شود برای افزایش تاکسونومی سؤال از سناریوی بالینی استفاده شود، یکی از نگرانی‌های رایج دشوار شدن سطح سؤال است. باید توجه داشت که میزان دشواری سؤال و سطح شناختی مورد ارزیابی لزوماً

1. Tractenberg et al.
2. Window dressing
3. Verbosity

در یک راستا نیستند. به عنوان مثال، یک سؤال تستی می‌تواند سخت باشد اما فقط بر قدرت حفظ کردن فراگیران تاکید داشته باشد و به سطوح پایین محدود بماند. همچنین سؤال ممکن است دشوار نباشد اما بر اساس سناریو باشد و دانشجو برای پاسخ به آن باید تحلیل کند و در آن واحد اطلاعات خود را از چند جنبه به کار گیرد. در همین راستا، پژوهشگران تلاش کرده‌اند ویژگی‌های روان‌سنجی سؤالات با سناریو را با سؤالات بدون سناریو مقایسه کنند که به نتایج متناقضی دست پیدا کرده‌اند. به نتایج مقالات زیر توجه کنید:

کِیس و همکاران ۱۹۹۶

در یک آزمون گواهینامه، سه نوع سؤال با موضوع مشابه اما طول‌های مختلف ارائه شد: بدون سناریو، با سناریوی کوتاه و با سناریوی بلند. در نوع آخر اطلاعات به صورت تفسیر نشده در اختیار دانشجو قرار می‌گرفت. در حالی که در نوع اول، نیازی به تفسیر اطلاعات از طرف دانشجو نبود. دانشجویان از نظر نمره کلی آزمون به دو دسته قوی (۲۰ درصد بالا) و ضعیف (۲۰ درصد پایین) تقسیم شدند. ۹۹ درصد دانشجویان قوی و ۹۰ درصد دانشجویان ضعیف به سؤال بدون سناریو پاسخ صحیح دادند. دو نوع سؤال دیگر، یعنی با سناریوی کوتاه و با سناریوی بلند، هر چند که برای دانشجویان قوی سخت‌تر نبودند اما تعداد کمتری از دانشجویان ضعیف به آنها پاسخ درست دادند (به ترتیب ۸۲ درصد و ۶۶ درصد). سؤالی که در آن یافته‌های بیمار تفسیر نشده بودند و خود دانشجو این کار را انجام می‌داد، باید به صورت داده شد، تعداد کمتری به آن پاسخ دادند یعنی سؤال دشوارتر بود. اما تفاوت معناداری در ضریب تمیز سؤالات مشاهده نشد.

ترکتنبرگ و همکاران ۲۰۱۳

در این مطالعه به منظور پیدا کردن رابطه بین میزان سختی و تاکسونومی سؤال، ۲۵۲ سؤال چندگزینه‌ای در یک کورس فیزیولوژی توسط متخصصان (رشته غیر فیزیولوژی) از نظر سطح شناختی مورد بررسی قرار گرفتند. ضریب شواری سؤالات توسط آنالیز راش محاسبه شد. تحلیل‌ها نشان داد از نظر آماری، پیچیدگی شناختی ۸۸ درصد سؤالات از میزان دشواری آنها مستقل بود و رسیدن به سطوح شناختی بالا بدون تغییر دادن میزان دشواری سؤال امکان‌پذیر است.

نگارش سؤال هدایت‌کننده

سؤال هدایت‌کننده در بسیاری از موارد فراموش می‌شود یا به صورت غیرشفاف نوشته می‌شود. در حالی که این بخش نقش مهمی دارد و به روشنی برای دانشجو مشخص می‌کند که به چه سؤالی باید پاسخ دهد. به عنوان مثال، به نمونه‌ای از سؤالات هدایت‌کننده توجه کنید.

مثالی از سؤال چندگزینه‌ای با هدایت‌کننده ضعیف

خانمی ۱۸ ساله به علت آمنوره اولیه مراجعه کرده است. در معاینه متوجه می‌شوید که قد بیمار ۱۴۸ سانتی‌متر است. علاوه بر این متوجه می‌شوید که اندام تناسلی خارجی بیمار نابالغ بوده و تکامل پستان ایجاد نشده است. گزینه صحیح کدام است؟

الف) سندرم ترنر ب) دیس‌ژنزی گنادی مرکب ج) دیس‌ژنزی گنادی خالص د) سندرم نونان

شکل بهتر

خانمی ۱۸ ساله به علت آمنوره اولیه مراجعه کرده است. در معاینه متوجه می‌شوید که قد بیمار ۱۴۸ سانتی‌متر است. علاوه بر این متوجه می‌شوید که اندام تناسلی خارجی بیمار نابالغ بوده و تکامل پستان ایجاد نشده است. با توجه به شرح حال بیمار، محتمل‌ترین تشخیص کدام یک از گزینه‌های زیر است؟

الف) سندرم ترنر ب) دیس‌ژنزی گنادی مرکب ج) دیس‌ژنزی گنادی خالص د) سندرم نونان

در مثال اول مشخص نیست که پاسخ‌دهنده چه کاری را باید انجام دهد. این نوع سؤال هدایت‌کننده معمولاً منجر به ساختن تست‌هایی مبهم یا فاقد جهت‌گیری مشخص می‌شود. در مثال دوم، انتظار از شرکت‌کننده مشخص است و در نتیجه سؤال جهت‌داری نیز طراحی شده است. برای اطمینان از درستی سؤال هدایت‌کننده از آزمون پوشاندن گزینه‌ها

استفاده می‌شود. معمولاً اگر سؤال هدایت‌کننده به این صورت نوشته شود که کدام گزینه صحیح است، با خطاهای دیگری از جمله پرسش بیش از یک موضوع همراه می‌باشد. به مثال زیر توجه کنید. در حالت اول با پوشاندن گزینه‌ها نمی‌توان به سؤال پاسخ داد.

سؤال ضعیف
کدام یک از گزینه‌های زیر درمورد آلزایمر صحیح است؟ الف) اولین علامت بیماری به صورت تیپیک اختلال حافظه بلندمدت است. ب) تشنج ممکن است در جریان بیماری رخ دهد. ج) اختلال راه رفتن جزء علایم نیست. د) مهم‌ترین شکایت خود بیمار، در ابتدا اختلال حافظه است.
سؤال بهتر
چه یافته‌ای در ابتدای سیر بیماری آلزایمر محتمل‌تر است؟ الف) افسردگی ب) هذیان ج) فراموشی د) تشنج

گاهی دو گزینه ارائه شده، هر دو به درجاتی حاوی پاسخ صحیح هستند اما یکی از آنها درست‌تر است. در این صورت، در سؤال هدایت‌کننده حتماً باید قید شود که صحیح‌ترین پاسخ، مورد نظر است.

در سؤال هدایت‌کننده نیز از به کاربردن افعال منفی باید پرهیز شود. افعال منفی باعث پیچیده شدن بیش از حد سؤال می‌شوند. علاوه بر این، هنگام طراحی سؤال هدف این است که ببینیم آیا داوطلب جواب صحیح را می‌داند یا خیر، نه این که بداند ضعیف‌ترین گزینه کدام است.

طراحی گزینه‌ها

تعداد گزینه‌ها از موارد چالش‌برانگیز در حوزه طراحی سؤال چندگزینه‌ای است و تحقیقات متعددی هم در این خصوص صورت گرفته است که در فصل اول به آن پرداخته شد. در هر حال، آنچه اهمیت دارد این است که تا جایی که می‌توان گزینه‌های مناسب و موثری طراحی کرد.

گزینه صحیح باید به وضوح درست و بدون ابهام باشد و گزینه‌های انحرافی باید کاملاً غلط اما در عین حال جذاب و محتمل باشند. برای طراحی گزینه‌های انحرافی می‌توان از اشتباهات شایع داوطلبان استفاده کرد یا به این فکر کرد که یک دانشجوی ضعیف، در مواجهه با سؤال چگونه فکر می‌کند. هدف از طرح گزینه‌های انحرافی این است که فراگیرانی که دانش ناکافی دارند و مطمئن نیستند، از جواب صحیح منحرف شوند. معمولاً الگوی اشتباهات فراگیران در پاسخ به سؤال چندگزینه‌ای و انتخاب گزینه‌ها تکراری است. مدرسان با تجربه‌ای که اشتباهات رایج دانشجویان را می‌شناسند، پس از برخوردهای مکرر با آنان، قادر هستند آنچه را که باعث سردرگمی دانشجویان می‌شود، به راحتی تشخیص دهند. بنابراین این استادان قادر هستند گزینه‌های انحرافی خوبی طراحی کنند.

همه گزینه‌های یک سؤال باید به موضوع واحدی مربوط باشند. اگر هر یک از گزینه‌ها اشاره به مطلب جداگانه‌ای داشته باشد، در حقیقت هر گزینه خود به صورت یک سؤال «درست–نادرست» در می‌آید که از گزینه‌های دیگر مستقل است. نمونه بارز این نوع سؤالات با این بدنه یا شبیه به آن شروع می‌شوند که «کدام یک از عبارات زیر درست است؟» و در ادامه، گزینه‌هایی می‌آیند که هر یک به مطالب جداگانه‌ای اشاره می‌کنند و همان‌طور که قبلاً اشاره شد، با پوشاندن گزینه‌ها نمی‌توان به سؤال پاسخ داد.

سؤال ضعیف

کدام‌یک از جملات زیر صحیح است؟

الف) سر گرد حجم کمتری از سر بیضی شکل دارد، حتی اگر دور سر یک اندازه باشد.
ب) افزایش قطر قدامی خلفی سر تأثیر کمتری روی سر دارد تا افزایش قطر biparietal.
ج) دور سر نوزادان ارتباط بیشتری با Postnatal age دارد تا Conceptional age
د) دور سر نوزادان نارس در سن سه ماهگی به حد دور سر نوزادان ترم هم سن می رسد

سؤال بهتر

دور سر نوزادان نارس در حدود چند ماهگی به حد نوزادان ترم هم‌سن می رسد؟

الف) سه
ب) شش
ج) نه
د) دوازده

همان‌طور که در قسمت خطاهای طراحی سؤال ذکر شد، گزینه‌ای که طولانی‌تر است، معمولاً گزینه جواب است و به دانشجویان سرنخ می‌دهد. بنابراین، باید تلاش شود تا طول گزینه‌ها تقریباً یکسان باشد و از تکرار مطالب در گزینه‌ها پرهیز شود. اگر مطلب یا موضوعی در پایه سؤال ذکر شده است، از ذکر آن در گزینه‌ها خودداری گردد تا باعث صرفه‌جویی در وقت نویسنده و خواننده سؤال شود.

سؤال ضعیف

در EEG دختر ۲/۵ ساله‌ای که به علت تشنج ناشی از تب مراجعه کرده است، امواج اپی لپتیک مشاهده می‌شود. کدام جمله زیر صحیح نیست؟

الف) احتمالاً این کودک در آینده دچار اپی‌لپسی خواهد شد.
ب) احتمال دارد که این پدیده یک ویژگی موروثی باشد.
ج) احتمال دارد که این پدیده اسپورادیک بوده و ارتباطی به بروز تشنج بعدی نداشته باشد.
د) احتمالاً این الگو یک Maturational pattern است.

سؤال بهتر

با وجود امواج اپی لپتیک در الکتروانسفالوگرافی دختر بچه ۲/۵ ساله مبتلا به Febrile seizure کدام احتمال کمتر مطرح است؟

الف) بروز بعدی اپی‌لپسی
ب) وجود یک خصوصیت توارثی
ج) یک پدیده اسپورادیک
د) الگوی مربوط به سن

گزینه ها نباید با یکدیگر هم‌پوشانی داشته باشند و باید در عین حال که در ارتباط با پایه سؤال هستند، مستقل از هم طراحی شوند. به عنوان مثال، در سؤالی که با هدف کنترل درد طراحی شده است اگر یک گزینه «تجویز مسکن» باشد و در عین حال یکی از گزینه‌های دیگر «تجویز استامینوفن» باشد، بین محتوای این دو گزینه هم‌پوشانی وجود دارد در نتیجه مستقل از هم محسوب نمی‌شوند. به این ترتیب، دانشجو هر دو گزینه را کنار می‌گذارد و به این ترتیب از ارزش سؤال کم خواهد شد.

به همین ترتیب، از طرح دو گزینه که نافی یکدیگر هستند، نیز باید پرهیز شود. در سؤالاتی که تنها باید یک گزینه صحیح داشته باشند، گزینه‌هایی که ماهیتاً متضاد یکدیگر هستند، نمی‌توانند گزینه صحیح باشند و در نتیجه تعداد گزینه‌های مطرح برای انتخاب توسط داوطلب عملا کاهش می‌یابد. برای درک بهتر موضوع به مثال زیر توجه کنید:

سؤال ضعیف

تغییرات pH مایع مغزی نخاعی نسبت به خون در کدام یک از موارد زیر مقاوم‌تر است؟
الف) آلکالوز متابولیک
ب) هیپرناترمی
ج) اسیدوز متابولیک
د) هیپوکلسمی

سؤال بهتر

تغییرات pH مایع مغزی نخاعی نسبت به خون در کدام یک از موارد زیر مقاوم‌تر است؟
الف) آلکالوز متابولیک ب) آلکالوز تنفسی ج) اسیدوز متابولیک د) اسیدوز تنفسی

همان‌طور که در قسمت خطاها توضیح داده شد، حتی‌المقدور از گزینه‌های «هیچ کدام از موارد فوق» و «همه موارد فوق» استفاده نشود. گاهی طراحی سه گزینه انحرافی جذاب دشوار است و به همین دلیل از این دو مورد استفاده می‌شود.

از به کار بردن قیود مطلق «همیشه» و «هرگز» در گزینه‌ها اجتناب شود. دانشجو می‌داند که در حوزه علوم پزشکی این کلمات زیاد کاربرد ندارند و وجودشان در گزینه به معنای نادرست بودن گزینه است. همچنین از به کار بردن قیود مبهم نظیر «معمولاً»، «غالباً»، «عمدتاً» و نظیر آن پرهیز شود. تعریف این کلمات مشخص نیست و در نتیجه باعث سردرگمی داوطلبان در هنگام پاسخ‌گویی به سؤال می‌شود.

سرانجام در مرتب کردن گزینه‌ها دقت شود که بر اساس ترتیب عددی یا منطقی فهرست شوند، به جای افقی به شکل عمودی مرتب شوند، جای گزینه صحیح در سؤالات مختلف تغییر کند و از نظر محتوایی و دستوری همگن شوند.

در دفترچه سؤالات باید دقت شود که تمام گزینه‌های یک سؤال در یک صفحه نوشته شوند تا دانشجو برای پاسخ به سؤال مجبور نباشد به صورت مداوم برگه‌ها را ورق بزند.

مرور و ارزیابی سؤالات

پس از اینکه سؤال طراحی شد، مهم است که یک بار خود طراح سؤال آن را مرور کند. برای ارزیابی پایه هر سؤال و گزینه‌های آن می‌توان از چک لیست‌هایی که به همین منظور تهیه شده است، استفاده کرد. نمونه‌ای از این چک‌لیست در جدول ۳-۶ آمده است. باید توجه شود که هدف این چک‌لیست، نمره‌دهی و محاسبه امتیاز سؤال نیست بلکه کمک می‌کند تا نقاط ضعف سؤال آشکار شوند تا در صورت لزوم بتوان سؤال را اصلاح نمود.

از آنجا که معمولاً دیگران بهتر از خود طراح سؤال متوجه اشتباهات می‌شوند، با نظر خواستن از همکاران قبل از اینکه سؤال به دست دانشجو برسد، از اشتباهات احتمالی جلوگیری می‌شود. در بسیاری از موارد، سؤالات در جلساتی که به همین منظور تشکیل می‌شود، مورد ارزیابی قرار می‌گیرند. در این جلسات، کارشناسان آموزش‌دیده در کنار خود طراحان سؤال به مرور سؤالات می‌پردازند. در اینجا هم می‌توان با استفاده از چک‌لیست، تک‌تک سؤالات را از نظر محتوا، سطح شناختی، خطاهای ساختاری، رعایت قواعد دستوری، نقطه گذاری، نکات نگارشی و ... بررسی کرد و در مورد کیفیت آن به طراح سؤال بازخورد داد تا بازبینی سؤال صورت گیرد یا با سؤال بهتر جایگزین شود.

همچنین، پس از برگزاری آزمون، شاخص‌های دشواری و تمیز و پایایی و روایی آزمون را باید محاسبه کرد و به طراح در مورد کیفیت سؤالات بازخورد داد.

جدول ۳-۶: چک‌لیست ارزیابی سؤال «چندگزینه‌ای با بهترین پاسخ»

ارزیابی کلی		
۱	آیا سؤال، مرتبط با موضوعات مندرج در بلوپرینت یا اهداف دوره است؟	بله
۲	آیا سطح دشواری سؤال متناسب با سطح فراگیران است؟	بله
۳	آیا قواعد دستوری، املایی، نقطه گذاری و آیین نگارش رعایت شده است؟	بله
۴	اگر سؤال دارای نمودار، شکل یا عکس است، آیا الصاق شده‌اند؟	بله
۵	آیا تمام محتوای یک سؤال (محور، پایه و گزینه‌ها) در یک صفحه چیده شده است؟	بله
ارزیابی پایه سؤال		
۶	آیا پایه سؤال کامل و شفاف است؟	بله
۷	آیا محتوای سؤال غنی است؟ مثلاً یک مورد بالینی مبنای طرح سؤال بوده است؟	بله
۸	آیا سؤال به جای ارزیابی محفوظات به ارزیابی توان به کارگیری اطلاعات می‌پردازد؟	بله
۹	آیا سؤال منفی است یا از عباراتی نظیر «به جز» استفاده شده است؟	خیر
۱۰	آیا واحد داده‌های آزمایشگاهی قید شده است؟	بله
۱۱	آیا از کلمات تخصصی غیرمعمول یا مخفف استفاده شده است؟	خیر
ارزیابی سؤال هدایت‌کننده		
۱۲	آیا سؤال هدایت‌کننده به روشنی مشخص کرده است که چگونه باید به سؤال پاسخ داده شود؟	بله
۱۳	آیا می‌توان بدون نگاه کردن به گزینه‌ها به سؤال پاسخ داد؟	بله
ارزیابی گزینه‌های سؤال		
۱۴	آیا گزینه‌ها از نظر جنس همگون هستند؟ (مثلاً همه گزینه‌ها اقدام درمانی هستند)	بله
۱۵	آیا آیا گزینه‌ها از نظر قواعد دستوری، طول گزینه و زبان تکنیکی مشابه هستند؟	بله
۱۶	آیا گزینه‌ها به ترتیب الفبایی یا منطقی و به صورت عمودی ردیف شده‌اند؟	بله
۱۷	آیا در میان گزینه‌ها یک گزینه کاملاً صحیح وجود دارد و سایر گزینه‌ها برای داوطلب ضعیف جذاب هستند؟	بله
۱۸	آیا گزینه‌ها به صورت یک کلمه یا یک عبارت کوتاه بدون فعل نوشته شده‌اند؟	بله
۱۹	آیا عباراتی نظیر «همه موارد فوق» یا «هیچ کدام از موارد فوق» به کار رفته است؟	خیر
۲۰	آیا قیود مبهم (غالباً، معمولاً، گاهی) یا قیود مطلق (اصلاً، همیشه، هرگز) در گزینه‌ها استفاده شده است؟	خیر

تصحیح سؤالات «چندگزینه‌ای با بهترین پاسخ»

تصحیح پاسخنامه سؤالات چندگزینه‌ای به کمک نرم‌افزار به آسانی صورت می‌گیرد و این یکی از دلایل استفاده فراوان از سؤالات چندگزینه‌ای است. همان‌گونه که قبلاً اشاره شد، روش‌های نمره‌دهی متنوعی برای سؤالات چندگزینه‌ای وجود دارد. در اینجا به صورت خلاصه به آنها اشاره می‌شود. برای جزئیات بیشتر به فصل اول همین بخش مراجعه کنید.:

- نمره‌دهی مثبت: تصحیح سؤالات چندگزینه‌ای معمولاً این طور انجام می‌شود که دانشجو به ازای هر پاسخ صحیح،

یک نمره دریافت می‌کند و به ازای سؤالاتی که جواب نداده یا غلط جواب داده است، صفر می‌گیرد. قبلاً ذکر شد که به این روش، نمره‌دهی مثبت گفته می‌شود.

- نمره‌دهی منفی: در برخی موارد، آزمون‌گیرندگان برای کاهش میزان حدسی جواب دادن دانشجویان به سؤالات، پاسخ‌های غلط را جریمه می‌کنند و نمره منفی در نظر می‌گیرند. به طوری که اگر دانشجو به یک سؤال پاسخ اشتباه دهد، از او نمره کسر می‌شود. چون در این مدل نمره سؤالات بر اساس فرمول محاسبه می‌شوند، به آن نمره‌دهی با فرمول هم می‌گویند. این فرمول‌ها متنوع هستند به گونه‌ای که گاهی به ازای هر پاسخ غلط در سؤال چهارگزینه‌ای، تنها $\frac{۱}{۳}$ نمره سؤال کم می‌شود و گاهی به اندازه تمام نمره یک سؤال. اینکه چه میزان نمره باید از دانشجو کم شود، محل بحث بوده است. برخی می‌گویند اگر واقعاً می‌خواهیم نمره منفی، تاثیرگذار باشد و دانشجویان ضعیف و قوی را از هم متمایز کنیم، حتما باید سهم نمره منفی به اندازه کافی بزرگ باشد تا به صورت جریمه موثر عمل کند.
- مدل ارزیابی اطمینان: در این نحوه نمره‌دهی که توسط گاردنر–مدوین[1] پیشنهاد شده است، به میزان اطمینان دانشجو از پاسخی که می‌دهد، توجه می‌شود. به این صورت که از دانشجو خواسته می‌شود هم‌زمان با پاسخگویی به هر سؤال، اعلام کند که از جواب خود به چه میزان مطمئن بوده است. سپس برای پاسخ‌های غلطی که دانشجو از آنها مطمئن بوده است، نمره منفی و برای پاسخ‌های درستی که دانشجو از آنها مطمئن بوده است نمره مثبت در نظر گرفته می‌شود. یک نمونه از این نحوه نمره‌دهی در فصل اول نشان داده شده است. این روش مختص سؤال چندگزینه‌ای نیست و برای سؤال «درست–نادرست» هم به کار رفته است.

سودمندی سؤال «چندگزینه‌ای با بهترین پاسخ»

بر اساس فرمول ون‌درولوتن، پنج معیار برای ارزیابی سودمندی یک ابزار ارزیابی مدنظر قرار می‌گیرد: روایی، پایایی، تاثیر آموزشی، هزینه و مقبولیت. در نظر گرفتن تمام این موارد با یکدیگر و برقراری تعادل اهمیت دارد. در هر امتحانی با توجه به اهدفی که دنبال می‌کند و میزان اهمیتی که دارد، ممکن است یکی از این معیارها بیشتر توجه شود.

روایی سؤال «چندگزینه‌ای با بهترین پاسخ»

برای اطمینان از روایی سؤالات چندگزینه‌ای باید مشخص شود که آیا مطابق اهداف دوره بوده‌اند؟ آیا از موضوعات مهم و اصلی[2] سؤال طرح شده است؟ آیا تمام مباحث مد نظر قرار گرفته‌اند و آیا سؤالات نمونه خوبی از تمام موضوعات هستند؟ بدیهی است که سؤالات چندگزینه‌ای برای سنجش اهداف مهارتی، عملی، بالینی و نگرشی جزء روا نیستند و اگر قرار است ارزیابی دوره‌ای که شامل طیفی از این اهداف است، به صورت روا انجام شود، باید از چندین نوع ابزار برای پوشش تمام اهداف دوره استفاده شود.

همچنین، همان‌طور که قبلاً مفصلاً توضیح داده شد، بررسی و اصلاح خطاهای ساختاری سؤال بر اساس چک‌لیست‌های موجود موجب بهبود روایی آزمون می‌شود. باید در نظر داشت که شرایط فیزیکی آزمون (نور، دما، صداهای اضافه و ..)، تقلب و خستگی دانشجویان از جمله مسائل دیگری هستند که روایی آزمون را تحت تاثیر قرار می‌دهند.

پایایی سؤال «چندگزینه‌ای با بهترین پاسخ»

پایایی آزمون چندگزینه‌ای در مقایسه با سایر ابزارهای ارزیابی نسبتاً بالا است. در حالی که پایایی یک آزمون

1. Gardner-Medwin
2. Core

چندگزینه‌ای ۴ ساعته ۰/۹۰ برآورده شده است، پایایی آزمون تشریحی محدود پاسخ و شفاهی در زمان مشابه به ترتیب ۰/۸۵ و ۰/۴۵ به دست آمده است (وس و همکاران ۲۰۰۱). یکی از دلایل این امر، تعداد سؤال در واحد زمان است. از آنجا که در یک زمان مشخص، تعداد بیشتری سؤال چندگزینه‌ای قابل طرح است تا سؤال تشریحی یا شفاهی، پایایی آزمون بالاتر است.

برای بررسی رابطه پایایی آزمون چندگزینه‌ای با تعداد سؤالات، مطالعه‌ای انجام شد که در آن از آزمونی با ۲۰۰ سؤال چندگزینه‌ای و پایایی ۰/۹۲ استفاده گردید. پژوهشگران به صورت تصادفی ۲۰ سؤال را حذف کردند و بار دیگر پایایی آزمون را محاسبه نمودند. آنها این کار را تا رساندن تعداد سؤالات به ۱۰۰ عدد ادامه دادند. ضریب پایایی کاهش پیدا کرد تا به ۰/۸۶ رسید. کاهش قابل توجه پایایی برای این آزمون هنگامی بود که کمتر از ۱۴۰ سؤال باقی ماند (ریموند و همکاران[1] ۲۰۰۳). البته این اعداد مربوط به آزمون مورد بررسی در همان پژوهش هستند و به راحتی قابل تعمیم به آزمون‌هایی که با موضوعات دیگری در سایر شرایط برگزار می‌شود.

سؤالات چندگزینه‌ای و به صورت کلی سؤالات بسته‌پاسخ، جواب‌های از قبل مشخصی دارند و دانشجو فقط باید از میان آنها انتخاب کند. بنابراین برخلاف آزمون‌های بازپاسخ، نمرات دانشجویان تحت تاثیر ذهنیت مصححان و تفاسیر و برداشت‌های مختلف آنان قرار نمی‌گیرد.

یکی از نگرانی‌های که در مورد پایایی سؤال چندگزینه‌ای وجود دارد، تاثیر حدس زدن تصادفی است که پایایی را کاهش می‌دهد. اما همان‌طور که قبلاً گفته شد، طبق نتایج شواهد، راه‌کارهایی که برای جلوگیری از پاسخ حدسی دانشجویان در نظر گرفته می‌شود، لزوماً منجر به بهبود چشم‌گیر پایایی نمی‌شوند. در هر حال، از بین سؤالات کتبی بسته‌پاسخ، میزان حدس زدن در سؤال چندگزینه‌ای نسبت به سؤال «درست–نادرست» کمتر و نسبت به سؤال «جورکردنی گسترده پاسخ» بیشتر است.

تاثیر آموزشی سؤال «چندگزینه‌ای با بهترین پاسخ»

اگر سؤال چندگزینه‌ای به گونه‌ای طراحی شود که امکان سنجش توانایی استدلال و حل مسأله دانشجویان را داشته باشد، می‌توان گفت یادگیری دانشجویان را به سمت درک و فهم و یادگیری عمیق، جهت‌دهی می‌کند در غیر این صورت بسیار محتمل است که آنان را صرفاً به حفظ کردن مطالب سوق دهد.

همچنین، ارائه کلید و پاسخنامه سؤالات بعد از آزمون به دانشجویان کمک می‌کند روی آموخته‌های خود تأمل کنند و اشتباهات خود را اصلاح کنند.

شایان ذکر است که اگر آزمون چندگزینه‌ای تنها روش ارزیابی دانشجویان باشد، دور از انتظار نیست که فقط مهارت‌های ذهنی و دانشی دانشجویان تقویت شود و انگیزه‌ای برای یادگیری مهارت‌های عملی و مسائل نگرشی نداشته باشند.

مقبولیت، هزینه و قابلیت اجرای سؤال «چندگزینه‌ای با بهترین پاسخ»

آزمون چندگزینه‌ای در مقاطع و رشته‌های گوناگون به کار می‌رود که نشان دهنده مقبولیت بالای آن برای مدرسان و فراگیران می‌باشد. هر چند که به صورت کلی هزینه این نوع امتحان، مخصوصاً در قیاس با سایر آزمون‌ها پایین است، اما اگر قرار باشد سؤال چندگزینه‌ای با کیفیت مطلوب طراحی گردد، طراحان سؤال باید وقت قابل توجه‌ای صرف کنند. به عنوان مثال، گفته می‌شود طرح یک سؤال چندگزینه‌ای برای طراح سؤال با تجربه، حدود ۱۵ دقیقه زمان می‌برد اما برای کسی که تجربه چندانی ندارد، سه الی چهار ساعت هم ممکن است طول بکشد (شوورث و ون‌درولوتن ۲۰۰۴). همچنین

1. Raymond et al.

برگزاری کارگاه‌های آموزشی و زمانی که هیأت علمی صرف مرور سؤالات می‌کنند، باید در نظر گرفته شود. بدیهی است که اگر آزمون فقط مربوط به یک رشته باشد، اجرای آن آسان‌تر است تا زمانی که محتوای آزمون به گونه‌ای است که استادان از چندین دیسیپلین و گروه مختلف باید به طراحی سؤال بپردازند.

برای تصحیح پاسخنامه‌ها پس از برگزاری آزمون، امروزه نرم‌افزارهای کامپیوتری فراوانی در دسترس هستند این نرم‌افزارها همچنین اجرای مراحل مختلف آزمون از تدوین دفترچه گرفته تا تصحیح برگه‌ها و تحلیل آزمون را تسهیل می‌کنند و به علاوه، امکان تهیه بانک سؤال را فراهم می‌سازند.

آزمون چندگزینه‌ای به ویژه برای ارزیابی تعداد زیاد شرکت‌کنندگان کاربرد زیادی دارد. هرچند که اصول کلی برگزاری آزمون در تعداد متفاوت دانشجویان، فرقی نمی‌کند اما مخصوصاً در آزمون‌های کشوری و مهم مانند آزمون پذیرش دستیار، رعایت مسائل مربوط به حفاظت سؤالات، امنیت آزمون، در نظر گرفتن فضاها و سالن‌های مناسب، اعلام به موقع نتایج و تضمین کیفیت آزمون باید شدیداً مورد توجه قرار گیرد.

منابع

1. Amin Z, Khoo HE, Chong YS. Practical Guide to Medical Student Assessment. World Scientific Pub Co Inc. 2006
2. Bandiera G, Sherbino J, Frank JR. The CanMEDS assessment tools handbook. An introductory guide to assessment methods for the CanMEDS competencies. Ottawa: The Royal College of Physicians and Surgeons of Canada; 2006
3. Boland RJ, Lester NA, Williams E. Writing Multiple-Choice Questions. Academic Psychiatry 2010; 34(4): 310-16
4. Burton SJ, Sudweeks RR, Merrill PF, Wood B. How to Prepare Better Multiple Choice Test Items: Guidelines for University Faculty, 1991.
5. Campbell DE. How to write good multiple-choice questions. J Paediatr Child Health 2011; 47(6):322-5.
6. Case SM, Swanson DB. Constructing Written Test Questions for the Basic and Clinical Sciences, Philadelphia, National Board of Medical Examiners; 2002
7. Cheung D, Bucat R. How can we construct good multiple-choice items? Presented at the Science and Technology Education Conference, Hong Kong 2002.
8. Collins J. Writing Multiple-Choice Questions for Continuing Medical Education Activities and Self-Assessment Modules. RadioGraphics 2006; 26:543-551.
9. Darling-Hammond L. School reform at the crossroads: Confronting the central issues of teaching. Educational Policy 1997;11(2):151–166
10. Downing SM. The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. Adv Health Sci Educ Theory Pract 2005;10(2):133-43.
11. Haladyna TM, Downing SM. Validity of a taxonomy of multiple-choice item-writing rules. Applied Measurement in Education 1989; 2(1): 51-78.
12. Haladyna TM. Developing and validating multiple-choice test items, 2nd edition. Lawrence Erlbaum Associates, 1999.
13. Haladyna TM, Downing SM, Rodriguez MC. A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. Applied Measurement in Education 2002; 15(3):309-334
14. Hill HC. Learning in the teaching workforce. The Future of Children 2007; 17(1): 111–128.
15. Jozefowicz RF, Koeppen BM, Case S, Galbraith R, Swanson D, Glew RH. The quality of in-house medical school examinations. Acad Med 2002 Feb;77(2):156-61.
16. Morrison S, Free K. Writing multiple-choice test items that promote and measure critical thinking. Journal of Nursing Education 2001; 40: 17-24.

17. Moss E. Multiple choice questions: their value as an assessment tool. Curr Opin Anaesthesiol 2001;14(6):661-6.
18. Naeem N, van der Vleuten C, Alfaris EA. Faculty development on item writing substantially improves item quality.Adv Health Sci Educ Theory Pract 2012;17(3):369-76.
19. Raymond M, Neiers B, Reid JB. Test-item development for radiologic technology. The American Registry of Radiologic Technologists 2003.
20. Schuwirth LW, van der Vleuten CP, Donkers HH. A closer look at cueing effects in multiple-choice questions. Med Educ 1996;30(1):44-9.
21. Schuwirth LWT, van der Vleuten CPM. ABC of learning and teaching in medicine: Written assessment. BMJ 2003; 326(7390):643-645.
22. Schuwirth LW, van der Vleuten CP. Different written assessment methods: what can be said about their strengths and weaknesses? Med Educ 2004; 38(9): 974-9.
23. Tarrant M, Knierim A, Hayes SK, Ware J. The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. Nurse Educ Pract 2006;6(6):354-63.
24. Wallach PM, Crespo LM, Holtzman KZ, Galbraith RM, Swanson DB. Use of a committee review process to improve the quality of course examinations. Adv Health Sci Educ Theory Pract 2006; 11(1):61-8.

۲۵. طاهری م، خوشرنگ ح، اسدی لویه ع، حیدرزاده ا. مقایسه کیفیت سؤالات آزمون ارتقای دستیاری قبل و بعد از مداخله آموزشی در دانشکده پزشکی گیلان طی سال‌های ۱۳۸۹ و ۱۳۹۰. جله ایرانی آموزش در علوم پزشکی ۱۳۹۲؛۵۷(۷):۵۵۱-۵۶۰
۲۶. سیف ع. اندازه گیری، سنجش و ارزشیابی آموزشی. نشر دوران، ۱۳۹۰
۲۷. گرک‌یراقی م، آویژگان م، ابراهیمی ا و همکاران. بررسی شاخص‌های کمی و کیفی آزمون‌های مقطع کارآموزی. مجله ایرانی آموزش در علوم پزشکی ۱۳۸۹؛ ۱۰(۵): ۵۳۳-۵۴۲.
۲۸. معیاری ا، بیگلرخانی م، زندی م، واحدی م، میر اسماعیلی ا. تاثیر مداخله آموزشی بر بهینه کردن طراحی سؤالات چندگزینه‌ای در آزمون های ارتقای دستیاری دانشکده دندانپزشکی. مجله ایرانی آموزش در علوم پزشکی ۱۳۹۱؛ ۱۲(۱): ۳۶ -۴۵
۲۹. شاهی ف، میرزازاده ع. راهنمای طراحی تست‌های چند جوابی. دفتر توسعه آموزش دانشکده پزشکی دانشگاه علوم پزشکی تهران ۱۳۸۴
۳۰. صبوری کاشانی ا، محمدی م، زینالو ع، رضوی م. استانداردسازی ارزشیابی دانشجو: معرفی دو نظریه کلاسیک و نوین. مرکز مطالعات و توسعه آموزش دانشگاه علوم پزشکی تهران. انتشارات تهران صدا. ۱۳۸۳
۳۱. میرزازاده ع. راهنمای نظارت بر سؤالات آزمون ارتقای دستیاران تخصصی. دفتر توسعه آموزش دانشکده پزشکی دانشگاه علوم پزشکی تهران ۱۳۸۵
۳۲. وفامهر و، دادگسترنیا م. بررسی نتایج آنالیز کمی و کیفی سؤالات چهارگزینه‌ای دوره مقدمات پزشکی بالینی. مجله ایرانی آموزش در علوم پزشکی ۱۳۸۹؛ ۱۰(۵):۱۱۴۶-۱۱۵۲

فصل ۷

سؤال «جور کردنی گسترده»

ساختار سؤال «جورکردنی گسترده»

با وجود اینکه سؤال «چندگزینه‌ای با بهترین پاسخ» مزایای بسیاری دارد و مخصوصاً اگر درست طراحی شود، قادر به ارزیابی سطوح بالای شناختی مانند فهم و درک و کاربرد اطلاعات است، همچنان نگرانی‌هایی در مورد احتمال بالای حدس زدن پاسخ توسط دانشجویان وجود دارد. از طرف دیگر اگر برای حل این مشکل از سؤالات تشریحی و کوته‌پاسخ استفاده شود، به دلیل مسائل و مشکلات مربوط به تصحیح، محدودیت‌هایی برای برگزارکنندکان ایجاد می‌شود.

برای حل این مشکلات، برخی از دانشگاه‌ها در چند سال اخیر سؤالات «جورکردنی گسترده» را جایگزین سؤال کوته‌پاسخ و چندگزینه‌ای کرده‌اند؛ زیرا پایایی و روایی آنها خوب است و همچنین تصحیح آنها آسان است و نظرات مصحح روی نمرات تاثیر نمی‌گذارد. در عین حال، در این نوع سؤالات، احتمال پاسخگویی حدسی و بازشناسایی پاسخ کمتر است (شوورث و ون‌درولوتن ۲۰۰۳، برتون ۲۰۰۹). تغییر رویه برخی از دانشگاه‌ها برای استفاده از سؤالات جورکردنی گسترده به عنوان جایگزین یا مکمل سؤالات دیگر، هم در مقطع پزشکی عمومی و هم تخصصی، مشهود بوده است. در این راستا، دستورالعمل‌هایی برای طراحی این سؤال توسط مراکز مختلف از جمله شورای پزشکی عمومی[1] انگلیس، بورد آموزش پزشکی تخصصی[2] و تعدادی از کالج‌های سلطنتی[3] تدوین شده است (برتون ۲۰۰۹) و توانمندسازی اساتید و فراهم کردن زمینه‌های سخت‌افزاری و نرم‌افزاری مورد توجه قرار گرفته است.

سؤال «جورکردنی گسترده» که در برخی از متون سؤال چندگزینه‌ای نوع R نامیده شده است، از نظر شکلی بسیار شبیه سؤال «چندگزینه‌ای با بهترین پاسخ» است اما با در اختیار گذاشتن لیستی از گزینه‌ها به جای تنها چهار گزینه، این امکان را به وجود می آورند که احتمال حدس زدن دانشجو کاهش پیدا کند. از طرف دیگر، با طراحی یک پایه خوب و مناسب که معمولاً به صورت یک مورد بالینی است، هدف سؤال از سطح ارزیابی یادآوری بالاتر می‌رود و قادر است میزان درک و قدرت تحلیل دانشجو را مورد سنجش قرار دهد. سؤالات «جورکردنی گسترده» از چهار قسمت تشکیل شده‌اند: محور، گزینه‌ها، سؤال هدایت‌کننده، و حداقل دو پایه سؤال.

1. General Medical Council (GMC)
2. Postgraduate Medical Education and Training Board (PMETB)
3. Royal College

نمونه ای از سؤال «جورکردنی گسترده»

محور: خستگی
الف) لوسمی حاد ب) آنمی بیماری مزمن ج) بیماری احتقانی قلب د) توبرکلوز ه) عفونت ویروس اپشتن بار
و) هایپوتیرو‌ییدی ز) فقر آهن ح) اسفروسیتوز ارثی ط) کمبود ویتامین B_{12}

برای هر یک از بیماران زیر کدام یک از تشخیص‌های فوق محتمل‌تر هستند؟
۱. خانم ۱۹ ساله با خستگی، تب و گلودرد از هفته قبل به شما مراجعه کرده است. در معاینه لنفانوپاتی گردنی و اسپلنومگالی دارد و درجه حرارت بدنش ۸۳/۳ درجه سانتی‌گراد است. جواب آزمایش اولیه او به صورت زیر است:

Leukocyte count: 5000/mm3 (80% lymphocytes, with many lymphocytes with atypical features).

AST: 200 U/L. Bilirubin: normal ALT: normal

۲. دختر ۱۵ ساله با درد شکم و خستگی که دو هفته قبل شروع شده است، مراجعه کرده است. در معاینه متوجه رنگ پریدگی، کبودی و تندرنس در ناحیه ستون مهره‌ها و هر دو فمور می‌شوید. جواب آزمایش خونی او به صورت زیر است:

Hemoglobin concentration: 7.0 g/dL Leukocyte count of 2000/mm3 Platelet count of 15,000/mm3

خطاهای طراحی سؤال «جورکردنی گسترده»

در مورد خطاهای طراحی سؤال قبلاً مفصلاً توضیح داده شد. در اینجا بر خطاهایی تمرکز می‌کنیم که هنگام نوشتن سؤال «جورکردنی گسترده» بیشتر رخ می‌دهند.

نبود سؤال هدایت‌کننده مشخص و شفاف، یکی از خطاهای طراحی سؤال «جورکردنی گسترده» است. به عنوان مثال، نمونه زیر یک سؤال هدایت‌کننده نادرست است و باید از آن اجتناب شود زیرا باعث سردرگمی دانشجو می‌شود و دقیقاً مشخص نمی‌کند که چه کاری باید انجام دهد.

سوال «جورکردنی گسترده» با هدایت‌کننده ضعیف

هر سؤال را با گزینه مربوطه جور کنید.

شکل بهتر

برای هر یک از بیماران زیر، مناسب‌ترین اقدام درمانی را از بین گزینه‌های فوق انتخاب کنید.

ناهمگونی در فهرست گزینه‌ها خطایی است که منجر به ایجاد گزینه غیرعملکردی می‌شود و باید از آن اجتناب شود. منظور از ناهمگونی گزینه‌ها این است که تمام گزینه‌ها به یک محور واحد اشاره نداشته باشند. مثلاً در سؤال زیر، برخی از گزینه‌ها نام ماده معدنی هستند. در حالی که محور سؤال، کمبود ویتامین است. دانشجو می‌تواند پنج گزینه را بدون زحمت زیاد حذف کند. بنابراین، شانس حدسی جواب دادن او زیاد می‌شود.

ناهمگونی در گزینه‌های سؤال «جورکردنی گسترده»

محور: ویتامین
الف) آهن ب) روی ج) فولات د) منیزیم ه) مس
و) ویتامین A ز) ویتامین B_1 ح) ویتامین B_6 ط) ویتامین C ی) ویتامین K

برای هر یک از موارد زیر، مشخص کنید که کدام ویتامین نقش دارد؟
۱. خانم ۷۰ ساله با اکیموز، پتشی پریفولیکولر و لثه‌های ملتهب مراجعه کرده که غذای معمول او نوشابه و ساندویچ بوده است.
۲. در فرایند تشکیل لخته نقش دارد.

به همین ترتیب، سرنخ‌های گرامری یا منطقی که ناشی از ناهماهنگی بین گزینه‌ها و پایه‌ها است، خطای دیگری است که نباید رخ دهد.

خطای دیگر هنگام نگارش گزینه‌ها، نوشتن آنها به صورت جملات فعل‌دار است. گزینه‌ها باید کلمات و عبارات کوتاه باشند. در مثال زیر، علاوه بر اینکه سؤال محور مشخصی ندارد، گزینه‌ها ناهمگون هستند. دانشجو متوجه نمی‌شود آنها از چه جهت به هم ارتباط دارند و مجبور است چند بار از ابتدا تا انتهای سؤال را مرور کند. از طرفی، چون سؤال هدایت کننده مشخصی وجود ندارد، درک کاری که باید انجام شود، برای دانشجو سخت و گیج‌کننده است. همچنین، پایه سؤال به تنهایی و بدون گزینه‌ها معنادار نیست و بیش از حد کوتاه است. در واقع آزمون پوشاندن گزینه‌ها منفی است.

مثالی از سؤال «جورکردنی گسترده» نامطلوب

الف) اثری ندارد. ب) ایجاد بیماری‌های حرکتی است. ج) می تواند روی بینایی اثر بگذارد.
د) کاملاً قابل کنترل است ه) به صورت غیر مستقیم افزایش می‌یابد. و) ریسک سرطان را زیاد می‌کند.
ز) افزایش بیماری‌های تنفسی است. ح) مرگ و میر را کاهش داده است.
۱. سموم تولید شده کارخانجات صنعتی ...
۲. خطرناک‌ترین اثر آلودگی هوا ....

یکی از خطاها که به محتویات سؤال برمی‌گردد، این است که پایه سؤال غنی نباشد و طوری نوشته شود که به جای سطوح بالای شناختی، حافظه دانشجو را ارزیابی کند. در مثال زیر، پایه سوم از این نوع است:

مثالی از پایه نامطلوب در «سؤال جورکردنی گسترده»

محور: تب
الف) آدنوویروس ب) اپشتن بار ویروس ج) استرپتوک پنومونیه د) استرپتوکوک پیوژنز ه) باسیلوس آنتراسیز
و) کاندیدیا آلبیکانس ز) مایکوباکتریومتوبرکلوزیس ح) مایکوپلاسما پنومونیه ط) نایسریا گنوره ی) نایسریا مننژایتیس
برای هر یک از بیماران زیر که با شکایت تب مراجعه کرده‌اند، کدام یک از میکروارگانیسم‌های فوق می‌توانند علت بیماری باشد؟
۱. دختر ۷ ساله با تب شدید و گلودرد مراجعه کرده است. در معاینه، ته حلق او قرمز است و لوزه ها ملتهب و پوشیده از اگزودا هستند. همچنین لنفادنوپاتی ساب‌مندیبولر دردناک در سمت راست دارد. کشت نمونه حلق، کلونی‌های همولیتیک متعدد ایجاد کرده که نسبت به باکتریسین حساس هستند.
۲. پسر ۸۱ ساله‌ای از یک هفته پیش تب، گلودرد و بی‌حالی داشته است. در معاینه، لوزه‌های بزرگ و اگزودا، لنفادنوپاتی گردنی و اسپلنومگالی مشهود است. در آزمایش خونی، لنفوسیتوز با ارجحیت لنفوسیت های آتیپیکال دیده می‌شود و تست آنتی‌بادی هتروفیل او مثبت است.
۳. یک میکروارگانیسم گرم مثبت کپسول‌دار که معمولاً به صورت جفتی یا زنجیره کوتاه رشد می‌کند.

اساساً سؤال «جورکردنی گسترده» مطلوب بهتر است به ارزیابی قدرت تحلیل و استدلال دانشجو بپردازد، نه اینکه دانش و اطلاعات را به صورت مجزا بسنجد و شبیه جدولی شود که از دانشجو بخواهد موارد متناظر دو ستون را به یکدیگر متصل کند. در سؤال زیر، پایه اول در یک بستر مرتبط پزشکی پرسشی را مطرح می‌کند در حالی که پایه دوم، این طور نیست.

مثالی از پایه نامطلوب در «سؤال جورکردنی گسترده»

محور: ویتامین
الف) ویتامین A ب) ویتامین B_1 ج) ویتامین B_6 د) ویتامین B_{12}
ه) ویتامین C و) ویتامین D ز) ویتامین E ح) ویتامین K
برای هر یک از موارد زیر، مشخص کنید که کدام ویتامین نقش دارد؟
۱. خانم ۷۰ ساله با اکیموز، پتشی پریفولیکولر و لثه‌های ملتهب مراجعه کرده که غذای معمول او نوشابه و ساندویچ بوده است.
۲. در فرایند تشکیل لخته نقش دارد.

گام‌های طراحی سؤال «جورکردنی گسترده»

برای ساخت یک سؤال «جورکردنی گسترده» مطلوب هر چهار قسمت (محور، گزینه‌ها، سؤال هدایت‌کننده و پایه سؤال) ضروری هستند. اگرچه شکل این سؤال بسیار شبیه سؤال چندگزینه‌ای است، ترتیب گام‌های طراحی آن کمی متفاوت است. برای طراحی سؤالات «جورکردنی گسترده» معمولاً ابتدا موضوع و محور مشخص می‌شوند، سپس سؤال هدایت‌کننده و گزینه‌ها نوشته می‌شوند و در ادامه از بین فهرست گزینه‌ها چند گزینه انتخاب می‌شوند تا پایه سؤال برای آنها طراحی شود (جدول ۱-۷).

جدول ۱-۷ : خلاصه مراحل طراحی سؤالات «جورکردنی گسترده»

ردیف	عنوان مرحله	توضیح
۱	انتخاب یک موضوع مناسب و مهم	موضوع قسمتی از اطلاعات و دانش داوطلب هستند که قرار است سؤال آن را بسنجد. این موضوعات در ارتباط با اهداف دوره هستند که قرار است مورد ارزیابی قرار گیرد.
۲	انتخاب محور مناسب برای سؤال	محور موقعیتی است که تعیین می‌کند چه اطلاعاتی (تفسیر داده‌ها، تشخیص، مراقبت از بیمار) باید در سؤال ارایه شود.
۳	نگارش سؤال هدایت‌کننده	سؤال هدایت‌کننده، رابط بین گزینه‌ها و پایه سؤال است و مشخص می‌کند که دانشجو چه کاری باید انجام دهد.
۴	طراحی فهرست گزینه‌ها	گزینه‌ها باید به صورت یک کلمه یا عبارت کوتاه و همگون باشند و بر اساس ترتیب الفبایی ردیف شوند.
۵	نگارش پایه‌های سؤال	بیش از یک پایه لازم است. پایه‌ها بهتر است بر مبنای موارد بالینی و از لحاظ ساختار یکسان باشند. علاوه بر این باید حاوی تمام اطلاعاتی باشد که برای پاسخ به سؤال ضروری است.
۶	مرور و ارزیابی سؤالات	سؤال توسط خود طراح، همکاران و کمیته آزمون توسط چک‌لیست ارزیابی می‌شود.

انتخاب یک موضوع مناسب و مهم برای سؤال

در فصل مربوط به سؤال چندگزینه‌ای در این مورد مفصلاً صحبت شد. از موضوعی باید سؤال طرح شود که با اهداف و بلوپرینت دوره مرتبط باشد و یادگیری آن برای دانشجو اهمیت و کاربرد داشته باشد.

انتخاب محور سؤال

محور سؤال همان چیزی است که سؤال حول آن پرسیده می‌شود. ممکن است شکایت بیمار، تظاهرات بیماری، نام دارو و ... باشد. این قسمت نیز تفاوت چندانی با انتخاب محور در سؤال چندگزینه‌ای ندارد. تنها باید توجه شود که اهمیت مشخص کردن محور در این است که طراح را وادار می‌کند گزینه‌ها را به صورت همگون طراحی کند. در صورت نبود محور مشخص، شانس پراکندگی و عدم تجانس در فهرست گزینه‌ها زیاد می‌شود.

نگارش سؤال هدایت‌کننده

چارچوبی که غالباً برای طراحی سؤال هدایت‌کننده استفاده می‌شود، می‌تواند به این صورت باشد: «برای هر یک از بیماران زیر، مناسب‌ترین [تشخیص/درمان] را انتخاب کنید» اما بهتر است سؤال هدایت‌کننده به صورت اختصاصی‌تر مطرح شود و محور سؤال نیز در آن بیاید. چارچوب‌هایی که می‌توان برای نگارش قسمت هدایت‌کننده سؤال «جورکردنی گسترده» از آنها استفاده کرد، به صورت زیر است:

- برای هر یک از بیماران زیر با [شکایت بیمار/بیماری]، مناسب‌ترین [تشخیص/درمان/علت] را انتخاب کنید
- برای هر یک از بیماران زیر با [شکایت بیمار/بیماری]، [ساختار]ی را که احتمالاً [آسیب دیده/کمبود دارد/ناقص است]، انتخاب کنید.
- برای هر یک از بیماران زیر با [شکایت بیمار/بیماری]، [یافته] مورد انتظار را انتخاب کنید.
- برای هر یک از بیماران زیر با [شکایت بیمار/بیماری]، اقدام مناسب بعدی را انتخاب کنید.

به دو نمونه سؤال زیر توجه کنید:

مثالی از دو هدایت‌کننده برای سؤال «جورکردنی گسترده»

برای هر یک از بیماران زیر با شکایت خستگی، مناسب‌ترین تشخیص را انتخاب کنید.
برای هر یک از بیماران زیر که مبتلا به کمبود آنزیم هستند، محتمل‌ترین نقص پروتیئنی را انتخاب کنید.

یکی از اشتباهات رایج این است که سؤال هدایت‌کننده نوشته نشود یا به شکل مخدوش نوشته شود. مثلاً «هر سؤال را با گزینه مربوطه جور کنید»، یک سؤال هدایت‌کننده نادرست است و باید از آن اجتناب شود زیرا باعث سردرگمی دانشجو می‌شود و دقیقاً مشخص نمی‌کند که چه کاری باید انجام دهد.

همانند سؤال چندگزینه‌ای، اینجا هم برای اطمینان از درستی سؤال هدایت‌کننده، می‌توان از آزمون پوشاندن گزینه‌ها استفاده کرد. سؤال هدایت‌کننده بدون دیدن گزینه‌ها باید قابل پاسخگویی باشد.

نگارش فهرست گزینه‌ها

تعداد گزینه‌ها بسته به موضوع و سؤال ممکن است متفاوت باشد. از سه گزینه تا بیش از بیست گزینه را می‌توان در اختیار دانشجو قرار داد. برخی متون توصیه کرده‌اند که حداقل هشت گزینه به ازای هر پنج سناریو ارائه شود تا احتمال پاسخ تصادفی به حداقل برسد. شواهد نشان می‌دهند که این نوع سؤال زمانی بهترین کارکرد را دارد که برای هر سناریو یک گزینه که بهترین پاسخ صحیح است، وجود داشته باشد (کیس و سوانسون ۲۰۰۲).

در خصوص نوع گزینه‌ها باید گفت که تقریباً هر موضوعی می‌تواند به عنوان گزینه در سؤال «جورکردنی گسترده» مطرح شود:

نمونه‌هایی از انواع گزینه‌های قابل استفاده برای سؤال «جورکردنی گسترده»

اجزای مولکولی	انواع پپیتیدها	نقایص متبولیک
انواع سلول	نام هورمون‌ها	بیماری‌های ایمنی
اجزای سلول	الکترولیت‌ها	نقایص مادرزادی
انواع بافت‌ها	نام آنزیم‌ها	مراحل پاتولوژیک
نام شریان‌ها	انواع نوروترانسمیترها	میکروارگانیسم‌ها
نام عصب‌ها	انواع ویتامین‌ها	توکسین‌ها
نام عضلات	انواع اسیدهای آمینه	تشخیص‌ها
ارگان‌ها و اجزای آنها	کاریوتایپ	داروها و طبقه‌بندی آنها

اما مهم است که در یک سؤال، تمام گزینه‌ها همگون و در یک راستا باشند. مثلاً همه نام بیماری باشند یا همه نام دارو باشند. از قرار دادن گزینه‌های ناشایع و نامرتبط باید اجتناب شود. گزینه‌ها از نظر شکل لغوی و دستور زبان نیز باید با یکدیگر تجانس داشته باشند. گزینه‌ها باید به صورت یک کلمه یا یک عبارت کوتاه نوشته شوند و فعل نداشته باشند و در مرتب کردن آنها بهتر است ترتیب الفبایی رعایت شود. در جدول ۲-۷ انواع سؤالات هدایت‌کننده‌ای که قبلاً ذکر شد، به همراه گزینه‌های پیشنهادی مناسب برای هر یک از آنها آورده شده است.

جدول ۲-۷ : انواع چارچوب‌های رایج برای نگارش سؤال «جورکردنی گسترده»

سؤال هدایت‌کننده	گزینه‌ها
برای هر یک از بیماران زیر، [ساختار]ی را که احتمالاً [آسیب دیده/کمبود دارد/ناقص است]، انتخاب کنید.	عصب، شریان، عضله، آنزیم، هورمون، پروتئین، نوروترانسمیتر، سلول
برای هر یک از بیماران زیر، [یافته] مورد انتظار را انتخاب کنید.	یافته‌های معاینات فیزیکی، جواب تست‌های آزمایشگاهی، یافته‌های پاتولوژی یا اتوپسی، نتایج آنالیز DNA
برای هر یک از بیماران زیر، محتمل‌ترین [علت] را انتخاب کنید.	نقایص آنزیمی، مکانیسم‌های التهابی، نقائص ژنتیکی، انواع مکانیسم بیماری‌ها، نام داروها
برای هر یک از بیماران زیر، [درمان] مناسب را انتخاب کنید.	انواع داروها، انواع آنزیم‌ها، انواع ویتامین‌ها
برای هر یک از بیماران زیر، اقدام مناسب بعدی را انتخاب کنید.	انواع درمان‌های دارویی، انواع تست‌های تشخیصی، انواع اقدامات حمایتی، ترکیب موارد بالا

نگارش پایه‌های سؤال

برای سؤال «جورکردنی گسترده» بیش از یک پایه سؤال لازم است. معمولاً تعداد پایه‌ها کمتر از تعداد گزینه‌ها است. به این ترتیب شانس حدس زدن کاهش پیدا می‌کند.

پایه‌ها باید از لحاظ ساختاری با یکدیگر همگون باشند. مثلاً اگر در یکی از آنها جواب آزمایش بیمار ارائه شده است، یا اگر یکی از پایه‌ها حاوی اطلاعات مربوط به نژاد یا شغل بیمار است، در بقیه هم باید مورد مشابه وجود داشته باشد. در همین راستا، توصیه می‌شود که از این کار اجتناب شود که یکی از پایه‌ها بر مبنای بیمار بزرگسال و پایه دیگر بر اساس بیمار کودک طرح شود زیرا دانشجو صرفاً با توجه به سن می‌تواند قسمت عمده‌ای از گزینه‌ها را حذف کند و از ارزش سؤال کاسته می‌شود.

گاهی برای گزینه‌هایی که مهم و شایع هستند، می‌توان بیش از یک پایه سؤال در نظر گرفت. در این حالت، باید در سؤال هدایت‌کننده برای دانشجو توضیح داده شود که می‌تواند بیش از یک بار از یک گزینه استفاده کند.

مانند آنچه در مورد سؤال «چندگزینه‌ای با بهترین پاسخ» گفته شد، پایه سؤالی که بر مبنای یک مورد بالینی طراحی می‌شود، زمینه بسیار مناسبی برای سنجش میزان درک دانشجو از کاربرد دانش ایجاد می‌کند. در حیطه علوم بالینی، اطلاعاتی که در پایه سؤال در مورد بیمار ارائه می‌شود، معمولاً به صورت شرح‌حال و شامل چندین مورد زیر است: سن، جنس، شکایت اصلی، محل مراجعه، شرح‌حال، سابقه خانوادگی، معاینات فیزیکی و نتایج آزمایش‌ها.

در اینجا هم طرح سؤال بر مبنای مورد بالینی، محدود به سنجش علوم بالینی نمی‌شود. برای ارزیابی دانش فرد در حوزه علوم پایه نیز توصیه می‌شود پایه سؤال بر اساس مشکل بیمار طراحی شود. البته باید توجه داشت که چون شرح‌حال معمول بیمار برای دانشجویان این مقطع خیلی شناخته شده نیست، معمولاً اطلاعات به صورت دیگری ارائه می‌شود.

نمونه سؤال آناتومی

محور: آسیب شریانی در مغز

الف) Left anterior cerebral artery
ب) Right anterior cerebral artery
ج) Left middle cerebral artery
د) Right middle cerebral artery
ه) Left lenticulostriate arteries
و) Right lenticulostriate arteries

برای هر یک از بیماران زیر با ضایعه نورولوژیک، مشخص کنید احتمالاً کدام شریان آسیب دیده است؟

۱. پیرمرد ۷۲ ساله، هایپررفلکسی و ضعف اندام تحتانی راست پیدا کرده است. قدرت دست راست و حرکات صورت او نرمال است.

۲. آقای ۶۸ ساله با همی‌پارزی اسپاستیک راست و فلج دو سوم تحتانی عضلات سمت راست صورت مراجعه کرده است. تکلم او فلوئنت است و درک شفاهی و کتبی نرمال دارد.

نمونه سؤال فارماکولوژی

محور: عوارض دارویی

الف) آسپیرین ب) آمیوداورون ج) پنی سیلین د) تتراسایکلین

ه) سولفاسالازین و) فوروزمایِد ز) مترونیدازول ح) وراپامیل

برای هر یک از بیماران زیر، مشخص کنید احتمالاً کدام دارو عارضه جانبی ایجاد کرده است؟

۱. آقای ۵۶ ساله با آریتمی بطنی که از ۵ ماه پیش داروی ضدآریتمی مصرف می کند و اکنون دچار تنگی‌نفس پیش‌رونده، سرفه و تب شده است. در آزمایش خونی، ESR افزایش یافته و گرافی سینه، پنومونی منتشر اینترستیشیال نشان می‌دهد. تست‌های عملکرد ریوی، نشان دهنده کاهش ظرفیت انتشار CO هستند.

۲. آقای ۶۲ ساله با بیماری انسدادی مزمن ریوی، تحت درمان با داروی فشارخون قرار گرفته و بعد از ۲ هفته، با شکایت تشدید تنگی نفس مراجعه کرده است و در سمع ریه ویز دارد.

نمونه سؤال فیزیولوژی

محور: گازهای خون شریانی

	pH	PO2 mm Hg	PCO2 mm Hg	HCO3 - mEq/L
الف)	7.15	89	33	11
ب)	7.15	89	42	8
ج)	7.30	08	65	38
د)	7.40	001	04	25
ه)	7.50	001	42	18
و)	7.50	65	33	25

برای هر یک از بیماران زیر، محتمل‌ترین ترکیب گازهای خونی را انتخاب کنید.

۱. آقای ۲۲ ساله با سابقه پرادراری و پرنوشی از سه هفته قبل، به علت استفراغ و کاهش سطح هوشیاری طی ۱۲ ساعت گذشته که در گلوکز و کتون آزمایش ادرار او ۴+ شده است.

۲. خانم ۲۵ ساله ۱۲ ساعت بعد از اقدام به خودکشی به اورژانس آورده شده است. او تقریباً ۱۰۰ قرص ۵۰۰ میلی گرمی آسپیرین مصرف کرده است.

برای بحث کامل‌تر در خصوص پایه سؤال، به قسمت سؤالات چندگزینه‌ای مراجعه کنید. بر این نکته تاکید می‌شود که سؤال «جورکردنی گسترده» قرار است میزان فهم و درک دانشجو را از کاربرد دانشی که فراگرفته است، بسنجد. بنابراین ساختار سؤال نباید شبیه سؤالات جورکردنی دبستان شود که از دانش‌آموز می‌خواهد کلمات و جملات کوتاه را با خطوط متقاطع به هم مرتبط کند.

مرور و ارزیابی سؤالات

مرور سؤال قبل از برگزاری آزمون توسط خود طراح و همکاران و سپس اصلاح آن در صورت لزوم اهمیت زیادی دارد. در بسیاری از موارد، سؤالات در جلساتی که به همین منظور تشکیل می‌گردد، مورد ارزیابی قرار می‌گیرند. در این جلسات، کارشناسان آموزش دیده در کنار خود طراحان سؤال به مرور سؤالات می‌پردازند. برای این کار می‌توان با استفاده از چک لیست‌هایی که نمونه‌ای از آنها در جدول ۳-۷ آمده است، سؤالات را از نظر محتوا، سطح شناختی، خطاهای ساختاری، رعایت قواعد دستوری، نقطه گذاری، نکات نگارشی و ... بررسی کرد و در مورد کیفیت آن به طراح سؤال بازخورد داد تا بازبینی سؤال صورت بگیرد یا با سؤال بهتر جایگزین شود. همچنین، پس از برگزاری آزمون، شاخص‌های دشواری و تمیز و پایایی و روایی آزمون را باید محاسبه کرد و به طراح در مورد کیفیت سؤالات بازخورد داد.

جدول ۳-۷ : چک لیست ارزیابی سؤالات «جورکردنی گسترده»

ارزیابی کلی		
۱	آیا سؤال، مرتبط با موضوعات مندرج در بلوپرینت یا اهداف دوره است؟	بله
۲	آیا سطح دشواری سؤال متناسب با سطح فراگیران است؟	بله
۳	آیا قواعد دستوری، املایی، نقطه گذاری و آیین نگارش رعایت شده است؟	بله
۴	اگر سؤال دارای نمودار، شکل یا عکس است، آیا الصاق شده‌اند؟	بله
۵	آیا ساختار سؤال شبیه سؤالات جورکردنی دبستان است که از دانش آموز می‌خواهد کلمات و جملات کوتاه را با خطوط متقاطع به هم مرتبط کند؟	خیر
۶	آیا تمام محتوای یک سؤال (محور، پایه و گزینه‌ها) در یک صفحه چیده شده اند؟	بله
ارزیابی سؤال هدایت‌کننده		
۷	آیا سؤال هدایت‌کننده به روشنی مشخص کرده است که چگونه باید به سؤال پاسخ داده شود؟ (شکل مخدوش: هر سؤال را با گزینه مربوطه جور کنید)	بله
۸	آیا می‌توان بدون نگاه کردن به گزینه‌ها به سؤال پاسخ داد؟	بله
۹	آیا اگر برای یک گزینه، بیش از یک پایه سؤال در نظر گرفته شده است، برای دانشجو توضیح داده شده که از یک گزینه می‌تواند بیش از یک بار استفاده کند؟	خیر
ارزیابی گزینه‌های سؤال		
۱۰	آیا گزینه‌ها همگون و در یک راستا هستند (مثلاً همه نام بیماری یا همه نام دارو)؟	بله
۱۱	آیا گزینه‌ها بر اساس ترتیب الفبایی ردیف شده‌اند؟	خیر
۱۲	آیا گزینه‌ها از نظر قواعد دستوری، طول گزینه و زبان تکنیکی مشابه هستند؟	بله
۱۳	آیا گزینه‌ها به صورت یک کلمه یا یک عبارت کوتاه بدون فعل نوشته شده‌اند؟	بله
ارزیابی پایه‌های سؤال		
۱۴	آیا پایه سؤال شفاف و کامل است؟	بله
۱۵	آیا بیش از یک پایه وجود دارد؟	بله
۱۶	آیا پایه سؤالات از لحاظ ساختاری با یکدیگر همگون هستند؟ (مثلاً اگر در یکی از آنها، جواب آزمایش بیمار ارائه شده، یا اگر یکی از پایه‌ها، حاوی اطلاعات مربوط به نژاد یا شغل بیمار است، در بقیه هم باید مورد مشابه وجود داشته باشد.)	بله
۱۷	آیا محتوای پایه‌ها غنی است؟ مثلاً یک مورد بالینی مبنای طرح سؤال بوده است؟	بله
۱۸	آیا سؤال به جای ارزیابی محفوظات به ارزیابی توان به کارگیری اطلاعات می‌پردازد؟	بله
۱۹	آیا واحد داده‌های آزمایشگاهی قید شده است؟	خیر
۲۰	آیا از کلمات تخصصی غیرمعمول یا مخفف استفاده شده است؟	خیر

تصحیح سؤال «جورکردنی گسترده»

سؤال جورکردنی گسترده چون بسته‌پاسخ است، به صورت کلی تصحیح ساده‌ای دارد. اما از آنجا که تعداد گزینه‌های آن زیاد است، نمی‌توان از پاسخنامه‌های معمول که چهار یا پنج‌گزینه‌ای هستند، استفاده کرد. اگر قرار است تصحیح این سؤالات به صورت ماشینی انجام شود، نیاز به پاسخنامه مخصوص دارد که خیلی رایج نیست.

به همین علت، در مواردی که تنها بخش محدودی از سؤالات آزمون از نوع «جورکردنی گسترده» هستند، شاید ترجیح داده شود که تصحیح به صورت دستی صورت گیرد. اما باید دقت داشت که صرف هزینه برای فراهم کردن زمینه‌های نرم‌افزاری و سخت‌افزاری به منظور تصحیح نرم‌افزاری، علی‌رغم مشکلاتی که دارد، به سرعت و دقت کار کمک می‌کند و علاوه بر آن تحلیل کمی و آماری آزمون را نیز امکان‌پذیر می‌کند (برتون ۲۰۰۹)

سودمندی سؤال «جورکردنی گسترده»

به نظر می‌رسد که امروزه سؤال «جورکردنی گسترده» جایگاه خود را در میان سؤالات کتبی به خوبی باز کرده است. همان‌طور که قبلاً عنوان شد، بورد ملی ارزیابان پزشکی که مسوول برگزاری آزمون‌های گواهینامه پزشکی در امریکا است، به سمت کاهش تنوع سؤالات و محدود کردن آن به انواع A و R در آزمون گواهینامه پزشکی امریکا پیش رفته است (کیس و سوانسون ۲۰۰۲).

در این قسمت سعی می‌کنیم بر اساس فرمول ون‌درولوتن و بررسی شواهد و مقالات، سودمندی سؤال «جورکردنی گسترده» را مورد بحث قرار دهیم. البته ذکر دو نکته لازم است. یکی اینکه تعداد مقالاتی که در مورد کیفیت سؤال جورکردنی گسترده در آموزش پزشکی منتشر شده‌اند، زیاد نیستند و دوم اینکه سؤال جورکردنی معمولی (هنگامی که دو ستون ارائه می شود و دانشجو باید موارد متناظر را به یکدیگر ارتباط دهد)، علی‌رغم اینکه مشابه سؤال «جورکردنی گسترده» است، دقیقاً مطابق دستورالعمل ارائه شده نیست. این شکل سؤال به دلیل عدم رعایت اصول توصیه شده مورد انتقاد قرار گرفته است و شواهدی در حمایت از سودمندی آن در دسترس نیست. بنابراین مطالبی که در اینجا ارائه می‌گردد، در خصوص سؤال «جورکردنی گسترده» با همان شکل و فرمتی است که در این فصل توضیح داده شد.

روایی سؤال «جورکردنی گسترده»

هر چند که هر دو نوع سؤال چندگزینه‌ای و جورکردنی به شرط اینکه خوب طراحی شوند، می‌توانند فراگیر را وادار کنند که برای پاسخ دادن به سؤال اجزای مختلف یک موضوع را به هم ارتباط دهد و در نتیجه می‌توانند به خوبی مهارت حل مسأله دانشجو را ارزیابی کنند، اما به نظر می‌رسد سؤال جورکردنی گسترده بهتر از سؤال چندگزینه‌ای قادر است استدلال بالینی رو به جلو[1] و یادگیری عمیق را بسنجد و همچنین دانشجویان توانمند را از دانشجویان با عملکرد مرزی[2] افتراق دهد (بورتون ۲۰۰۹). البته این ادعا که گفته می‌شود این نوع سؤال می‌تواند سطوح شناختی بالا و توانایی استدلال بالینی فراگیران را مورد سنجش قرار دهد، انتظاری است که با توجه به ویژگی‌های آن ایجاد شده و تا حد زیادی نظری است. برای بررسی تجربی این ادعا، به نتایج دو مطالعه می‌توان اتکا کرد:

- وس و همکاران[3] در پژوهش خود روایی سازه[4] سؤالات «جورکردنی گسترده» را بررسی کردند. به این منظور، ضریب

1. Forward reasoning
2. Borderline
3. Wass et al
4. Construct Validity

همبستگی ۲۵ سؤال جورکردنی گسترده با سؤالات کوته‌پاسخ، درست–نادرست، تشریحی، OSCE و Long Case را محاسبه کردند که به ترتیب ۰/۶۰، ۰/۴۳، ۰/۰۸–، ۰/۸۳ و ۰/۴۸ به دست آمد. نویسندگان به دلیل ارتباط خوب این نوع سؤال با آزمون‌های بالینی و آزمون‌های حل مسأله نتیجه گرفتند که سؤال «جورکردنی گسترده» قادر است مهارت حل مسأله بالینی را در فراگیران بسنجد (وس و همکاران ۲۰۰۱).

- در سال ۲۰۰۵ مطالعه‌ای توسط بولنز و همکاران[1] برای ارزیابی فرایند تفکر در هنگام پاسخ به سؤال جورکردنی گسترده ترتیب داده شد. این کار با مقایسه نحوه حل مسأله دانشجویان و دستیاران صورت گرفت. از ۲۵ دانشجوی سال آخر و ۲۰ دستیار داخلی سال چهارم یا پنجم خواسته شد بیست سؤال «جورکردنی گسترده» را حل کنند. تمام سؤالات در مورد تشخیص یا پاتوژنز بیماری بودند. جلسات ضبط و سپس گفته‌ها پیاده و تحلیل شدند. پژوهشگران با بررسی نحوه پاسخگویی به سؤالات و مقایسه نحوه استدلال دستیاران و دانشجویان به این نتیجه رسیدند که سؤال «جورکردنی گسترده» قادر است توانایی استدلال بالینی رو به جلو را به خوبی بسنجد. با این فرض که دانش طب داخلی دو گروه نباید تفاوت زیادی با یکدیگر باشد و انتظار می‌رود تفاوت بین نحوه استدلال دو گروه نیز کم باشد، نویسندگان ذکر کردند که این نوع سؤال برای تشخیص قدرت استدلال بالینی آن‌قدر حساس بوده که توانسته تفاوت اندک بین دو گروه را مشخص کند (بولنز و همکاران ۲۰۰۵).

پایایی سؤال «جورکردنی گسترده»

گفته می‌شود یکی از مسائلی که روی پایایی اثر می‌گذارد، پاسخ حدسی است. به نظر می‌رسد از آنجا که تعداد گزینه‌ها در سؤال «جورکردنی گسترده» بسیار بیشتر از سؤال چندگزینه‌ای (سه، چهار یا پنج گزینه) است، احتمال اینکه دانشجو به صورت تصادفی با حدس زدن یا رد گزینه‌ها به جواب برسد، بسیار اندک است و بنابراین تاثیر چندانی روی پایایی ندارد. در این نوع سؤال، اگر دانشجو جواب را نداند و بخواهد به صورت تصادفی یکی از گزینه‌ها را انتخاب کند، احتمال اینکه پاسخ نادرست بدهد، خیلی زیاد است (کیس و سوانسون ۲۰۰۲).

به صورت کلی، نتایج مطالعات صورت گرفته تاکنون نشان می‌دهند که ثبات و پایایی «سؤال جورکردنی گسترده» همانند یا بیشتر از «سؤال چندگزینه‌ای با بهترین پاسخ» است (کیس و سوانسون ۲۰۰۲، برتون ۲۰۰۹).

- در یک مطالعه، پایایی ۱۸ سؤال پنج‌گزینه‌ای با پایایی ۱۸ سؤال «جورکردنی گسترده پاسخ» که بین ۹ تا ۲۳ گزینه داشتند، مقایسه شد. ضریب تعمیم‌پذیری سؤالات پنج‌گزینه‌ای و «جورکردنی گسترده» به ترتیب ۰/۴۲ و ۰/۵۵ بود (کیس و همکاران[2] ۱۹۹۴).
- در مطالعه دیگر، سؤال پنج‌گزینه‌ای با سؤال «جورکردنی گسترده» ۲۰ گزینه‌ای مقایسه شد. آلفای کرونباخ ۲۴۰ سؤال پنج گزینه‌ای و ۲۲۰ سؤال «جورکردنی گسترده» به ترتیب ۰/۸۳ و ۰/۹۰ به دست آمد (فندرسون و همکاران[3] ۱۹۹۷).
- نتایج پژوهش فندرسون و همکاران نشان می‌دهد که سؤال جورکردنی گسترده با ۲۰ گزینه به اندازه سؤال بازپاسخ که ۱۰۰ پاسخ احتمالی دارد، روا و پایا است (فندرسون و همکاران ۱۹۹۷).
- یک مطالعه به این نتیجه رسید که ۵۲ سؤال جورکردنی گسترده برای دستیابی به پایایی ۰/۷۵ لازم است و چنانچه این تعداد به ۱۰۵ سؤال برسد، پایایی ۰/۸۵ خواهد شد (کریتر و همکاران[4] ۱۹۹۹).

1. Beullens et al.
2. Case et al.
3. Fenderson et al.
4. Kreiter et al.

تاثیرآموزشی سؤال «جورکردنی گسترده»

اگر به نتیجه مطالعات در مورد امکان سنجش توانایی استدلال و حل مسأله توسط سؤال «جورکردنی گسترده» اتکا کنیم، می‌توان گفت پیغامی که این سؤال به دانشجویان منتقل می‌کند، این است که به جای حفظ کردن مطالب، به مطالعه عمیق بپردازند. البته هنوز شواهد تجربی برای این موضوع در دسترس نیست.

در هر حال، داشتن کلید سؤالات بعد از آزمون به دانشجویان کمک می‌کند روی آموخته‌های خود تأمل کنند و اشتباهات خود را اصلاح کنند.

مقبولیت، هزینه و قابلیت اجرای سؤال «جورکردنی گسترده»

طراحی سؤال «جورکردنی گسترده» سخت‌تر و زمان‌برتر از سؤال چندگزینه‌ای است و چنانچه طراح قبلاً با آن آشنایی نداشته باشد، برای طراحی آن لازم است وقت به مراتب بیشتری صرف کند (برتون ۲۰۰۹). همچنین، از آنجا که فرمت نسبتا جدیدی است، لزوم توجه توانمندسازی استادان، در نظر گرفتن جلسات تمرین و مرور و ارائه بازخورد به سؤالاتی که طراحی کرده‌اند، حائز اهمیت است (شوورث و ون‌درولوتن ۲۰۰۳). همچنین، توصیه شده است که دانشجویان قبلاً با این نوع سؤال آشنا شده باشند و در جریان جزئیات آزمون قرار بگیرند.

تصحیح سؤال «جورکردنی گسترده» به صورت کلی ساده است. همان‌طور که قبلاً اشاره شد، باید توجه داشت که در صورت تصحیح ماشینی، پاسخنامه مخصوص مورد نیاز است که چندان رایج نیست. البته همان‌گونه که پیشتر نیز تأکید شد باید دقت داشت که صرف هزینه برای نرم‌افزاری کردن تصحیح دارای علاوه بر افزایش سرعت و دقت فرایند واجد مزایای دیگری مانند تحلیل کمی و آماری آزمون نیز است. (برتون ۲۰۰۹).

مسأله دیگری که به عنوان محدودیت اجرایی مطرح می‌شود، این است که شاید نتوان این نوع سؤال را برای تمام موضوعات طرح کرد. یافتن گزینه‌های همگون و موثر به تعداد مناسب مهم است و برای همه مباحث کار ساده‌ای نیست. استفاده از سؤال «جورکردنی گسترده» مخصوصاً زمانی که قرار است تعداد زیادی سؤال در مورد تصمیم‌گیری درباره یک موضوع خاص پرسیده شود، مناسب‌تر است (شوورث و ون‌درولوتن ۲۰۰۳).

مطالعه بولنز ۲۰۰۲

در سال ۲۰۰۲ مقاله‌ای منتشر شد که به تشریح تجربه دانشکده پزشکی دانشگاه لوون در بلژیک در خصوص آزمون نهایی پرداخت که قسمتی از آن به صورت «جورکردنی گسترده» و بر مبنای سناریو و بیمار طراحی شده بود. دانشجویان باید گزینه صحیح را از بین ۷ تا ۲۶ پاسخ احتمالی انتخاب می‌کردند. اعضای هیأت علمی از ده رشته مختلف حدود ۹۰۰ سؤال طراحی کردند و برای سنجش کیفیت سؤالات، آنها را در دسته‌های ۱۰۰ تایی به ۲۵۱ دانشجوی سال آخر که قبلاً با این شکل سؤال آشنایی نداشتند، ارائه دادند. مدت آزمون ۴ ساعت طول کشید. خستگی در طول آزمون تاثیری در نمرات نداشت اما افزایش آشنایی با نوع سؤال در طول آزمون مشهود بود. سؤالاتی که ضریب همبستگی آنها با نمره کل آزمون منفی بود، حذف شدند تا آلفای کرونباخ ۰/۸۰ به دست آمد. روایی صوری سؤالات با استفاده از پرسشنامه تایید شد. روایی ملاکی از طریق مقایسه با نمرات آزمون سالانه به دست آمد. نویسندگان نتیجه‌گیری کردند که با ارائه ۱۰۰ سؤال «جورکردنی گسترده» به هر دانشجو، می‌توان آزمون نهایی روایی و پایایی برای سنجش دانش بالینی برگزار نمود (بولنز ۲۰۰۲).

منابع

1. Alcolado J, Mir MA. Extended-matching questions for finals. Churchill Livingstone, Edinburgh; 2002
2. Amin Z, Seng CY, Eng KH. Practical guide to medical student assessment. World Scientif- ic, Singapore, 2006
3. Anderson J. Multiple-choice questions revisited. Medical Teacher 2004;26(2):110–113
4. Baird AS. The new Extended Matching Question (EMQ) paper of the MFSRH Examination. J Fam Plann Reprod Health Care 2010;36(3):171-3
5. Beullens J, Damme BV, Jaspaert h, Janssen PJ. Are extended-matching multiple-choice items appropriate for a final test in medical education? Medical Teacher 2002;24(4): 390–395
6. Beullens J, Struyf E, Damme BV. Do extended matching multiple-choice questions measure clinical reasoning? Medical Education 2005; 39: 410–417
7. Burton JL. How to write and how to answer EMQs? Obstetrics, gynaecology and reproductive medicine 2009;19(12):359-361
8. Case SM, Swanson DB. Extended matching items: A practical alternative to free response questions, Teaching and Learning in Med 1993;5(2): 107-115
9. Case SM, Swanson DB, Ripkey DR. Comparison of items in five-option and extended-matching formats for assessment of diagnostic skills. Academic Medicine 1994;69(10 Suppl.):S1-S3.
10. Case SM, Swanson DB. Constructing Written Test Questions for the Basic and Clinical Sciences, Philadelphia, National Board of Medical Examiners; 2002
11. Chandratilake M, Davis M, Ponnamperuma G. Assessment of medical knowledge: the pros and cons of using true/false multiple choice questions. Natl Med J India. 2011 Jul-Aug;24(4):225-8.
12. Duthie S, Hodges P, Ramsay I, Reid W. EMQs: a new component of the MRCOG Part 2 exam. The Obstetrician & Gynaecologist 2006;8:181–185
13. Duthie S, Fiander DM A, Hodges P. EMQs: a new component of the MRCOG Part 1 examination. The Obstetrician & Gynaecologist 2011; 9(3):189-194
14. Fenderson BA, Damjanov I, Robeson MR, Veloski JJ, Rubin E. The virtues of extended matching and uncued tests as alternatives to multiple choice questions, Human Pathology 1997;28:526-532
15. Haladyna M, Downing SM, Rodriguez MC. A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. Applied Measurement in Education 2002;15(3):309-34
16. Kreiter CD, Ferguson K, Gruppen LD. Evaluating the usefulness of computerized adaptive testing for medical in-course assessment. Acad Med 1999; 74:1125-1128.
17. Schuwirth LWT, van der Vleuten CPM. ABC of learning and teaching in medicine: Written assessment. BMJ 2003;22;(326):643645.

18. Schuwirth LW, van der Vleuten CP. Different written assessment methods: what can be said about their strengths and weaknesses? Med Educ 2004; 38(9): 974-9.
19. Wass V, McGibbon D, van der Vleuten CPM. Composite undergraduate clinical examinations: how should the components be combined to maximize reliability? Medical Education 2001;35:326-330.

فصل | ۸ |

سؤال «درست- نادرست»

ساختار سؤال «درست–نادرست»

سؤالات «درست–نادرست» که در منابع مختلف با اسامی متفاوتی از آنها یاد شده است، سؤالات شایعی هستند و زیاد استفاده می‌شوند. بر اساس مقاله‌ای که بیش از یک دهه قبل منتشر شده، سؤال «درست–نادرست» شایع‌ترین فرمی است که در آن زمان در انگلیس مورد استفاده قرار می‌گرفت (فوول و جولی[1] ۲۰۰۰). از آنجا که سؤال، یا درست یا غلط است، به آن دوگزینه‌ای یا انتخاب دوحالته[2] هم گفته می‌شود. همان‌طور که در فصل اول همین بخش ذکر شد، سؤالات مختلفی را می‌توان از این نوع در نظر گرفت: «درست–نادرست ساده»، «درست–نادرست متعدد» یا نوع X، نوع J یا K، نوع C و نوع E. همه انواع فوق دارای یک ویژگی مشترک می‌باشند: در همه آنها دانشجو باید گزینه «کاملاً صحیح» را انتخاب کند. در این فصل تنها به نوع درست–نادرست ساده و متعدد می‌پردازیم.

سؤال «درست–نادرست ساده» از یک هدایت‌کننده تشکیل شده است که از دانشجو می‌خواهد مشخص کند پایه سؤال که در پی میاید، درست است یا غلط. پایه به صورت جمله‌ای است که باید کاملاً غلط یا کاملاً صحیح باشد. گزینه‌ها هم که در دو حالت درست و نادرست نوشته می‌شوند.

نمونه سؤال «درست–نادرست»

مشخص کنید که جمله زیر درست است یا نادرست.
سیستیک فیبروزیس یک بیماری وابسته به x مغلوب است.

درست ☐ نادرست ☐

البته ذکر این نکته لازم است که در برخی از متون، اجزای سؤال «درست–نادرست» به صورت دیگری ذکر شده‌اند. به این معنا که جمله کاملاً درست یا غلط به عنوان گزینه سؤال در نظر گرفته شده است و پایه سؤال همان قسمتی که در اینجا تحت عنوان هدایت‌کننده آمده است، نام است.

در سؤال «درست–نادرست متعدد» یا همان سؤال نوع X معمولاً مجموع چند «سؤال درست–نادرست ساده» با هم نوشته می‌شوند. در ابتدا سؤال هدایت‌کننده مطرح می‌شود که از دانشجو می‌خواهد پایه‌های صحیح و غلط را مشخص کند. به دنبال آن، پایه‌ها نوشته می‌شوند که به صورت چند جمله هستند که یک یا چند تا از آنها درست و بقیه غلط هستند. در حقیقت، این سؤالات، ترکیبی از سؤالات چندگزینه‌ای و «درست–نادرست» هستند.

1. Fowell & Jolly
2. Binary choice

سؤال زیر یک مورد سؤال «درست–نادرست» قابل‌قبول با پایه شفاف است که گزینه‌های آن کاملاً غلط یا کاملاً صحیح هستند:

نمونه‌ای از سؤال «درست–نادرست» متعدد

در مورد هر یک از جملات زیر مشخص کنید که درست یا نادرست است:

۱. سیستیک فیبروزیس یک بیماری وابسته به X مغلوب است.	درست ☐	نادرست ☐
۲. هموفیلی A یک بیماری وابسته به X مغلوب است.	درست ☐	نادرست ☐
۳. دیستروفی عضلانی دوشن یک بیماری وابسته به X مغلوب است.	درست ☐	نادرست ☐
۴. بیماری تای–ساکس یک بیماری وابسته به X مغلوب است.	درست ☐	نادرست ☐

برای جلوگیری از تکرار کلمات مشابه، این سؤال به این صورت نیز مطرح می‌شود:

شکل دیگر سؤال «درست–نادرست» متعدد

مشخص کنید کدام بیماری‌ها، وابسته به X مغلوب هستند؟

۱. سیستیک فیبروزیس ۲. هموفیلی A ۳. دیستروفی عضلانی دوشن ۴. بیماری تای–ساکس

احتمالاً شکل سؤال فوق بسیار آشناست. مخصوصاً اگر تنها یک گزینه از بین چهار گزینه درست باشد. با اینکه ظاهر آن بسیار شبیه سؤال «چندگزینه‌ای با بهترین پاسخ» است، اما در واقع با آن تفاوت دارد زیرا روی یک موضوع واحد متمرکز نمی‌شود و هر گزینه به تنهایی و مستقل از بقیه قابل بررسی است. با مرور سؤالات امتحانی به سادگی می‌توان دریافت که بسیاری از سؤالاتی که در حال حاضر به عنوان سؤال چندگزینه‌ای برای دانشجویان علوم پزشکی طرح می‌شود، از همین فرمت پیروی می‌کنند. اما همان‌طور که گفته شد، این سؤال از نظر مفهومی یک سؤال «درست–نادرست» است و ویژگی‌های روان‌سنجی آن متفاوت از سؤال چندگزینه‌ای است و نباید به جای آن به کار رود. در دستورالعمل‌های طراحی سؤال چندگزینه‌ای که قبلاً به آن پرداختیم، به کرات تاکید شده است که از طراحی این نوع سؤال به جای سؤال چندگزینه‌ای اجتناب شود. طراحان به شدت از نوشتن پایه سؤال به صورت «کدام یک از گزینه‌های زیر درست است؟» منع شده‌اند.

نه تنها استفاده از این فرمت به جای سؤال چندگزینه‌ای توصیه نمی‌شود، بلکه امروزه از استفاده از سؤال «درست–نادرست» در بسیاری از آزمون‌های مهم و سرنوشت‌ساز، بنا به دلایلی که بعدا توضیح داده می‌شود، اجتناب می‌شود (کیس و سوانسون ۲۰۰۲، کامپبل ۲۰۱۱). تشخیص این دو نوع سؤال به آسانی با آزمون پوشاندن گزینه‌ها قابل انجام است.

توجه

هنگام طراحی سؤال چندگزینه‌ای، از طرح سؤال به صورت «کدام یک از گزینه‌های زیر درست است؟» اجتناب کنید. این سؤال، شامل چند سؤال «درست–نادرست» است و در واقع یک سؤال «چندگزینه‌ای با بهترین پاسخ» نیست.

خطاهای طراحی سؤال «درست–نادرست»

خطاهای طراحی سؤال «درست–نادرست» کمابیش مشابه سؤال چندگزینه‌ای هستند اما طراحی این نوع سؤال بدون خطا کار دشواری است (آلبنز ۱۹۹۳، شوورث و ون‌درولوتن ۲۰۰۳). در روند مرور سؤالات، لزوم حذف یا بازنویسی سؤالات «درست–نادرست» خیلی بیشتر از سایر سؤالات اتفاق می‌افتد (کیس و سوانسون ۲۰۰۲). رایج‌ترین خطایی که مختص سؤال «درست–نادرست» است، استفاده از قیود مطلق (همیشه، هرگز، اصلاً) و قیود مبهم (معمولاً، غالباً، گاهی) است

(چاندراتیلکه و همکاران ۲۰۱۱). در اینجا به صورت اجمالی به آنها اشاره می‌شود. برای بحث‌های تکمیلی مرتبط به فصل دوم، بخش خطاهای طراحی مراجعه کنید.

خطاهای سؤال «درست–نادرست» از آنجا ناشی می‌شود که طراحان سؤال، برای نوشتن عبارت «کاملاً درست» یا «کاملاً غلط» با مشکل مواجه می‌شوند. زیرا مطالبی در ذهن دارند و بر اساس آن فکر می‌کنند که گزینه خوبی نوشته‌اند اما معمولاً مرور دقیق سؤال توسط کسی که کلید را ندارد، نشان می‌دهد که جنبه‌های دیگری نیز وجود داشته است که طراح به آنها توجه نکرده است و باعث شده است که عبارت کاملاً درست یا کاملاً غلط نباشد. بنابراین عموم سؤالاتی که به عنوان «درست–نادرست» طرح می‌شوند، به جای اینکه «کاملاً» صحیح باشند، «تا حدودی» درست هستند. این مسأله موجب می‌شود فراگیر برای پاسخگویی به سؤال، فراتر از دانشی که باید بلد باشد، مجبور شود به این مسأله هم فکر کند که در ذهن ممتحن چه گذشته است و قضاوت کند که عبارت باید تا چه میزان صحیح باشد تا آن را به عنوان عبارت درست اعلام کند. همان‌طور که مشخص است این امر ربط چندانی به محتوایی که قرار بود ارزیابی شود، ندارد و در واقع از روایی سؤال می‌کاهد.

به سؤال زیر که به عنوان یک سؤال «درست–نادرست» طراحی شده است، اما قابل قبول نیست و واجد خطا است، توجه کنید.

نمونه‌ای از خطا در سؤال «درست–نادرست» متعدد

در مورد سیستیک فیبروزیس مشخص کنید که کدام از جملات زیر صحیح و کدام نادرست است؟

۱. بروز سیستیک فیبروزیس یک در ۲۰۰۰ است.	درست ☐	نادرست ☐
۲. کودکان مبتلا به سیستیک فیبروزیس در نوجوانی می‌میرند.	درست ☐	نادرست ☐
۳. مردان مبتلا به سیستیک فیبروزیس نازا هستند.	درست ☐	نادرست ☐
۴. سیستیک فیبروزیس یک بیماری وابسته به x مغلوب است.	درست ☐	نادرست ☐

دانشجو باید هر یک از جملات را به صورت مستقل از بقیه بررسی کند. اگر یکی درست باشد، به معنای این نیست که بقیه غلط هستند. بنابراین همان‌طور که قبلاً گفته شد، هر جمله باید به تنهایی و به صورت کامل درست یا غلط باشد. اما در مثال فوق، قابل قضاوت نیست که موارد ۱ و ۲ و ۳ کاملاً درست هستند یا غلط. به عنوان مثال، میزان بروز سیستیک فیبروزیس عددی قطعی نیست. اساسا مشخص نیست که بروز بیماری در امریکا منظور است یا در همه جوامع. وقتی متخصصان روی موضوعی توافق ندارند، چه دلیلی دارد که از دانشجو بخواهیم مطالب این چنینی را حفظ کند. یا در مورد جمله سوم، مشخص نیست که منظور نازایی در تمام مردان مبتلاست یا اینکه منظور طراح سؤال این بوده که مردان مبتلا می‌توانند نازا باشند. تنها جمله چهارم است که می‌توان قطعاً گفت صحیح است.

در این موارد که مرور سؤال توسط شخص دیگر روشن می‌کند سؤال گنگ است و مفهوم را نمی‌رساند، وقتی سؤال به طراح برگردانده می‌شود تا آن را ویرایش کند و ابهام موجود در سؤال خود را کاهش دهد، احتمال اینکه سؤال خیلی آسان شود یا صرفاً محفوظات را ارزیابی کند، زیاد می‌شود. به همین دلیل سؤال درست–نادرست، غالباً حافظه و سطوح پایین شناختی دانشجویان را مورد ارزیابی قرار می‌دهد. این موضوع برعکس قابلیت سؤالات چندگزینه‌ای است که امکان سنجش توانایی به کاربستن اطلاعات، تلفیق داده‌ها، حل مسأله، استدلال و قضاوت را دارند (کیس و سوانسون ۲۰۰۲، شوورث و ون‌درولوتن ۲۰۰۳).

در همان مثال قبلی، برای اینکه خطا اصلاح شود و جمله سوم به صورت کاملاً درست باشد، احتمالاً به شکل زیر نوشته می‌شود:

شکل دیگر سؤال

در مورد سیستیک فیبروزیس مشخص کنید که جمله زیر صحیح است یا غلط.

برخی از مردان مبتلا به سیستیک فیبروزیس نازا هستند.	درست ☐	نادرست ☐

همان‌طور که مشخص است، ابهام موجود رفع شده است و اکنون می‌توان ادعا کرد که جمله کاملاً درست است. اما مشکل دیگری پیش آمده است. کلمه «برخی» به دانشجو سرنخ می‌دهد. دانشجویی که اطلاعات زیاد و قطعی در مورد بیماری سیستیک فیبروزیس ندارد، احتمالاً با این شکل سؤال، بهتر می‌تواند حدس بزند. پس سؤال آسان شده و احتمال پاسخ حدسی زیاد شده است. از طرفی، کلمه «برخی» مبهم است و معلوم نیست به چه درصدی از افراد اطلاق می‌شود. اگر جمله به صورتی نوشته شود که کاملاً غلط باشد، احتمالاً شبیه سؤال زیر می‌شود:

شکل دیگر سؤال

در مورد سیستیک فیبروزیس مشخص کنید که جمله زیر صحیح است یا غلط.
تمام مردان مبتلا به سیستیک فیبروزیس نازا هستند. درست ☐ نادرست ☐

برای دانشجو تقریباً واضح است که در حوزه سلامت و پزشکی، معمولاً قیود مطلق «همیشه» و «هرگز» صدق نمی‌کنند و برای هر موضوعی، استثنایی وجود دارد. بنابراین، باز هم به راحتی و بدون اینکه بر محتوا مسلط باشد، با حدس زدن به پاسخ می‌رسد و تشخیص می دهد که عبارت غلط است. به این ترتیب سطح سؤال آسان می‌شود. برای توضیح بیشتر، نمونه دیگری را بررسی می‌کنیم:

نمونه‌ای از خطا در سؤال «درست–نادرست» متعدد

مشخص کنید که در مورد ارزیابی درد کدام جملات درست هستند؟
۱. نگرش شخصی پزشک در مورد درد ممکن است روی قضاوت پزشکی او اثر بگذارد.
درست ☐ نادرست ☐
۲. احساسات نامطلوب ممکن است به شکل شکایت درد جسمانی بروز کنند.
درست ☐ نادرست ☐
۳. درد ممکن است معنای نمادین داشته باشد.
درست ☐ نادرست ☐
۴. تظاهرات چهره یا بدن، معمولا سرنخی در مورد شدت درد به پزشک می‌دهند.
درست ☐ نادرست ☐

در این سؤال، عبارات مبهم به کار رفته است و کلمه ممکن در سه عبارت ۱ و ۲ و ۳ این سرنخ را به فراگیر می دهد که احتمالاً درست هستند. عبارت چهارم این مشکل را دارد که روی تعریف کلمه «معمولاً» اتفاق نظر وجود ندارد و برداشت‌های مختلفی می‌توان از آن کرد که موجب می‌شود عبارت صحیح با غلط به نظر برسد.

خطای محتمل دیگر این است که تعداد عبارت درست زیاد باشد زیرا نوشتن عبارت نادرستی که در عین حال برای عده‌ای باورکردنی باشد، سخت است. بنابراین این تمایل وجود دارد که تعداد سؤالات درست زیاد شود لذا باید توجه شود که تعادل بین تعداد گزینه‌های نادرست و درست حفظ شود (موس ۲۰۰۱، شوورث و ون‌درولوتن ۲۰۰۳).

گام های طراحی سؤال «درست–نادرست»

هرچند که بسیاری از طراحان سؤال معتقد هستند نوشتن این سؤالات آسان‌تر از طرح سؤالات چندگزینه‌ای است، حقیقت این است که طراحی سؤال «درست–نادرست» خوب و بدون ابهام کار دشواری است. اصول کلی طراحی این سؤال مانند سؤال چندگزینه‌ای است و همان ملاحظات در اینجا هم صدق می‌کند. بنابراین برای اطلاع از اصول طراحی سؤال، به فصل دوم، بخش گام‌های طراحی سؤال مراجعه کنید اما به منظور مرور قوانین و برای تاکید بر دو ویژگی سؤال «درست–نادرست» که تفاوت اساسی آن با سؤال «چندگزینه‌ای با بهترین پاسخ» است (احتمال درستی بیش از یک گزینه و لزوم درستی گزینه به صورت کامل)، در اینجا

به صورت خلاصه به چند نکته اجمالی اشاره می‌شود. مراحل زیر باید در طراحی سؤالات «درست–نادرست» انجام شوند. البته مورد اول و آخر برای تمام سؤالات این گروه یکسان است و تنها نگارش پایه است که باید با دقت صورت گیرد:

نگارش سؤال هدایت‌کننده

نگارش سؤال هدایت‌کننده «درست–نادرست» ساده است و تقریباً شکل ثابت و یکسانی دارد. مهم است که این قسمت شفاف و دقیق باشد و از عبارات مبهم اجتناب شود. معمولاً چیزی شبیه به این جمله نوشته می‌شود: «در مورد هر یک از جملات مشخص کنید که درست یا نادرست هستند.»

نگارش پایه سؤال

پایه‌ها همان عبارات و جمله‌هایی هستند که دانشجو باید در مورد درستی یا نادرستی آنها تصمیم‌گیری کند. از آنجا که پایه‌ها مستقل از یکدیگر هستند و معمولاً ارتباطی چندانی با هم ندارند، طراحی پایه در سؤال «درست–نادرست» پیچیدگی کمتری دارد اما باید توجه شود که پایه‌ها:

- عباراتی کاملاً صحیح یا کاملاً غلط باشند.
- کلماتی که به نحوی برای تشخیص درست یا غلط بودن عبارت سرنخ می‌دهند، مانند «احتمال دارد» و «ممکن است» نداشته باشند.
- کلمات مبهم مانند «معمولاً» و «غالباً» نداشته باشند.

نگارش گزینه‌ها

گزینه‌ها همیشه یکسان و خود عبارات درست یا نادرست هستند.

تصحیح سؤالات «درست–نادرست»

برای تصحیح این نوع سؤال، روش‌های مختلفی به کار رفته‌اند که اکثراً تلاشی در راستای کاهش حدس زدن دانشجویان بوده‌اند:

- نمره‌دهی مثبت: به صورت ساده، منظور کردن نمره یک برای پاسخ درست و نمره صفر برای پاسخ غلط
- نمره‌دهی منفی: به این صورت که برای پاسخ درست، نمره یک و برای پاسخ اشتباه، نمره منفی یک را در نظر گرفته شود.
- افزودن گزینه «نمی‌دانم» و محاسبه نمره منفی به این صورت که به پاسخ درست، پاسخ نادرست و پاسخ «نمی‌دانم»، به ترتیب نمره‌های یک، منفی یک و صفر تعلق می‌گیرد.
- ارزیابی اطمینان: در این مدل، دانشجو علاوه بر اینکه گزینه صحیح را مشخص می‌کند، باید درجه اطمنیان خود را از پاسخی که می‌دهد، اعلام کند.

برای توضیحات بیشتر در خصوص روش‌های نمره‌دهی فوق به قسمت «سؤالات رایج» در فصل اول و قسمت تصحیح سؤال در فصل دوم مراجعه کنید.

سودمندی سؤالات «درست–نادرست»

از اواخر دهه ۱۹۸۰ سؤالات «درست–نادرست» به دلیل محدودیت‌ها و نقاط ضعف بسیار مورد انتقاد قرار گرفتند و استفاده از آنها در بسیاری از آزمون‌های مهم و سرنوشت‌ساز پزشکی مخصوصاً در امریکا و انگلیس قطع شد (داونینگ ۱۹۹۲، کیس و سوانسون ۲۰۰۲، کمپبل ۲۰۱۱). البته این گونه نیست که در حال حاضر هیچ امتحانی به این صورت برگزار نشود. از خلال متون منتشرشده، می‌توان به آزمون عضویت کالج سلطنتی متخصصان زنان[1] اشاره کرد که برای ارزیابی دانش و محفوظات فراگیران از سؤال «درست–نادرست» در کنار سؤالات دیگر استفاده می‌کند (دوتیه[2] ۲۰۰۶). بِندر نیز در گروه بیوشیمی و بیولوژی مولکولی کالج دانشگاه لندن[3] در مقاله خود به این موضوع اشاره می‌کند که استفاده از روش‌های جدید مانند سؤال «جورکردنی گسترده» نباید باعث شود که روش‌های قبلی به کلی کنار گذاشته شوند. در عوض، او توصیه می‌کند که آزمون‌های مشابه از هر دو نوع سؤال برگزار و سپس شاخص‌های روان‌سنجی سؤالات ارزیابی شود تا بتوان با شواهد کافی تصمیم بهتری گرفت.

در اینجا بر اساس فرمول ون‌درولوتن به مواردی که در متون آموزش پزشکی به عنوان سودمندی سؤالات «درست–نادرست» مطرح شده‌اند، می‌پردازیم.

روایی سؤال «درست–نادرست»

قبلاً گفته شد که سؤال «درست–نادرست» در اکثر موارد صرفاً محفوظات را ارزیابی می‌کند و سطوح شناختی پایین دانشجویان را مورد ارزیابی قرار می‌دهد (کیس و سوانسون ۲۰۰۲، شوورث و ون‌درولوتن ۲۰۰۳). با این حال، برخی معتقدند سؤال «درست–نادرست» اگر خوب طراحی شود، می‌تواند سطوح شناختی بالاتر را هم بسنجد؛ هرچند که نمونه‌های کمی از این موضوع در دسترس است (بندر ۲۰۰۳). واقعیت این است که عملاً در استدلال بالینی، پزشک در شرایط عدم قطعیت[4] و با احتمالات بیماری سر و کار دارد تا موقعیت‌های قطعی و کاملاً درست. ارزیابی قضاوت بالینی در سؤال «درست–نادرست» اگر هم وجود داشته باشد، معمولاً به مواقعی محدود می‌شود که مقایسه بین چند مورد صورت می‌گیرد. به عنوان مثال، در درمان بیماری X، داروی A از داروی B بهتر است (موس ۲۰۰۱، کیس و سوانسون ۲۰۰۲). حتی بندر به این موضوع اشاره کرده است که می‌توان با سایر سؤالات مانند چندگزینه‌ای و «جورکردنی گسترده» به ارزیابی «کاربرد دانش» پرداخت و با سؤال «درست–نادرست» حقایق و محفوظات را سنجید و سپس از مقایسه آنها با یکدیگر دریافت که کدام یک از دانشجویان مطالب پایه را فراگرفته‌اند ولی توانایی به کاربستن آنها را ندارند (بندر ۲۰۰۳).

نکته دیگر اینکه وقتی دانشجو به درستی تشخیص می‌دهد عبارت نادرست است، نمی‌توان دریافت که آیا جواب صحیح را می‌دانسته است یا فقط توانسته تشخیص دهد که این جمله اشتباه است. شاید اصلاً جواب درست را بلد نباشد و واقعاً اطلاعات اشتباهات دیگری دارد (داونینگ ۱۹۹۲، شوورث و ون‌درولوتن ۲۰۰۳).

این چند موضوع باعث می‌شوند که سؤال «درست–نادرست» قادر به افتراق بین دانشجویان با توانمندی بالا و پایین نباشد. زیرا اکثراً یا در عبارات نوشته شده ابهام وجود دارد یا در تلاش برای رفع ابهام، طوری نوشته می‌شوند که پاسخگویی به آنها بسیار آسان است یا با حدس زدن می‌توان به پاسخ رسید. مطالعات تجربی در زمینه قابلیت سؤال «درست–نادرست» در شناسایی و افتراق دانشجویان قوی و ضعیف نشان می‌دهند که این سؤالات در مقایسه با سؤالات چندگزینه‌ای

1. The Royal College of Obstetricians and Gynaecologists (RCOG)
2. Duthie
3. University College of London
4. Uncertainty

ضریب تمیز کمتری دارند (اووسترهوف و گلاسنپ[1] ۱۹۷۴، ابل[2] ۱۹۸۰) و نمره منفی این قابلیت را بیشتر کاهش می‌دهد (چاندراتیلکه و همکاران ۲۰۱۱).

شاید یک نکته مثبت در مورد روایی سؤال «درست–نادرست» این باشد که دانشجو به سرعت می‌تواند به این سؤال پاسخ دهد. بنابراین تعداد سؤالاتی که می‌توان در واحد زمان ارائه کرد زیاد خواهد بود و امکان نمونه‌گیری مناسب از سؤالات وجود دارد (شوورث و ون‌درولوتن ۲۰۰۳). اما در مجموع شواهد نشان می‌دهند که سؤال «درست–نادرست» روایی مطلوبی ندارد (بریج[3] و همکاران ۲۰۰۳، مک کوبریه ۲۰۰۸، کامپبل ۲۰۱۱).

پایایی سؤال «درست–نادرست»

در مورد پایایی، نتایج کمی متناقض هستند. در یک مطالعه که پایایی از طریق تکرار سؤالات در دو آزمون[4] به دست آمد، ضریب پایایی بین دو نوبت بالا گزارش شد (شوارتز و همکاران[5] ۱۹۸۶). مطالعه دیگری در دانشگاه آمستردام انجام شد و آلفای کرونباخ پنج نوبت آزمون شامل سؤالات «درست–نادرست» و کوته‌پاسخ را با یکدیگر مقایسه کرد. نتایج نشان داد که ضریب پایایی سؤالات کوته‌پاسخ بیشتر از سؤالات «درست–نادرست» است و برای دستیابی به پایایی معادل سؤالات کوته‌پاسخ، به تعداد ۶ تا ۱۰ برابر بیشتر سؤال «درست–نادرست» نیاز است (تن‌کیت[6] ۱۹۹۷). بر اساس مطالعه دیگری که در آزمون پایان دوره تخصص داخلی انجام شد، مشخص شد که سؤال «درست–نادرست» نسبت به سؤال چندگزینه‌ای نتیجه پایاتری دارد، هرچند که سطوح شناختی مورد ارزیابی در آن پایین‌تر است (داونینگ و همکاران ۱۹۹۵). موس معتقد است این سؤالات در واحد زمان پایایی کمتری دارند ولی در تعداد بالا (۶۰ تا ۹۰ سؤال) پایایی آنها جبران می‌شود (موس ۲۰۰۱).

در سؤال «درست–نادرست» مانند سایر آزمون‌های کتبی بسته‌پاسخ، احتمال حدس زدن و اثر سرنخ دادن[7] وجود دارد که در اینجا شدیدتر از سؤالات چندگزینه‌ای است و روی پایایی اثر دارد. گاهی تلاش می‌شود تا با اختصاص دادن نمره منفی یا ارزیابی اطمینان با آن مقابله شود که البته بر اساس شواهد، این اقدامات نیز اثر چندانی روی پایایی آزمون نداشته‌اند (داونینگ ۱۹۹۲، برتون ۲۰۰۲، مک کوبریه[8] ۲۰۰۸).

تاثیر آموزشی سؤال «درست–نادرست»

احتمال اینکه این نوع سؤال، دانشجویان را به سمت حفظ کردن مطالب سوق دهد، وجود دارد. با در اختیار گذاشتن کلید سؤالات می‌توان امیدوار بود که دانشجویان به اشتباهات خود پی ببرند و یادگیری خود را اصلاح کنند. هر چند نتایج یک مطالعه این مسأله را تأیید نکرده است و تاثیر آموزشی سؤال «درست–نادرست» را در قیاس با سایر سؤالات کمتر به دست آورده است (ریز[9] ۱۹۸۶).

مقبولیت، هزینه و قابلیت اجرای سؤال «درست–نادرست»

ساختار سؤال «درست–نادرست» برای دانشجویان و استادان آشناست. شاید مسأله مهم‌تر تأکید بر خطاهای طراحی آن و آموزش نحوه طراحی درست سؤال به طراحان باشد. همچنین، این نوع سؤال تصحیح آسانی دارد که به صورت غیردستی قابل انجام است. بنابرین، به صورت کلی، قابلیت اجرای آن خوب است.

1. Oosterhof & Glasnapp
2. Ebel
3. Bridge et al.
4. Test-retest
5. Schwartz et al.
6. Ten Cate et al.
7. Cueing effect
8. McCoubrie
9. Rees

منابع

1. Albanese MA. Type K and other complex multiple-choice items: An analysis of research and item properties. Educ Meas: Issues and practice 1993; 12:28–33.
2. Bandaranayake R, Payne J, White S. Using multiple response true±false multiple choice questions. Aust N Z J Surg 1999; 69:311-315.
3. Bridge D, Musial J, Frank R, Roe T, Sawilowsky S. Measurement practices: Methods for developing content valid student examinations. Med Teach 2003;25: 414–21
4. Bender DA. MCQ, EMSQ or multiple true/false questions? Bioscience Education e-journal 2003; 2. available at: http://bio.ltsn.ac.uk/journal/vol2/beej-2-L1.htm
5. Burton RF. Misinformation, partial knowledge, and guessing in true/false tests. Med Educ 2002;36:805-11
6. Case SM, Swanson DB. Constructing Written Test Questions for the Basic and Clinical Sciences. 3rd (revised) edition, Philadelphia, National Board of Medical Examiners, 2002.
7. Campbell DE. How to write good multiple-choice questions. Paediatrics and Child Health 2011;47:322–325
8. Dixon RA. Evaluating and improving multiple choice papers: true–false questions in public health medicine. Medical Education408–28:400;1994
9. Downing SM. True-false, alternate-choice, and multiple-choice items. Educ Meas: Issues and practice 1992;11:27–30.
10. Downing SM, Baranowski RA, Grosso LJ, Norcini JJ. Item type and cognitive ability measured: the validity evidence for multiple true-false items in medical specialty certification. Applied Measurement in Educ97–8:187;1995 .
11. Duthie S, Hodges P, Ramsay I, Reid W. EMQs: a new component of the MRCOG Part 2 exam. The Obstetrician & Gynaecologist 2006;8:181–185
12. Ebel RL. Are true–false items useful? In: Ebel RL (ed). Practical problems in educational measurement. Lexington, MA: D.C. Heath; 1980:145–56
13. Frisbie DA. The multiple true-false item format: A status review. Educational Measurement: Issues and Practice 1992;5(4):21–26.
14. Fowell S, Jolly B. Combining marks, scores and grades: reviewing common practices reveals some bad habits. Med Educ 2000; 34:785±786
15. Haladyna TM, Downing SM, Rodriguez MC. A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. Applied Measurement in Education 2002;15(3):309-34
16. Gardner-Medwin AR. Updating with Confidence: Do your students know what they don't know? Health Informatics 1998;4:45-46

17. McCoubrie P. Improving the fairness of multiple-choice questions: a literature review. Medical Teacher 2004; 26(8):709–712
18. Moss E. Multiple choice questions: their value as an assessment tool. Current Opinion in Anaesthesiology 2001,14:661-666
19. Oosterhof AC, Glasnapp DR. Comparative reliabilities and difficulties of the multiple-choice and true-false formats. J Experimental Educ 1974; 42:62–4.
20. Rees PJ. Do medical students learn from multiple choice examinations? Med Educ1986 ;5–20:123 .
21. Schwartz PL, Crooks TJ, Sein KT. Test–retest reliability of multiple true–false questions in preclinical medical subjects. Medical Education 1986; 20:399-406.
22. Schuwirth LWT, van der Vleuten CPM. ABC of learning and teaching in medicine: Written assessment. BMJ 2003; 22; (326):643-645.
23. Schuwirth LW, van der Vleuten CP. Different written assessment methods: what can be said about their strengths and weaknesses? Med Educ 2004; 38(9): 974-9.
24. Ten Cate Th.J. Comparing reliabilities of true/false and short answer questions in written problem solving tests. In: Scherpbier A, et al. Advances in Medical Education. Springer Netherlands 1997:193–196

۳
آزمون‌های کتبی باز پاسخ

فصل ۹

خانواده سؤالات باز پاسخ

تاریخچه سؤالات باز پاسخ

برخلاف خانواده آزمون‌های «بسته‌پاسخ» که در بخش قبلی مورد بحث قرار گرفت، در آزمون‌های «بازپاسخ»[1] فراگیر خود باید جواب مورد نظر را تدوین کند و ارائه دهد. تا پیش از معرفی آزمون‌های چندگزینه‌ای و سایر اَشکال آزمون‌های بسته‌پاسخ، آزمون‌های بازپاسخ از معدود روش‌های ارزیابی بودند که استفاده می‌شدند. برای سال‌های متمادی، در دانشکده‌های پزشکی سراسر جهان از سؤالات تشریحی به عنوان ابزار اصلی ارزیابی فراگیران استفاده می‌شد اما به‌دنبال مشکلات متعدد حاصل از اجرای این سؤالات، از جمله نارضایتی شدید فراگیران در مورد زمان‌بربودن پاسخگویی به این سؤالات، پایایی پایین نمره‌دهی در بین ارزیابان مختلف و پوشش کم محتوای دوره درسی به وسیله آن‌ها، طولی نکشید که تقریباً در اواسط قرن نوزدهم با معرفی نسل جدیدی از آزمون‌های کتبی به نام سؤالات چندگزینه‌ای، استفاده از سؤالات چندگزینه‌ای به منظور غلبه بر مشکلات موجود شیوع پیدا کرد. با این حال، دیری نپایید که محدودیت‌های سؤالات چندگزینه‌ای شناسایی شد و مشخص گردید که علی‌رغم مزایای زیاد سؤالات چندگزینه‌ای از قبیل پایایی بالای و امکان گنجاندن تعداد زیادی سؤال در یک آزمون، همچنان نگرانی‌هایی از جمله امکان حدس‌زدن دانشجویان برای انتخاب بهترین گزینه وجود دارد. به طوری که در سال ۱۹۲۸ فردیک کلی[2] در مقاله خود پس از شرح پاره‌ای از محدودیت‌های سؤالات چندگزینه‌ای، طرحی برای بازنگری ارائه نمود که از جمله در آن توصیه شده بود سؤالات چندگزینه‌ای به شکل محدودتری در موسسات آموزشی مورد استفاده قرار گیرند. به این ترتیب مجدداً تمایل به استفاده از سؤالات بازپاسخ افزایش یافت. چندی نگذشت که استفاده از سؤالات کوتاه‌پاسخ به عنوان شکلی از سؤال که مابین سؤال چندگزینه‌ای و تشریحی بود، رواج یافت. با توجه به این امر که هدف از ارزیابی تنها بازشناسی اطلاعات نیست، در طول زمان گرایش به استفاده از سؤالات کوته‌پاسخ در دانشکده‌های پزشکی سراسر دنیا فزونی یافت. به عنوان مثال دانشکده پزشکی اتریش در اصلاح برنامه درسی دوره پزشکی عمومی خود سعی کرده است از سؤالات کوته‌پاسخ و بازپاسخ در کنار سایر ابزارهای اندازه‌گیری به منظور ارزیابی فراگیران استفاده کند (رادمیکرز[3] ۲۰۰۵).

شاید بتوان گفت که یکی از علل مهمی که باعث شد استفاده از سؤالات بازپاسخ در طول سالیان مختلف حفظ گردد، پایین بودن احتمال حدس زدن و بروز تقلب در این نوع از سؤالات است. این موضوع به قدری مهم است که علی‌رغم وجود مزایای بسیار زیاد حاصل از آزمون‌های بسته‌پاسخ، جایگاه استفاده از سؤالات تشریحی و کوته‌پاسخ در فرایند ارزیابی فراگیران همچنان حفظ گردیده است. حتی تلاش‌های بسیاری در راستای رفع محدودیت‌های این دسته از آزمون‌ها و

1. Open-ended question
2. Frederick J. Kelly
3. Rademakers

معرفی شکل جدیدی از سؤالات تشریحی صورت گرفته است. مشکلات موجود در سؤالات تشریحی منجر گردید تا ایده تغییر این سؤالات در قالب سؤالات تشریحی تغییریافته ارائه گردد. تلاش‌های جسته گریخته صورت گرفته برای استفاده مجدد از آزمون‌های بازپاسخ در اواخر دهه ۱۹۶۰ منجر به ارائه شکل جدیدی از آزمون‌های کتبی تحت عنوان آزمون‌های تشریحی تغییریافته گردید. هر چند برای اولین بار، این شکل از سؤالات توسط بورد ممیزی کالج سلطنتی پزشکان عمومی[1] ارائه شد. با این حال، اولین مستندات در خصوص استفاده از این سؤالات در تسهیل یادگیری مبتنی بر مسأله به هوچکین و ناکس[2] نسبت داده می‌شود. آن‌ها برای اولین بار از سؤالات تشریحی تغییریافته به منظور ارزیابی عملکرد پزشکان عمومی استفاده نمودند (هوچکین و ناکس ۱۹۷۵). از زمان معرفی این شکل از سؤالات زمان زیادی نگذشته بود که اقبال عمومی به استفاده از این شکل از آزمون‌ها در برنامه‌های درسی آموزش پزشکی به سرعت فزونی یافت، به نحوی که از سال ۱۹۷۸ به بعد دانشگاه‌های مختلفی از جمله دانشگاه نیوکاسل سعی کردند از این سؤالات در ارزیابی رسمی توانایی استدلال و حل مسأله بالینی دانشجویان سال اول و دوم پزشکی استفاده نمایند (فلتی[3] ۱۹۸۰). به صورت کلی، این نوع از سؤالات یکی از انواع متعدد ابزارهای موجود هستند که به منظور ارزیابی توانایی استدلال و تصمیم‌گیری فراگیران به جای ارزیابی صرف حقایق طراحی شدند.

در جدول ۱-۹ نوع سؤال بازپاسخ که به صورت شایع مورد استفاده قرار می‌گیرند، با خانواده بسته‌پاسخ مقایسه شده‌اند. ویژگی مشترک خانواده بازپاسخ این است که فهرستی از گزینه‌ها در اختیار دانشجو قرار نمی‌دهند. به این ترتیب دانشجو باید جواب را تولید کند و نمی‌تواند با دیدن گزینه‌های موجود، سرنخی از جواب پیدا کند. یکی دیگر از ویژگی‌های آزمون‌های این دسته که باعث تمایز آن‌ها از سایر انواع آزمون‌های بسته‌پاسخ می‌شود، تاثیر نظر شخصی مصحح هنگام ارزیابی و نمره‌دهی به این سؤالات است. همچنین همواره در ذکر تفاوت بین آزمون‌های بازپاسخ و بسته‌پاسخ، این موضوع مطرح می‌شود که از طریق سؤالات تشریحی می‌توان به ارزیابی سایر مهارت‌های فراگیران از قبیل شیوه نگارش، سازمان‌دهی و نحوه ارائه مطالب و حتی خلاقیت و نوآوری فراگیران در ارائه پاسخ پرداخت، در حالی‌که در آزمون‌های بسته‌پاسخ این امکان فراهم نیست. این سؤالات از سنجش سطح یادآوری بالاتر رفته و می‌توانند میزان درک و قدرت حل مساله را در دانشجو مورد سنجش قرار دهند. نکته شایان توجه در این خصوص این است که این ویژگی در بین آزمون‌های مختلف خانواده سؤالات بازپاسخ متغیر است، به گونه‌ای که آزمون‌های کوته‌پاسخ و تشریحی گسترده‌پاسخ به ترتیب از کمترین و بیشترین فرصت برای سنجش خلاقیت و نوآوری برخوردار هستند. البته با این موضوع باید با احتیاط برخورد شود. همواره باید در نظر داشت که علی‌رغم آنکه سؤالات تشریحی، فرصت ارائه پاسخ‌های خلاقانه را برای فراگیران فراهم می‌سازند اما این موضوع هیچ تضمینی برای این نیست که این سؤالات ذاتاً توانایی ارزیابی سطوح بالای حیطه شناختی را دارند. در واقع، همان‌طور که در بخش دوم کتاب اشاره شد، شکل و فرمت سؤال تعیین‌کننده قابلیت سؤال در ارزیابی سطوح مورد سنجش نیست، بلکه محتوای سؤال است که تعیین می‌کند، سؤال مورد نظر چه سطحی از حیطه دانش را مورد ارزیابی قرار می‌دهد. توجه به این امر ضروری است که سنجش سطوح بالای شناختی، با طراحی یک پایه خوب و مناسب که معمولاً به صورت یک مورد بالینی است، امکان‌پذیر می‌شود و تنها بر اساس شکل و فرمت سؤال نمی‌توان به این نتیجه رسید که سؤال طراحی شده بدون شک به ارزیابی سطوح بالای حیطه شناختی خواهد پرداخت. بنابراین ارزیابان در هنگام طراحی سؤالات بازپاسخ باید به این نکته توجه داشته باشند که استفاده از سؤالات بازپاسخ لزوماً به معنای سنجش سطوح بالای شناختی نیست، بلکه نحوه پرسش است که باعث فراهم‌سازی فرصتی جهت ارزیابی سطوح بالای حیطه شناختی می‌شود.

همواره باید به این نکته توجه داشت که علی‌رغم مزایای سؤالات بازپاسخ، استفاده از این سؤالات برای هر موقعیتی

1. The Board of Censors of the Royal College of General Practitioners
2. Hodgkin & Knox
3. Feletti

مناسب نیست و در هنگام تصمیم‌گیری در خصوص انتخاب آزمون‌های کتبی به منظور داشتن آزمونی روا[1] باید به تجانس بین روش ارزیابی با هدف آموزشی موردنظر توجه نمود. زمانی که هدف، سنجش توانایی فراگیر در ارائه پاسخ‌های محتمل موجود یا ارائه پاسخ‌های خلاقانه است، استفاده از سؤالات تشریحی یا کوته‌پاسخ در مقایسه با سؤالات بسته‌پاسخ مناسب‌تر است و اطلاعات دقیق‌تری از عملکرد فراگیر فراهم می‌کند. از این موضوع مهم تحت عنوان اصل هم‌راستایی ساختاری[2] در فرایند یاددهی- یادگیری نام برده می‌شود. بر طبق این اصل انتخاب اهداف آموزشی، محتوا، روش‌های تدریس و در نهایت شیوه ارزیابی باید در یک راستا و منطبق با پیامد مورد انتظار آموزشی باشد. تنها در این شرایط است که می‌توان انتظار داشت که تاثیر آموزشی[3] آزمون‌های به‌عمل آمده در بالاترین سطحِ ممکن باشد، چرا که در این شرایط دانشجویان تنها به سمت یادگیری مطالبی هدایت خواهند یافت که اطمینان دارند به‌وسیله آن مورد ارزیابی قرار می‌گیرند.

جدول ۱-۹: مقایسه سؤالات تشریحی و کوته‌پاسخ یا انواع سؤالات بسته‌پاسخ

ویژگی	سؤال تشریحی	سؤال کوته‌پاسخ	خانواده سؤالات بسته‌پاسخ
نوع پاسخ	تولید پاسخ	تولید پاسخ	انتخاب پاسخ
نمره‌دهی	نمره‌دهی ذهنی	نمره‌دهی نسبتاً عینی	نمره‌دهی بسیار عینی
ساختار پاسخ	غیرساختارمند	نسبتاً ساختارمند	بسیار ساختارمند
ثبات نمره‌دهی	دشواری در ثبات نمره‌دهی	ثبات قابل قبول	ثبات بالا
پایایی بین ارزیابان	پایایی ضعیف	پایایی خوب	پایایی بالا
سطح شناختی قابل سنجش	حل مسأله و خلاقیت	شناسایی و کاربرد	شناسایی و کاربرد

اصل هم‌راستایی ساختاری (برگرفته از گیبس و تانگ ۲۰۱۱)

- هم راستایی ساختاری شکلی از آموزش مبتنی بر پیامد است.
- در هم راستایی ساختاری ارتباط بین میزان یادگیری مورد انتظار، فعالیت‌های آموزشی و یادگیری و ارزیابی از طریق زنجیره‌ای که بین این‌ها وجود دارد برقرار می‌گردد، در حالی‌که در سایر مدل‌های مبتنی بر پیامد، همترازی فقط بین یادگیری مبتنی بر پیامد و ارزیابی به وجود می‌آید.
- هم‌راستایی ساختاری چارچوبی را برای تصمیمات آموزشی و ارزیابی فراهم می‌سازد که میزان دستیابی به پیامدهای مورد انتظار یادگیری بر اساس آن افزایش می‌یابد.
- از طریق اجرای هم‌راستایی ساختاری هماهنگی بین اجزای مختلف برنامه درسی از جمله فعالیت‌های آموزشی و ارزیابی افزایش می یابد.

(برگرفته از گیبس و تانگ ۲۰۱۱)

انواع سؤالات بازپاسخ

با توجه به قدمت بالای استفاده از سؤالات بازپاسخ در دانشکده‌های مختلف سراسر دنیا، طی این مدت تلاش برای رفع پاره‌ای از محدودیت‌های این سؤالات، منجر به ارائه اشکال جدیدتری از آن‌ها شده است. هرچند در حال حاضر، رایج‌ترین نوع سؤالات در خانواده بازپاسخ، سؤالات تشریحی و سؤالات کوته‌پاسخ هستند، اَشکال متنوعی از آنها با نام‌های مختلف در دسترس است. در ادامه به طور مختصر در مورد هر یک از آنها توضیحاتی ارائه می‌دهیم و در فصول بعدی، به ذکر جزئیات بیشتر در خصوص سه نوع از آن‌ها خواهیم پرداخت.

1. Valid
2. Constructive alignment
3. Educational Impact

سؤال تشریحی

سؤالات تشریحی[1] سؤالاتی هستند که در آن‌ها از فراگیر خواسته می‌شود با توجه به مطالب فراگرفته‌شده، پاسخ را تولید و ارائه نماید. این سؤالات را با توجه به آزادی عمل فراگیر در ارائه پاسخ به دو دسته سؤالات گسترده‌پاسخ[2] و محدودپاسخ[3] تقسیم‌بندی کرده‌اند. به صورت کلی، در آزمون‌های گسترده‌پاسخ هیچ محدودیتی برای ارائه پاسخ در نظر گرفته نمی‌شود و به فراگیر این اجازه داده می‌شود تا هر طور که تمایل دارد پاسخ مورد نظر را ارائه دهد. به عبارت دیگر در سؤالات تشریحی گسترده‌پاسخ، جواب سؤالات محدود نیست و خود دانشجو در جواب دادن به سؤال و سازمان‌دهی آن آزادی عمل دارد. استفاده از این سؤالات در صورت طراحی مناسب، معمولاً برای سنجش سطوح بالای حیطه شناختی مناسب است. در حالی‌که در سؤالات تشریحی محدودپاسخ، محدودیت‌هایی برای ارائه پاسخ در نظر گرفته می‌شود. با توجه به زمان‌بر بودن پاسخ دادن به این سؤالات برای دانشجو و همچنین دشواری تصحیح برای استادان، در یک جلسه امتحان، تنها می‌توان تعداد محدودی از این نوع سؤال طرح کرد که موجب می‌شود مشکل مربوط به ویژگی محتوا[4] رخ دهد.

نمونه‌ای از سؤال تشریحی گسترده‌پاسخ

با توجه این موضوع که یکی از علل اصلی بروز فشارخون بالا، بیماری کلیوی است، مکانیسم بروز فشارخون بالا را در این حالت شرح دهید؟

نمونه‌ای از سؤال تشریحی محدودپاسخ

علت تفاوت در میزان بقای نوزادان با ناهنجاریهای آمفالوسل و گاستروشیزی را توضیح دهید؟ (ذکر دو دلیل کافی است)

سؤال تشریحی تغییریافته

سؤالات تشریحی تغییریافته[5] انواعی از سؤالات تشریحی هستند که برای سنجش توانایی حل مسائل به کار می‌روند. در این شکل از سؤالات، پرسش به شکل یک سناریو یا مورد بالینی مطرح می‌شود که معمولاً مرحله به مرحله معرفی می‌گردد و در هر مرحله از دانشجو خواسته می‌شود تا به سؤال یا سؤالاتی که از جنبه‌های مختلف به مورد بالینی پرداخته‌اند، پاسخ دهد. طرح تنه سؤال به صورت سناریوی بالینی منجر به تشویق فراگیران به یادگیری موضوعات به شکل کاربردی می‌شود. بدین ترتیب از یادگیری مطالب به شکل طوطی‌وار ممانعت می‌شود. این شکل از سؤالات، امکان ارزیابی توانایی حل مسأله و نیز درک فراگیران از ارتباط بین علوم پایه و بالینی را فراهم می‌سازند.

نمونه‌ای از سؤال تشریحی تغییریافته

آقای ۵۲ ساله با درد ناحیه تحتانی شکم به اورژانس مراجعه کرده است. با گرفتن شرح حال، متوجه می‌شوید که درد او از ۸۴ ساعت پیش شروع شده، در ابتدا نامنظم، ناپیوسته، مبهم و حوالی ناف بوده و به تدریج که به سمت ربع تحتانی راست متمرکز شده، ماهیت دائمی پیدا کرده است. شما به تشخیص آپاندیسیت حاد فکر می‌کنید.

۱. توضیح دهید که چرا درد ابتدا در اطراف ناف احساس شده است (۱ نمره)؟
۲. با توجه به بافت شناسی آپاندیس توضیح دهید که چرا درد در ابتدا نامنظم و ناپیوسته بوده است (۱ نمره)؟
۳. به جز روده، سه ساختار را در شکم نام ببرید که ممکن است درد با چنین ماهیتی ایجاد کنند (۰/۵۷ نمره).

شما بیمار را معاینه می‌کنید. سه نشانه‌ای که در لمس شکم او پیدا کرده‌اید، تندرنس مک بورنی، گاردینگ و ریباند تندرنس است.

۴. با رسم شکل، نقطه مک بورنی را نشان دهید (۰/۵ نمره).

شما برای بیمار مشاوره جراحی می‌دهید. جراح پس از ویزیت از شما می‌خواهد که بیمار را برای اتاق عمل آماده کنید. اما بیمار از پذیرفتن جراحی امتناع می‌کند و شما باید عواقب این تصمیم را برای او توضیح دهید.

۵. سه عارضه‌ای که در صورت عدم جراحی به موقع ممکن است برای بیمار رخ دهد، نام ببرید (۰/۵۷ نمره).
۶. چه تغییراتی در بافت آپاندیس ملتهب رخ می‌دهد که منجر به عوارض فوق می‌شود (۱ نمره)؟

1. Essay
2. Extended response
3. Restricted response
4. Content specificity
5. Modified Essay Question (MEQ)

سؤال کوته‌پاسخ

در سؤال کوته‌پاسخ[1] همان‌طور که از نامش پیداست، از دانشجو خواسته می‌شود پاسخ خود را در حد یک کلمه یا یک عبارت کوتاه و مختصر ارائه دهد. پاسخ ارائه شده در آزمون‌های کوته‌پاسخ آنقدر مختصر است که این آزمون‌ها را به کلی از آزمون‌های تشریحی مجزا می‌سازد و به آزمون‌های بسته‌پاسخ نزدیک می‌کند. با توجه به اینکه تعداد سؤالاتی از این نوع که می‌توان در یک جلسه امتحان از دانشجویان پرسید نسبتاً خوب است، نگرانی در خصوص ویژگی محتوا در این سؤالات در مقایسه با سؤالات تشریحی کمتر است. در منابع مختلف، دسته‌بندی‌های متفاوتی از این سؤالات ارائه شده است، اما به‌طور کلی آنها را به چهار دسته «پرسشی»، «تکمیلی»، «هدایتی» و «تشخیصی» طبقه‌بندی می‌کنند. نمونه سؤال ارائه شده در زیر مثالی از سؤالات کوته‌پاسخ نوع پرسشی است.

نمونه‌ای از سؤال کوته‌پاسخ

از اندازه‌گیری فشار وج ریه (PCWP) برای تخمین فشار در کدام ناحیه استفاده می‌شود؟

سایر انواع آزمون‌هایی که در ادامه به شرح آنها خواهیم پرداخت، برخلاف انواعی که در بالا به آن‌ها اشاره شد، معمولاً چندان متداول نیستند. به‌همین دلیل به شرح مختصر آنها اکتفا می‌کنیم و از ذکر جزئیات بیشتر در فصول مجزا برای هر یک از آن‌ها خودداری می‌کنیم.

سؤال یادآوری آزاد

سؤال یادآوری آزاد[2] شکلی از خانواده بازپاسخ است که در آن از فراگیر خواسته می‌شود هر آنچه را که در خصوص یک موضوع به یاد می‌آورد، یادداشت کند. از این سؤالات معمولاً برای تعیین میزان دانسته‌های فراگیران قبل از شروع آموزش (پیش‌آزمون) یا سنجش توانایی به خاطرسپاری و قدرت حافظه فراگیران در یادآوری مطالب بلافاصله بعد از آموزش استفاده می‌شود. پاسخ به این سؤالات بیشتر حالت انشایی داشته و استادان از طریق بررسی پاسخ این سؤالات، میزان توجه و یادگیری فراگیران را به مبحث ارائه شده مورد سنجش قرار می‌دهند. از این سؤالات نمی‌توان برای تصمیم‌گیری در خصوص وضعیت تحصیلی فراگیران استفاده نمود.

نمونه‌ای از سؤال یادآوری آزاد

با توجه به مطالب بیان شده در کلاس، هر آنچه را که در خصوص فرایند انعقاد خون به خاطر دارید، شرح دهید.

شبیه‌سازهای نوشتاری

شبیه‌سازهای نوشتاری[3] موارد بالینی را گام به گام و به صورت نوشتاری ارائه می‌کنند و اغلب به منظور بررسی مهارت پزشکان در تصمیم‌گیری بالینی مورد استفاده قرار می‌گیرند. این نوع از آزمون‌ها بر پایه این پیش‌فرض استوار هستند که پاسخ‌های نوشتاری پزشکان به موارد بالینی، بسیار به عملکردی که آنان در محیط واقعی انجام می‌دهند، نزدیک است. با این حال همچنان این ادعا در حد فرضیه باقی مانده است و مطالعات کمی به بررسی آن پرداخته‌اند. در مرور نظامند صورت گرفته توسط جونز و همکاران[4] در سال ۱۹۹۰ مشخص شد از حدود ۷۴ مقاله چاپ شده در خصوص شبیه‌سازهای

1. Short answer question
2. Free recall test
3. Written simulations
4. Jones et al.

نوشتاری تنها ۱۱ مورد (۱۵ درصد) روایی محتوایی این نوع از سؤالات را مورد ارزیابی قرار داده‌اند. نویسندگان همچنین ذکر کردند که بین این ۱۱ مقاله در مورد اینکه چگونه پاسخ‌های کتبی افراد می‌تواند معرف عملکرد واقعی آنها باشد، توافق نظر وجود ندارد.

نمونه‌ای از سؤال شبیه ساز نوشتاری

خانم ۵۳ ساله به علت داشتن سابقه خانوادگی دیابت، آزمایشات روتین انجام داده است که به قرار زیر است:
FBS=140mg/dl 2hpp=200mg/d TG = 220 Cholesterol = 210 HDL = 46 LDL= 130
در معاینه قد بیمار ۱۵۵ سانتیمتر و وزن ۷۵ کیلوگرم بدست آمد. سابقه PCOS دارد. با توجه به اطلاعات ارائه شده، در هر مرحله پاسخ صحیح را ارائه نمایید.
الف) اولین اقدام درمانی شما چیست؟
ب) در صورت عدم پاسخ به درمان قدم دوم درمان شما چیست؟
ج) اعداد هدف کنترل قند خون بیمار چیست؟
FBS: 2hpp: HBA:
د) اعداد هدف کنترل لیپید در این بیمار چیست و با چه اقدامات درمانی به آن می رسید؟

پروژه

یکی دیگر از روش‌های ارزیابی کتبی در قالب ارائه پروژه[1] تحقیقاتی یا کلاسی است. در این حالت فعالیت‌های فراگیر در خصوص موضوع مورد نظر در قالب یک گزارش کتبی به استاد ارائه می‌گردد. در فرهنگ لغات آکسفورد[2] پروژه به عنوان تکلیفی در دانشگاه یا مدرسه که برای آن باید تحقیق انجام شود و در نهایت گزارش آن به صورت متنی نگاشته شود، تعریف شده است. به طور کلی، پروژه به مجموعه فعالیت‌هایی که به صورت پی در پی و هدف‌دار برای رسیدن به یک هدف از پیش تعیین شده انجام می‌پذیرد گفته می‌شود. هرچند برخی از مستندات، از پروژه به عنوان یک روش ارزیابی کتبی نام برده‌اند، در این خصوص در بین صاحب نظران مختلف توافق نظر وجود ندارد و برخی این شیوه ارزیابی را از آزمون‌های تشریحی و مبتنی بر عملکرد مجزا نموده‌اند (استچر و همکاران[3] ۱۹۹۷).

مزایا و محدودیت‌های سؤالات باز پاسخ

مزایای سؤالات بازپاسخ

همان‌طور که پیشتر نیز اشاره شد، وجود برخی از نقاط قوت در سؤالات بازپاسخ باعث گردیده است تا در حال حاضر نیز در بسیاری از دانشکده‌های پزشکی سراسر دنیا، در کنار سایر روش‌های ارزیابی، همچنان از این سؤالات استفاده شود. در ادامه به برخی از این مزایا اشاره شده است.

- **طراحی آسان:** یکی از مشکلات مربوط به سؤالات بسته‌پاسخ به ویژه سؤالات چندگزینه‌ای این است که طراحی سؤالاتی که به ارزیابی سطوح بالای حیطه شناختی بپردازند و در عین حال فاقد خطا باشند نسبتاً دشوار است. در سؤالات بازپاسخ، به دلیل عدم نیاز به طراحی گزینه‌های انحرافی طراحی سؤالات آسان‌تر خواهد بود این موضوع باعث شده است تا اعضای هیأت علمی به طرح سؤالات بازپاسخ به دلیل طراحی ساده آنان متمایل شوند. به‌علاوه به دلیل آشنایی بیشتر اعضای هیأت علمی با این سؤالات از دیرباز، طراحی آن‌ها اغلب در محدوده زمانی کوتاه‌تری صورت می‌گیرد.

1. Project
2. Oxford dictionary
3. Stecher et al.

- **نبود مشکل بازشناسی[1]:** در مقایسه با سؤالات چندگزینه‌ای، مشکل بازشناسی در این سؤالات وجود ندارد زیرا لیستی از پاسخ‌های احتمالی در اختیار آزمون‌شوندگان قرار نمی‌گیرد.
- **نبود احتمال حدس زدن:** در این سؤالات امکان حدس زدن و تقلب در مقایسه با سؤالات بسته‌پاسخ بسیار کمتر است. اغلب اعضای هیأت علمی در هنگام نگارش سؤالات چندگزینه‌ای دچار خطاهایی می‌شوند که باعث هدایت فراگیران به سمت پاسخ درست می‌شود. این موضوع به قدری تاثیرگذار است که حتی پس از مدتی، فراگیران از طریق کشف سرنخ‌های موجود در سؤال چندگزینه‌ای، می‌توانند پاسخ صحیح سؤالات را حدس بزنند (برای کسب اطلاعات بیشتر به بخش دوم کتاب مراجعه کنید). البته باید در نظر داشت که امکان حدس زدن و تقلب در سؤالات کوته‌پاسخ اندکی بیشتر از سؤالات تشریحی است و علاوه بر آن، سؤالات تشریحی زمینه را برای شکل دیگری از حدس زدن فراهم می‌کنند که در فصل دوم این بخش مورد اشاره قرار خواهد گرفت.
- **امکان ارزیابی سطوح بالای حیطه شناختی و خلاقیت:** پاسخ‌گویی به سؤالات تشریحی نیازمند استفاده از فرایندهای فکری و استدلال است. تعدادی از محققان معتقدند که سؤالات بازپاسخ مخصوصاً تشریحی، امکان ارزیابی پیامدهای پیچیده یادگیری را که نمی‌تواند به شکل موثر، به‌وسیله سایر آزمون‌های کتبی مورد سنجش قرار گیرد، فراهم می‌سازند (بریگمن و لویس[2] ۱۹۹۴، اسکولر[3] ۱۹۹۸، اسکولر و پروسر[4] ۱۹۹۴، والستد و بیکر[5] ۱۹۹۴). طراحی سؤالات بازپاسخی که به ارزیابی سطوح بالای حیطه شناختی بپردازند، به دلیل تجربه و آشنایی بیشتر اعضای هیأت علمی با این سؤالات آسان‌تر است. از طرف دیگر امکان ارزیابی توانایی تفسیر و سازمان‌دهی مطالب به شکل خلاقانه و با استفاده از مثال‌های متعدد از این طریق امکان‌پذیر است. با این حال، بر خلاف تصور رایجی که وجود دارد، سؤالات بازپاسخ ذاتاً توانایی ارزیابی سطوح بالای حیطه شناختی را ندارند، بلکه تنها در صورت طراحی صحیح و انتخاب محتوای مناسب است که می‌توان انتظار داشت این سؤالات به ارزیابی سطوح بالای حیطه شناختی بپردازند. در واقع در سؤالات بازپاسخ امکان ارزیابی فرایند تولید پاسخ از قبیل توانایی تلفیق ایده‌ها و سازمان‌دهی و ترکیب مطالب نیز فراهم است. بنابراین، می‌توان گفت که به صورت بالقوه، امکان ارزیابی خلاقیت، تفسیر و سازمان‌دهی مطالب در آزمون‌های بازپاسخ مهیا است. همچنین در سؤالات تشریحی تغییریافته، امکان ارزیابی فرایند استدلال و حل مسأله فراگیران به شکل بهتری وجود دارد. استفاده از سناریوی بالینی در این سؤالات کمک می‌کند تا سطوح بالای شناختی مورد ارزیابی قرار گیرند زیرا از دانشجو خواسته می‌شود تا معلومات و محفوظات خود را در بستر یک موقعیت نزدیک به واقعیت به کار برد. البته باید توجه داشت که در برخی از مواقع ممکن است تعدادی از طراحان سؤال که تجربه زیادی در طرح سؤالات تشریحی تغییریافته ندارند، صورت سؤال را به گونه‌ای تنظیم نمایند که هیچ ارتباطی با سؤالاتی که در ادامه مطرح می‌شود، نداشته باشد. در این حالت، معمولاً سؤالات طراحی شده تنها به ارزیابی سطوح پایین حیطه شناختی می‌پردازند، به نحوی که آزمون شوندگان حتی بدون مطالعه صورت سؤال می‌توانند از طریق یادآوری حقایق و دانسته‌های علمی به سؤالات مطرح شده پاسخ دهند. در خصوص سؤالات کوته‌پاسخ نیز باید گفت که امکان سنجش قدرت سازمان‌دهی و خلاقیت فراگیران و بررسی هدف‌های سطح بالای شناختی مانند ترکیب و تصمیم‌گیری به وسیله این سؤالات میسر نیست. استفاده از این سؤالات تنها در ارزیابی موضوعاتی مناسب است که پاسخ آن محدود است. این امر باعث شده است که در عمل از سؤالات کوته‌پاسخ به منظور ارزیابی سطوح پایین حیطه شناختی استفاده شود که ممکن است فراگیران را به یادگیری طوطی‌وار مطالب تشویق نماید. در حالی که به وسیله سؤالات کوته‌پاسخ امکان ارزیابی اهداف تا سطح کاربرد حیطه شناختی وجود دارد و نباید این سؤالات را تنها به ارزیابی دانسته‌ها و حقایق فراگیران محدود کرد.

1. Recognition
2. Bridgeman & Lewis
3. Scouller
4. Scouller & Prosser
5. Walstad & Becker

- **تشویق به یادگیری عمیق:** یکی از مزایای بالقوه سؤالات بازپاسخ، تشویق فراگیران به یادگیری عمیق مطالب است. همان‌طور که در بالا نیز اشاره شد، این سؤالات امکان ارزیابی توانایی سازمان‌دهی بین موضوعات مختلف را فراهم می‌سازند، لذا فراگیر برای پاسخ‌دهی به این سؤالات نیازمند درک عمیق مطالب است. این موضوع باعث گردیده است تا ارزیابان از این سؤالات بیشتر در طی دوره آموزشی و به شکل تکوینی استفاده نمایند. به عبارت دیگر، یکی از موارد کاربرد سؤالات تشریحی و کوته‌پاسخ در مواردی است که هدف از ارزیابی، هدایت فراگیران به یادگیری عمیق‌تر مطالب است. در این شرایط فراگیران به منظور پاسخ‌دهی صحیح به سؤالات نیازمند آن هستند که به خوبی مطالب مطالعه شده را درک کنند و در نهایت با برقراری ارتباط بین مطالب مختلف فرا گرفته شده به ارائه پاسخ‌های منحصر به فرد و خلاقانه بپردازند. از این رو در مواردی که هدف آموزشی به دنبال تقویت قدرت تفکر فراگیران در ارائه و سازمان‌دهی پاسخ و فراهم‌سازی فرصت یادگیری عمیق‌تر است، استفاده از این سؤالات به شکل تکوینی توصیه می‌شود. به‌علاوه، استفاده از سؤالات بازپاسخ به شکل تکوینی باعث می‌شود تا مشکل ویژگی محتوا که در سؤالات تشریحی مطرح است، تا حدودی از طریق انجام ارزیابی‌های مکرر دوره‌ای برطرف گردد (برای کسب اطلاعات بیشتر در خصوص مشکل اختصاصی بودن محتوا به بخش اول کتاب مراجعه کنید).
- **امکان ارزیابی تکوینی و تراکمی:** همان‌طور که پیشتر نیز اشاره شد، یکی از اهداف اصلی به‌کارگیری سؤالات بازپاسخ، تشویق فراگیران به یادگیری عمیق‌تر مطالب است. لذا در مواقعی که هدف آزمون تقویت میزان یادگیری فراگیران باشد، از این سؤالات مخصوصاً نوع کوته‌پاسخ به شکل تکوینی در طول دوره آموزشی می‌توان استفاده کرد. با وجود این استفاده از این سؤالات در قالب آزمون پایان دوره متداول است.
- **مطابقت با شرایط واقعی:** در سؤالات بازپاسخ امکان طراحی سؤالات مشابه با موقعیت‌های نسبتاً واقعی فراهم است. در واقع، در این سؤالات استفاده از سناریوی بالینی کمک می‌کند سطوح شناختی بالا مورد ارزیابی قرار گیرند، زیرا از دانشجو خواسته می‌شود تا معلومات و محفوظات خود را در بستر یک موقعیت نزدیک به واقعیت به کار برد. از این رو در انواع مختلف سؤالات بازپاسخ به ویژه سؤالات تشریحی تغییریافته، تنه سؤال به شکل سناریو و یا مورد بالینی مطرح می‌گردد. به‌علاوه در سؤالات بازپاسخ، فراگیران خود باید پاسخ‌های محتمل موجود را ارائه نمایند که این موضوع نیز با شرایط واقعی شباهت بیشتری دارد، و این در حالی است که در سؤالات بسته‌پاسخ، فهرست پاسخ‌های احتمالی در اختیار آزمودنی قرار داده می‌شود و او کافی است از بین مجموعه پاسخ‌های ارائه شده یکی را انتخاب نماید. لذا بدیهی است که سؤالات تشریحی از این جهت با شرایط واقعی که فرد تجربه می‌کند، همخوانی بیشتری دارد.

محدودیت‌های سؤالات بازپاسخ

علی‌رغم مزایای سؤالات بازپاسخ، همواره این سؤالات با پاره‌ای از محدودیت‌ها از سوی منتقدان مواجه هستند که در زیر به مهم‌ترین معایب به صورت کلی اشاره می‌شود.

- **محدودیت در تنوع کاربردها[1]:** سؤالات بازپاسخ صرفاً امکان ارزیابی سطوح مختلف حیطه دانشی را فراهم می‌سازند. بنابراین در شرایطی که هدف از ارزیابی سنجش هدف‌های حیطه حرکتی و یا نگرشی است، استفاده از این سؤالات به تنهایی کفایت نمی‌کند.
- **زمان‌بر بودن پاسخ‌دهی:** برای پاسخ‌دهی به این سؤالات به ویژه سؤالات تشریحی گسترده‌پاسخ به زمان زیادی نیاز است، این موضوع باعث گردیده است تا تعداد معدودی از اهداف و محتوای آموزشی را بتوان در هر بار اجرای آزمون ارزیابی نمود. البته در مقایسه با سؤالات تشریحی گسترده‌پاسخ، کوتاه بودن زمان پاسخ دهی به سؤالات کوته‌پاسخ منجر شده است تا تعداد بیشتری از این سؤالات را بتوان در هر بار اجرای آزمون گنجاند.

1. Versatility

- **روایی پایین:** زمان‌بر بودن پاسخ‌دهی به این سؤالات و محدودیت تعداد سؤالات مورد ارزیابی بر پوشش محتوا تاثیر می‌گذارد و موجب کاهش روایی محتوایی آزمون می‌شود. البته سؤالات کوته‌پاسخ، محتوای دوره را بهتر از سؤالات تشریحی پوشش می‌دهند.
- **مشکل ویژگی محتوا:** همان‌طور که گفته شد، مشکل نمونه‌گیری محدود این سؤالات یک مشکل جدی است که منجر به کاربرد کمتر آن‌ها شده است. در نتیجه لازم است به هنگام کاربرد این آزمون‌ها، تا حد امکان از سؤالات تشریحی محدودپاسخ به جای سؤالات تشریحی گسترده‌پاسخ استفاده کرد، به این طریق امکان طرح تعداد بیشتری از آنها در یک جلسه امتحانی فراهم می‌شود. به‌علاوه می‌توان به منظور رفع این مشکل چندین نوبت امتحان در طول دوره به عمل آورد تا به این طریق نمونه بزرگتری از محتوا و هدف‌های آموزشی سنجش شوند.
- **تاثیر قضاوت ذهنی در نمره‌دهی و عینیت پایین:** اگر ساختار و معیارهای نمره‌دهی مشخص در پاسخ‌های تشریحی طولانی وجود نداشته باشد، ذهنیت مصححان بر فرایند تصحیح اثرگذار خواهد بود و منجر به بروز تفاوت چشمگیر در نمرات و پایایی پایین نمرات می‌شود. البته سؤالات کوته‌پاسخ تا حدودی به دلیل کوتاه بودن پاسخ مورد نظر در مقایسه با انواع سؤالات تشریحی از عینیت بالاتری برخوردارند.
- **میزان پایین پایایی، توافق و ثبات در نمره‌دهی:** در سؤالات بازپاسخ مخصوصاً تشریحی منابع خطا متعدد هستند و عواملی به جز دانش فراگیر از قبیل سبک نگارش، دست‌خط و شیوه سازمان‌دهی و بیان مطلب نیز در فرایند نمره‌دهی تاثیرگذار خواهند بود. مجموعه موارد ذکر شده می‌توانند در نهایت منجر به پایایی پایین‌تر این سؤالات در مقایسه با سؤالات بسته‌پاسخ شود. از جمله انتقاداتی که پیوسته به سؤالات تشریحی وارد می‌شود این است که اگر برگه یک دانشجو توسط دو فرد مختلف تصحیح شود، نمرات ارائه شده تفاوت دارد که که این مسأله تحت عنوان پایین بودن پایایی بین ارزیابان[1] مطرح می‌گردد. علاوه بر پایین بودن پایایی بین مصححان، احتمال ارائه نمرات متفاوت در سؤالات تشریحی در صورتی که یک سؤال توسط یک مصحح در زمان‌های مختلف نیز ارزیابی شود، وجود دارد که این مشکل تحت عنوان پایین بودن پایایی درون ارزیاب[2] مطرح می‌گردد. بر اساس نتایج حاصل از مطالعه پاورز و همکاران[3] مشخص گردید که پاسخ‌های دست‌نویس فراگیران معمولاً نمره بالاتری را نسبت به پاسخ‌های مشابه تایپ شده دریافت می‌کنند. به عبارت دیگر در سؤالات تشریحی دست‌نویس، مصححان عمدتاً بخشنده‌تر عمل می‌کنند. به اعتقاد این محققان، علت بروز این امر از یک سو، نادیده گرفتن خطاهای دستوری و املایی در پاسخ‌های دستی است و از طرف دیگر به دلیل آن است که پاسخ‌های تایپ شده فضای کمتری را در برگه امتحانی می‌گیرند و ممکن است ناقص به نظر برسند (پاورز و همکاران ۱۹۹۴).

چیس [1] ۱۹۹۰

چیس در پژوهشی به بررسی ارتباط بین انتظارات معلمان از فراگیران با کیفیت دست خط‌های آنان پرداخت. در این مطالعه دو مجموعه یکسان از پاسخ‌های تشریحی تهیه شد که یکی از آن‌ها با دست خط خوب و دیگری با دست خط ضعیف‌تر نگاشته شده بود. پیش از تصحیح برگه‌های پاسخنامه، به مصححان اطلاعاتی در خصوص ویژگی‌های آزمون شوندگان ارائه شد که هدف از این کار افزایش انتظار ارزیابان از عملکرد آنان بود. نتایج حاصل از مطالعه نشان داد که پاسخ‌های ارائه شده با دست خط ضعیف نمره بالاتری را در مقایسه با پاسخ‌های ارائه شده با دست خط خوب دریافت کردند. وی بر اساس نتایج حاصل از مطالعه خود پیشنهاد می‌کند که از نظر مصححان، انتظار کسب موفقیت در آزمون در کسانی که دست خط ضعیف دارند، بیش از آن‌هایی است که دست‌خط خوب دارند.

1. Chase

1. Interrater reliability
2. Intrarater reliability
3. Powers et al

دی و همکاران [۲] ۱۹۹۰

این محققان در مطالعه‌شان خوانا بودن دست‌خط دستیازان سال اول و کارورزان پزشکی را با استفاده از مقیاس لیکرت سه تایی در هر دو روش نمره‌دهی «کلی» و «تحلیلی» ارزیابی کردند. بر اساس نتایج حاصل از این مطالعه، همبستگی بین نمرات با دست‌خط آزمون‌شوندگان در روش تحلیلی برابر با ۰/۱۲ (p=0.12) و در روش کلی برابر با ۰/۸۱ (p=0.3) بود. هر چند نتایج حاصل از آماری معنی دار نبود اما نتایج نشان‌دهنده تاثیر دست خط در روش نمره‌دهی کلی بود. به عبارت دیگر، مصححان در روش کلی تحت تاثیر دست خط فراگیران قرار گرفتند.

2. Day et al

- **دشواری تصحیح:** یکی از بزرگترین محدودیت‌های سؤالات تشریحی تغییریافته، مشکل تصحیح دستی برگه‌های امتحانی است. این موضوع غیر از صرف انرژی و زمان زیاد برای تصحیح، نیازمند دقت بالای مصححان نیز است. به منظور افزایش پایایی فرایند نمره‌دهی بهتر است پیش از اجرای آزمون مصححان در خصوص پاسخ‌های احتمالی مورد نظر به توافق نظر برسند. نبود الگوی پاسخ‌دهی، منجر به دشواری تصحیح برگه‌های امتحانی می‌شود. برای کاهش میزان دشواری تصحیح سؤالات بازپاسخ و همچنین به حداکثر رساندن میزان عینیت این سؤالات، باید سعی شود تا از روبریک پاسخ‌دهی از پیش تعیین شده استفاده شود. البته بدیهی است که این موضوع در سؤالات کوته‌پاسخ و تا حدودی در سؤالات تشریحی محدودپاسخ قابل اجرا است و در سؤالات تشریحی گسترده‌پاسخ استفاده از کلید پاسخ چندان عملی نیست. به‌علاوه یکی از ویژگی‌های اصلی سؤالات بازپاسخ این است که تصحیح این سؤالات وابسته به فرد است. به عبارت دیگر برخلاف سؤالات چندگزینه‌ای که در آنها تصحیح سؤالات با استفاده از رایانه به راحتی امکان‌پذیر است، در این سؤالات یک نفر باید پاسخ‌ها را خوانده و پس از مقایسه با راهنماهای نمره‌دهی و روبریک پاسخ‌دهی مشخص تصمیم بگیرد که آیا پاسخ ارائه شده قابل قبول است یا خیر. از طرف دیگر در این سؤالات هر چه پاسخ‌ها پیچیده‌تر می‌شوند، قضاوت در خصوص نمره‌دهی نیز دشوارتر می‌شود که نیازمند صرف فعالیت ذهنی بیشتر ارزیابان در فرایند تصحیح سؤالات است.
- **قابلیت اجرای پایین:** به طور کلی برای داشتن پوشش محتوایی وسیع، باید از تعداد زیادی از سؤالات تشریحی استفاده کرد. انجام این کار باعث می‌شود که آزمون غیر کاربردی شده و اداره و تصحیح آن بسیار دشوار شود. این مسأله به ویژه هنگامی که تعداد فراگیران زیاد باشد و یا محتوا و اهداف دوره آموزشی زیاد باشد، بغرنج‌تر می‌شود. مسأله دیگر اینکه اجرای آزمون‌های تشریحی تغییریافته تا اندازه‌ای دشوارتر است. در ابتدای دفترچه این نوع آزمون باید نکاتی را که از دانشجو انتظار انجام آن می‌رود، به شکل شفاف و دقیق شرح داد. گاهی اوقات شرح این موارد به دلیل عدم آشنایی فراگیران با این نوع از سؤالات باعث گیج شدن و منحرف شدن آن‌ها از ارائه پاسخ صحیح می‌شود. این مشکل ممکن است در فرایند اجرا نیز باعث سر در گم شدن مراقبان آزمون شود.

به طور خلاصه می‌توان گفت که سؤالات بازپاسخ مزایای زیادی دارند که این موضوع بالتبع بر کاربرد آن تاثیرگذار است، اما باید در نظر داشت که برای هر موقعیتی مناسب نیستند و در انتخاب این سؤالات تطابق با اهداف دوره و ملاحظات اجرایی باید همواره در نظرگرفته شود. در جدول شماره ۲–۹ مزایا و محدودیت‌های سؤالات بازپاسخ به صورت خلاصه آورده شده است.

جدول ۲-۹: خلاصه نکات مثبت و منفی آزمون‌های کتبی بازپاسخ

نکات مثبت	نکات منفی
امکان ارزیابی خلاقیت	دشواری تصحیح و نمره‌دهی
نبود احتمال حدس زدن	عینیت و پایایی پایین
احتمال تشویق فراگیران به یادگیری عمیق مطالب	قابلیت اجرای پایین (در مواقعی که تعداد فراگیران زیاد است)
امکان پرسش از موقعیت‌های بالینی و تطابق با شرایط واقعی	روایی محتوایی پایین (اگر به عنوان تنها ابزار ارزیابی استفاده شود)
امکان سنجش سطوح بالای حیطه شناختی	احتمال تاثیر ذهنی ارزیاب در نمره دهی
طراحی نسبتاً آسان	مقبولیت پایین از سوی فراگیران

سودمندی سؤالات بازپاسخ

همان‌طور که در بخش اول کتاب اشاره شد، ابزارهای ارزیابی ویژگی‌های ذاتی و ثابتی ندارند که بتوان در مورد کاربرد آنان در موقعیت‌های مختلف نسخه واحد تجویز کرد، بلکه در هر شرایطی باید تصمیم گرفت که از کدامیک بهتر می‌توان استفاده کرد. بر همین اساس برای ارزیابی سودمندی[1] یک ابزار ارزیابی همواره باید فاکتورهای متعددی را در ذهن داشت. بر اساس فرمول ون‌درولوتن[2] در نظر داشتن معیارهای روایی[3]، پایایی[4]، تاثیر آموزشی[5]، هزینه[6] و مقبولیت[7] در انتخاب ابزار کارآمد بسیار کمک کننده است. اگرچه ممکن است که در هر آزمون، با توجه به اهدافی که دنبال می‌شود، سطح اهمیتی که آن آزمون دارد و حتی موقعیت مورد ارزیابی و ویژگی‌های آزمون‌شوندگان یکی از این معیارها بیشتر مورد توجه قرار گیرد اما در هر حال، همواره باید این نکته را در ذهن داشت که تمامی این معیارها در ارتباط با یکدیگر باید لحاظ شوند.

روایی سؤالات بازپاسخ

در مورد روایی سؤالات بازپاسخ می‌توان گفت که به دلیل طولانی بودن فرایند پاسخ‌دهی به این سؤالات در مقایسه با سؤالات بسته‌پاسخ، تعداد کمتری از آن‌ها را در هر آزمون می‌توان ارزیابی نمود. از این رو در هر بار اجرای آزمون روایی محتوایی خوبی را نمی‌توان انتظار داشت. با این حال همان‌طور که پیشتر نیز تاکید شد، استفاده از بلوپرینت و انجام آزمون‌های تکوینی مکرر به بهبود روایی محتوایی این آزمون‌ها کمک می‌کند. در مورد روایی معیاری سؤالات بازپاسخ نیز همواره این سؤال مطرح بوده است که آیا بین نمرات حاصل از این نوع آزمون‌ها با سایر انواع آزمون‌های دیگر ارتباط و همبستگی وجود دارد یا خیر. مطالعات انجام شده یافته‌های متناقضی در این مورد ارائه کرده‌اند که در فصول بعدی به تعدادی از مطالعات انجام شده خواهیم پرداخت.

مطالعه کاظمی و احسان پور ۱۳۸۹

این مطالعه به بررسی روایی آزمون‌های دروس تخصصی نظری دانشجویان مامایی دانشگاه علوم پزشکی اصفهان پرداخت. این پژوهش توصیفی که بر روی آزمون‌های هجده درس تخصصی نظری مقطع کارشناسی مامایی انجام شد. ابزار مطالعه چک لیستی بود که بعد از تدوین به وسیله محققین در اختیار کارشناسان آموزش پزشکی قرار گرفت. روایی محتوایی سؤالات بر اساس توان سؤال در سنجش و اندازه‌گیری اهداف آموزشی طرح درس به وسیله دو کارشناس درس تخصصی مربوطه ارزیابی می‌شد.

1. Utility
2. Van der Vleuten
3. Validity
4. Reliability
5. Educational impact
6. Cost
7. Acceptability

در صورت وجود اختلاف نظر بین دو کارشناس، سؤالات در اختیار کارشناس سوم قرار داده می‌شد و نتیجه ارزیابی که بین دو نفر از کارشناسان مشترک بود ثبت می‌شد. نمونه‌گیری از نوع سرشماری بود. سطح سؤالات و سطح اهداف یادگیری مورد انتظار بر اساس طبقه بندی بلوم ارزیابی می‌شد و انطباق آن‌ها با یکدیگر از طریق مقایسه سطح اهداف آموزشی با سطح تک تک سؤالات صورت می‌گرفت. از تقسیم تعداد سؤالاتی که سطح آن‌ها با سطح اهداف منطبق بود بر تعداد کل سؤالات، میزان تطابق سطح سؤال با سطح اهداف آموزشی محاسبه شد. روایی صوری سؤالات نیز بر اساس رعایت اصول طراحی سؤال میلمن ارزیابی شد. ارزیابی ۱۰۱۳ سؤال از مجموع هجده درس نشان داد، ۹۲/۸۳ درصد از سؤالات روایی محتوایی و ۸۰/۵۴ درصد از سؤالات روایی صوری داشتند. آزمون اسپیرمن نشان داد در ۶۱/۱۱ درصد آزمون‌ها بین سطح سؤالات و اهداف آموزشی ارتباط معنادار وجود دارد.

پایایی سؤالات بازپاسخ

همان‌طور که قبلاً اشاره شد عوامل متعددی از قبیل شیوه نمره‌دهی، تعداد سؤالات امتحانی، تجانس بین سؤالات آزمون، شرایط آزمون و ... در پایایی یک آزمون نیز موثر است. در سؤالات بازپاسخ ممکن است مصححان مختلف، پاسخ یک فراگیر به یک سؤال را به صورت یکسان ارزیابی نکنند و این به دلیل تاثیر ذهنیت‌های مختلف آنان در فرایند نمره‌دهی است. هر چند با استفاده از راهکارهایی از قبیل استفاده از الگوی پاسخ[1] و یا آموزش مصححان می‌توان این خطا را تا حد امکان کاهش داد اما در مجموع سؤالات بازپاسخ در مقایسه با سؤالات بسته‌پاسخ از این نظر، از پایایی پایین‌تری برخوردار هستند.

عامل دیگر تاثیرگذار، تعداد سؤالات آزمون است. با توجه به آنکه در سؤالات بازپاسخ تعداد کمتری از این سؤالات را می‌توان در یک آزمون مطرح کرد، لذا در یک زمان مشخص، تعداد سؤالات بازپاسخی که می‌توان ارائه کرد در مقایسه با سؤالات بسته‌پاسخ کمتر است که این موضوع هم بر روایی و هم بر پایایی آزمون اثر می‌گذارد.

تاثیر آموزشی سؤالات بازپاسخ

دانشجویان در آمادگی برای آزمون از رویکردهای متفاوت مطالعه بهره می‌گیرند (توماس[2] ۱۹۸۴). گفته شده است که دانشجویان زمانی که خودشان را برای امتحان چندگزینه‌ای آماده می‌کنند بیشتر از رویکردهای سطحی مطالعه استفاده می‌کنند و زمانی که خود را برای آزمون تشریحی آماده می‌کنند بیشتر از رویکردهای عمیق مطالعه استفاده می‌کنند (دیویدسون[3] ۲۰۰۲). بر این اساس یکی از باورهای متداولی که در بین ارزیابان وجود دارد این است که در سؤالات بازپاسخ به دلیل آنکه فراگیر مجبور است خود پاسخ را ارائه نماید، مطالب با عمق بیشتری مطالعه می‌شود. هر چند شواهد تجربی چندانی برای اثبات این ادعا وجود ندارد. با این حال، به نظر می‌رسد بهترین نتیجه زمانی حاصل می‌شود که بتوان محتوای آموزشی را با اَشکال متعدد آزمون مورد اندازه‌گیری قرار داد (رسائیان ۱۳۸۱).

مقبولیت، هزینه و قابلیت اجرای سؤالات بازپاسخ

به دلیل آشنایی زیاد استادان با طراحی سؤالات بازپاسخ و نیز سهولت طراحی آنها، این نوع از سؤالات همواره از مقبولیت بالایی در بین ارزیابان برخوردار بوده‌اند اما سؤالات بازپاسخ به ویژه سؤالات تشریحی گسترده‌پاسخ، این نوع سؤالات چندان از سوی فراگیران استقبال نمی‌شود.

اجرای یک آزمون از نظر هزینه، از دو بُعد اقتصادی و غیراقتصادی قابل بررسی است. از بُعد هزینه‌های اقتصادی، هزینه آزمون مربوط به طراحی سؤال، اجرای آزمون و تصحیح برگه‌های امتحانی است. باید توجه داشت که امکان استفاده و اجرای یک آزمون به تمامی این عوامل در ارتباط با یکدیگر وابسته است. به عنوان مثال هر چند در یک آزمون بازپاسخ ممکن است هزینه کمتری برای طراحی یک سؤال مورد نیاز باشد، اما برای اینکه روند تصحیح آن‌ها به

1. Model Answer
2. Thomas
3. Davidson

درستی انجام شود، پاسخنامه‌ای نیاز خواهد بود که به تفصیل کلیه پاسخ‌های احتمالی موجود را شامل شود و این موضوع قاعدتاً کار آسانی نیست و نیازمند صرف زمان و انرژی زیادی است. هزینه‌های غیراقتصادی، هزینه‌هایی است که یک سیستم آموزشی ممکن است در صورت اجرای نامناسب یک آزمون متحمل شود که شاید مهم‌ترین آن اعتبار آن موسسه باشد. نکته شایان توجه این است که در بُعد هزینه‌های غیراقتصادی همواره باید به بحث موقعیت و هدف آموزشی مورد اندازه‌گیری توجه داشت. از نظر هزینه‌های غیر اقتصادی یک ابزار ارزیابی می‌تواند برای اندازه‌گیری یک هدف آموزشی دارای هزینه–اثربخشی و برای هدف دیگر نامناسب باشد. به عنوان مثال، هرگز با استفاده از آزمون‌های کتبی از جمله سؤالات چهار گزینه‌ای، تشریحی و ... نمی‌توان به بررسی مهارت‌های عملکردی فراگیران پرداخت اما چنانچه به دنبال ارزیابی توانمندی‌هایی از قبیل خلاقیت، نوآوری و ... هستیم، استفاده از سؤالات بازپاسخ در مقایسه با سؤالات بسته‌پاسخ هزینه‌های کمتری بر سیستم آموزشی تحمیل خواهد کرد.

ویژگی‌های سؤالات بازپاسخ از نقطه نظر قابلیت اجرا در جدول شماره ۳–۹ ارائه شده است.

جدول۳–۹: ویژگی‌های سؤالات بازپاسخ

ویژگی	به ندرت	گاهی اوقات	معمولاً
طراحی آسان			❊
اجرای آسان		❊	
نمره دهی آسان	❊		
شباهت با موقعیت‌های واقعی زندگی			❊
بازده زمانی نمره دهی (نیازمند زمان محدود)	❊		
موثر برای ارزیابی موقعیت‌های پیچیده دانش			❊
مقبولیت از دیدگاه دی نفعان		❊	

منابع

1. Bridgeman B, Lewis C. The relationship of essay and multiple-choice scores with grades in college courses. Journal of Educational Measurement 1994; 31(1): 37-50.
2. Biggs J, Tang C. Teaching for quality learning at university, 4th edition, 2011, Open University press.
3. Chase C. Essay test scoring: Expectancy and handwriting quality. Psychology: A Journal of Human Behavior 1990.
4. Davidson RA. Relationship of study approach and exam performance. Journal of Accounting Education 2002; 20(1): 29–44.
5. Day SC, Norcini J, Disernes D, Cebul R, Schwartz S, Beck LH, Webster GD, Schnabel TG, Elstein A. The validity of an essay test of clinical judgment. Academic Medicine 1990; 65(9): 39-40.
6. Dogan CD, Atmaca S, Yolcu F. The correlation between learning approaches and assessment preferences of eight-Grade Students. Elementary Education On line. 2012; 11(1): 264 -72
7. Feletti, Grahame I. Reliability and validity studies on modified essay questions. Academic Medicine 1980; 55(11): 933-41.
8. Irwin, WG, and J. H. Bamber. An evaluation of a course for undergraduate teaching of general practice. Medical Education 1978; 12(1): 20-25.
9. James PW. The merits of measurement-driven instruction. Phi Delta Kappa 1987; 68(9): 679-682.
10. Jones TV, Gerrity MS, Earp J. Written case simulations: Do they predict physicians' behavior? Journal of clinical epidemiology 1990; 43(8): 805-15.
11. Powers DE., et al. They Think Less of My Handwritten Essay If Others Word Process Theirs? Journal of Educational Measurement 1994; 31(3): 220-233.
12. Rademakers JJDJM, Th J. ten Cate, and P. R. Bär. Progress testing with short answer questions. Medical Teacher 2005; 27(7): 578-582.
13. Scouller K. The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. Higher Education 1998; 35(4): 453-472.
14. Scouller KM., Prosser M. Students' experiences in studying for multiple choice question examinations. Studies in Higher Education 1994; 19(3): 267-279.
15. Stecher BM, Rahn ML, Ruby A, Naomi Alt M, Robyn A. Using Alternative Assessments in Vocational Education. 1997. Rand publishing
16. Thomas PR, Bain JD. Contextual dependence of learning approaches: The effects of assessment. Human Learning 1984; 3: 227-240.
17. Vleuten CP, van Luijk SJ, Beckers HJ. A written test as an alternative to performance testing. Medical Education 1989; 23: 97–107
18. Walstad, William B., and William E. Becker. "Achievement differences on multiple-choice and es-

say tests in economics." The American Economic Review 1994: 193-196.

۱۹. رساییان ن، نخعی س، صادقی قندهاری ن. مقایسع روش های آزمون های چندگزینه ای، صحیح غلط و کوته‌پاسخ در دانشجویان پزشکی، مجله ایرانی آموزش در علوم پزشکی ۱۳۸۱؛ ۵ (۴): ۲۷۱–۲۷۸.

۲۰. صاحب الزمانی م، زیرک آ. راهبردهای مطالعه و یادگیری دانشجویان دانشگاه علوم پزشکی اصفهان و ارتباط آن با سطح ارتباط امتحان. مجله ایرانی آموزش در علوم پزشکی ۱۳۹۰؛ ۱۱ (۱): ۵۸–۶۸.

۲۱. کاظمی ا، احسان پور س. تحلیل سؤالات دروس تخصصی نظری دانشجویان مامایی دانشگاه علوم پزشکی اصفهان، مجله ایرانی آموزش در علوم پزشکی ۱۳۸۹؛ ۱۰ (۵): ۶۴۳–۶۵۰

فصل ۱۰

سؤالات تشریحی

ساختار سؤالات تشریحی

سؤالات تشریحی از دسته سؤالات بازپاسخ هستند که می‌توانند قدرت فهم مطالب، درک عمیق محتوای درس، پردازش پاسخ‌ها و در نهایت طبقات بالای حیطه شناختی را اندازه بگیرند. پاسخ‌دهی به این سؤالات نیازمند استفاده از فرایند فکری عمیق و سیستماتیک است در حالی که در سؤالات کوته‌پاسخ، ارزیابی دانسته‌های آزمون‌شوندگان نهایتاً تا سطح کاربرد حیطه شناختی امکان‌پذیر است. یکی از تعریف‌های رایج در مورد سؤالات تشریحی به این صورت است: «سؤالات تشریحی انواعی از آزمون‌های کتبی هستند که در آن‌ها پاسخ توسط فراگیر تولید می‌شود و معمولاً پاسخ ارائه شده بیش از یک جمله است. ماهیت سؤال به گونه‌ای است که هیچ جواب منحصر به فرد و الگوی پاسخ‌دهی مشخصی برای آن وجود ندارد، طوری که صحت و کیفیت پاسخ ارائه شده تنها می‌تواند به وسیله یک فرد ماهر و متخصص مورد ارزیابی قرار گیرد» (استالنیکر[1] ۱۹۵۱). بنا به این تعریف، یک سؤال تشریحی باید از معیارهای زیر برخوردار باشد:

- **پاسخ به این سؤالات نیازمند تولید و ارائه پاسخ توسط دانشجو است:** برخلاف سؤالات بسته‌پاسخ به ویژه سؤالات چندگزینه‌ای که فراگیران پاسخ صحیح را از میان مجموعه‌ای از گزینه‌ها انتخاب می‌کنند، در سؤالات تشریحی فراگیران تشویق می‌شوند که تا خود هر طور که تمایل دارند پاسخ را سازمان‌دهی کنند و بپرورانند.
- **دانشجو استنباط و درک خود را از سؤال ارائه می‌دهد:** از سؤالات تشریحی معمولاً به منظور سنجش سطوح بالای حیطه شناختی استفاده می‌شود. به منظور پاسخ‌دهی به این سؤالات دانشجو باید دانسته‌ها و اطلاعات خود را به گونه‌ای منطقی در کنار یکدیگر قرار دهد و در نهایت اندیشه‌های خود را به نحو مناسب بر روی کاغذ بیاورد.
- **امکان ارائه پاسخ صحیح به اشکال گوناگون وجود دارد:** در سؤالات تشریحی این اختیار به دانشجو داده می‌شود تا مطالب را به گونه‌ای که به نظر او منطقی است سازمان‌دهی و ارائه کند. این موضوع منجر به ایجاد الگوهای مختلف پاسخ می‌شود.
- **تعیین صحت و کیفیت پاسخ‌های ارائه شده نیازمند انجام قضاوت ذهنی است:** از آنجا که در این نوع سؤالات نمی‌توان هیچ الگوی مشخصی برای پاسخ تعیین کرد، به منظور تصحیح این سؤالات باید از افراد متخصص که به موضوع مورد ارزیابی تسلط کامل دارند، استفاده نمود.

استالنیکر معتقد است در صورتی که یک سؤال تشریحی منطبق با چهار معیار بالا باشد، در این صورت می‌توان اطمینان حاصل کرد که سؤال طرح شده قابلیت ارزیابی قدرت تفکر و سطوح بالای دانسته‌های فراگیران را خواهد داشت. همان‌طور که پیشتر نیز اشاره شد، علاوه بر رعایت نکات تکنیکی بالا باید به انتخاب دقیق محتوای مورد سنجش نیز

1. Stalnaker

توجه داشت. تنها در این صورت می‌توان اطمینان حاصل نمود که سؤال طراحی شده توانایی سنجش سطوح بالای حیطه شناختی را دارد. همچنین توصیه شده است که سؤالات تشریحی به صورت ساختارمند طراحی گردند. مطالعات تجربی نیز نشان داده‌اند که این مسأله بر پایایی آزمون تاثیرگذار است (ورما و همکاران[1] ۱۹۹۷).

ورما و همکاران ۱۹۹۷

در مطالعه‌ای به بررسی پایایی آزمون‌های تشریحی و تاثیر سازمان‌دهی آنان پرداختند. ۶۲ دانشجوی سال آخر MBBS[1] به دو دسته ۳۱تایی تقسیم شدند. گروه الف دانشجویان به وسیله ۵ سؤال تشریحی که از بانک سؤالات آزمون‌های قبلی برداشته شده بود، به مدت ۲ ساعت مورد ارزیابی قرار گرفتند. در حالی که گروه ب همان سؤالات را به شکل ساختارمند دریافت کردند. برگه‌های پاسخ به تعداد مصححان تکثیر و توسط ۷ مصحح ارزیابی شد. نتایج حاصل از مطالعه نشان داد که پراکندگی نمرات در گروه الف به شکل معنی‌داری بیشتر از گروه ب است. همچنین همبستگی بین نمرات فردی ارزیابان در گروه الف پایین‌تر بود. همسانی درونی برای گروه الف ۰/۳۱ ($p > 0.05$) در حالی که برای گروه ب ۰/۶۹ ($p < 0.05$) بود. ضریب توافق بین ارزیابان در فرایند نمره‌دهی برای گروه ب بهتر بود. یافته‌های مطالعه، پیشنهاد کننده این مسأله است که پایایی سؤالات تشریحی سنتی می‌تواند به وسیله ساختارمند کردن آن‌ها بهبود یابد.

1. Bachelor of Medicine/ Bachelor of Surgery (MBBS)

برای روشن شدن مطالب فوق در خصوص ساختارمندی سؤال تشریحی به مثال زیر توجه کنید.

نمونه‌ای از سؤال تشریحی

هفت گام در اتخاذ تصمیمات اخلاق پزشکی را لیست کنید.

با نگاهی اجمالی به مثال فوق مشخص می‌گردد که برای پاسخ به این سؤال دانشجو لازم نیست درک عمیقی از موضوعات مطرح در حوزه اخلاق پزشکی داشته باشد و تنها کافی است که هفت گام اساسی در این زمینه را به خاطر بیاورد. بنابراین سؤال فوق، تنها در سطح یک یادآوری ساده است و کاری که دانشجو باید انجام دهد، خیلی پیچیده نیست. با کمی دقت می‌توان دریافت که سؤال مطرح شده از ویژگی‌های یک سؤال تشریحی موثر برخوردار نیست:

- دانشجو در پاسخ به این سؤال ملزم نیست که حتماً پاسخ خود را در قالب جمله ارائه نماید و می‌تواند پاسخ خود را در قالب عبارت ذکر نماید.
- با توجه به آنکه ممکن است فراگیران مراحل را به خاطر سپرده باشند، احتمالاً تمامی پاسخ‌ها یکسان هستند. بنابراین امکان ارائه پاسخ‌های خلاقانه در این سؤال وجود ندارد.
- پاسخ به این سؤال نیاز به استفاده از فرایندهای فکری پیچیده و پیشرفته ندارد بنابراین ارزیابی و نمره‌دهی به این سؤال می‌تواند توسط یک فرد معمولی که چندان از دانش تخصصی برخوردار نیست نیز انجام گیرد.

انواع سؤالات تشریحی

به طور کلی آزمون‌های تشریحی را بر اساس میزان اختیار و آزادی عمل فراگیر در ارائه پاسخ به دو دسته کلی سؤالات تشریحی «گسترده‌پاسخ» و «محدودپاسخ» تقسیم می‌کنند. در سؤالات تشریحی گسترده‌پاسخ به دانشجو از نظر میزان پاسخ و مدت زمان، اختیار کامل داده می‌شود. در حالی‌که در سؤالات تشریحی محدودپاسخ برای حجم پاسخ ارائه شده به وسیله دانشجو محدودیت‌هایی در نظر گرفته می‌شود.

سؤالات تشریحی گسترده‌پاسخ

دانشجو در پاسخ دادن به این سؤالات کاملاً آزاد است تا هر طور که خود در نظر دارد، پاسخ را تولید و ارائه نماید. در

1. Verma et al

مواردی ممکن است محدودیت‌های زمانی برای ارائه پاسخ منظور گردد اما در مورد متن و مقدار پاسخ هیچ محدودیتی وجود ندارد. در این سؤالات دانشجو باید بتواند با ترکیب اطلاعات و دانسته‌های موجود، به نحوی منطقی و به شکل خلاقانه پاسخ را ارائه دهد. به همین دلیل از سؤالات تشریحی گسترده‌پاسخ عمدتاً با هدف سنجش سطوح بالای حیطه شناختی (ترکیب و ارزشیابی در تاکسونومی بلوم) استفاده می‌شود. همچنین در این سؤالات بیشتر می‌توان به بررسی طرز تفکر، سبک نگارش و توانایی دانشجو در ارائه نتیجه‌گیری منطقی از سطح وسیعی از دانسته‌ها پرداخت.

در همین راستا، یکی از انتقادهای جدی که همواره به این سؤالات وارد می‌شود این است که در یک آزمون تشریحی، جدا از اینکه سطح سواد دانشجو چقدر باشد، نحوه نگارش، سبک و شیوه ارائه مطالب یا دست‌خط او ممکن است در نمره‌ای که می‌گیرد، تاثیر بگذارد. بنابراین نکته‌ای که در اینجا باید به آن توجه نمود، بحث تجانس روش ارزیابی با هدف آموزشی است. در مواقعی که سنجش توانمندی‌های ذکر شده بخشی از هدف آموزشی باشد، استفاده از این سؤالات توصیه می‌گردد. در غیر این صورت، این عوامل موجب زیاد شدن خطای نمره‌دهی می‌شوند.

نمونه‌ای از سؤال تشریحی گسترده‌پاسخ

مکانیسم افزایش قند خون را متعاقب کاهش سطح انسولین در خون شرح دهید؟

سؤالات تشریحی محدودپاسخ

برخلاف سؤالات تشریحی گسترده‌پاسخ، در سؤالات تشریحی محدودپاسخ دانشجو از آزادی کامل برای ارائه پاسخ برخوردار نیست. طراح سؤال با به کار بردن کلماتی خاص در صورت سؤال و یا با ارائه دستورالعمل‌های موردنظر، فراگیر را ملزم به ارائه پاسخ در یک قالب مشخص می‌کند. به عنوان مثال، تعداد پاراگراف یا کلمات مورد نیاز برای پاسخ ذکر می‌شود. ذکر شرایط خاص در صورت سؤال باعث ایجاد محدودیت برای دانشجو در تولید پاسخ می‌شود که بالتبع باعث کاهش خلاقیت در ارائه پاسخ‌های احتمالی می‌شود. به‌همین دلیل در این آزمون‌ها در مقایسه با نوع گسترده‌پاسخ، کمتر می‌توان شاهد نوآوری و خلاقیت از سوی فراگیران بود. اما قاعدتاً این موضوع باعث سهولت در تصحیح پاسخ‌های ارائه شده و هماهنگ‌تر شدن شیوه نمره‌دهی مصححان خواهد شد.

نمونه‌ای از سؤال تشریحی محدودپاسخ

کاهش کلسیم خارج سلولی موجب ایجاد تشنج و انقباضات عضلانی می‌شود. علت این امر را در یک پاراگراف توضیح دهید.

نمونه‌ای از سؤال تشریحی محدودپاسخ

تفاوت آگنوزی و آفازی نوع اسمی را توضیح دهید؟ (لطفاً پاسخ خود را در نیمی از صفحه بنویسید.)

ضرورت و کاربرد سؤالات تشریحی

به طور کلی دو هدف عمده استفاده از سؤالات تشریحی شامل «ارزیابی توانایی درک و تفکر فراگیران در خصوص موضوع موردنظر» و «سنجش توانایی نگارش و سازمان‌دهی مطالب» است. اما در مجموع از سؤالات تشریحی در موارد زیر استفاده می‌شود:

- ارزیابی توانایی استنتاج و استدلال فراگیران در مواجهه با موضوعات پیچیده

نمونه‌ای از سؤال تشریحی

در تبدیل پرواریتروبلاست به اریتروبلاست که دارای مراحل زیر است، هسته سلول چه تغییراتی می‌کند و سرنوشت آن چیست؟
Polychromatophilic → Basophilic erythroblast → Proerythroblast erythroblast → orthochromat philic erythroblast → Erythrocyte → Reticulocyte

- ارزیابی توانایی نگارش و خلاقیت فرد در ارائه مطالب علاوه بر دانش وی
- کشف دیدگاه‌ها و نظرات جدید دانشجویان در خصوص یک موضوع

نمونه‌ای از سؤال تشریحی

دیدگاه خود را در خصوص نظام ارائه خدمات سلامت در کشور در یک صفحه بنویسید.

- ارزیابی سطوح بالای حیطه شناختی (تحلیل، ترکیب و ارزشیابی)

نمونه‌ای از سؤال تشریحی

نوزاد دختری دارای ضایعات صورتی رنگ موضعی می‌باشد. علاوه بر این وی دارای عوارض ثانویه‌ای شامل زخم و انسداد مجاری هوایی نیز است. در معاینه توسط پزشک همانژیومای مویرگی تشخیص داده شد. علت تکثیر غیرطبیعی عروق در ضایعه را شرح دهید؟

- ارزیابی یک موضوع از جوانب گوناگون و به طرق مختلف

نمونه‌ای از سؤال تشریحی

شما پزشک کشیک بیمارستان هستید. مهدی مرد ۳۴ ساله است که مبتلا به دیستروفی عضلانی است و اینک به علت پنومونی به اورژانس مراجعه کرده است. خواب آلوده ولی هوشیار است. پزشک خانوادگی وی در نامه‌ای نوشته است: مهدی از حیث قوای ذهنی هوشیار و طبیعی است و اخیراً یک دوره تحصیلی دانشگاهی را به پایان رسانده است، اما وضعیت جسمانی وی رو به زوال بوده به طوری‌که برای انجام بیشتر فعالیت‌های خود نیازمند والدینش می‌باشد. می‌خواهید او را در بخش بستری کنید. پدرش با او به بیمارستان آمده است. شما او را در راهرو می‌بینید و او از شما احوال پسرش را می‌پرسد، چگونه پاسخ می‌دهید؟ (لطفاً پاسخ خود را در دو پارگراف تنظیم نمایید)

- سنجش قدرت سازمان‌دهی و نتیجه‌گیری دانشجو از طیف وسیعی از اطلاعات

نمونه‌ای از سؤال تشریحی

با توجه به فرایندهای فیزیولوژیک و آناتومیک، تغییراتی که در بدن نوزاد بلافاصله بعد از تولد صورت می‌گیرد را شرح دهید.

- تشویق دانشجویان به درک عمیق مطالب آموزشی

گاهی اوقات، ارزیابی اهداف مورد انتظار یادگیری هم با استفاده از آزمون‌های تشریحی و هم آزمون‌های چندگزینه‌ای امکان‌پذیر است. در این صورت در موارد زیر استفاده از سؤالات تشریحی توصیه می‌شود:

- زمانی که طراح سؤال از توانایی کافی برای طراحی سؤالات مناسب چندگزینه‌ای برخوردار نیست و در مقابل منابع و زمان لازم برای تصحیح و نمره‌دهی در دسترس است؛ به عنوان مثال زمانی که تعداد دانشجویان کم است.
- زمانی که هدف اصلی از اجرای آزمون، ارزیابی فرایند و نحوه استدلال فراگیران است.

به طور کلی با توجه به محدودیت‌های حاصل از اجرای آزمون‌های تشریحی برخی پژوهشگران توصیه کرده‌اند

که از این سؤالات به شکل تکوینی استفاده شود و اگر در نظر است که از این سؤالات در آزمون تراکمی استفاده شود، برای گرفتن نتیجه بهتر حتماً سؤالات چندگزینه‌ای نیز همزمان با آن‌ها در آزمون استفاده شود. به عبارت دیگر به منظور دستیابی به حداکثر پایایی و روایی در آزمون توصیه می‌شود که از این سؤالات به شکل تکمیلی و همراه با سایر آزمون‌های بسته‌پاسخ استفاده شود (رسائیان ۱۳۸۱).

گام‌های طراحی سؤال تشریحی

بسیاری از اعضای هیأت علمی تجربه طراحی سؤالات تشریحی برای ارزیابی فراگیران را دارند اما مانند سایر آزمون‌ها، رعایت اصول و قواعدی که برای طراحی نوع مطلوب این سؤالات توصیه شده است، منجر به افزایش صحت و کارایی ارزیابی می‌گردد. خلاصه مراحل طراحی سؤال تشریحی در جدول ۱-۱۰ آمده است.

جدول ۱-۱۰: خلاصه مراحل طراحی یک سؤال تشریحی

ردیف	عنوان مرحله	توضیح
۱	انتخاب یک موضوع و هدف آموزشی مناسب	موضوعات قسمتی از اطلاعات و دانش داوطلب هستند که قرار است سؤال آن را بسنجد. این موضوعات در ارتباط با توان‌مندی‌هایی است که قرار است در داوطلبان مورد ارزیابی قرار گیرد.
۲	طراحی صورت سؤال به شکل مناسب	در هنگام طراحی صورت سؤالات تشریحی باید از به کاربردن افعال مبهم خودداری کنید.
۳	تصمیم‌گیری در خصوص پاسخ‌های صحیح و نحوه نمره‌دهی	از آنجا که در این نوع سؤالات نیاز است تا فراگیران خود پاسخ‌ها را تولید و ارائه کنند، احتمالاً پاسخ‌های متنوعی از دانشجویان دریافت خواهد شد. به همین منظور ضروری است که در خصوص نحوه نمره‌دهی به هر سؤال و روش مورد استفاده، پیش از اجرای آزمون تصمیمات لازم را اتخاذ کنید.
۴	مرور و ارزیابی سؤال	از چک‌لیست ارزیابی سؤالات تشریحی استفاده کنید. از همکاران خود بخواهید به سؤال پاسخ دهند. معمولا دیگران متوجه اشتباه‌هایی می‌شوند که طراح سؤال ممکن است در هنگام طراحی سؤال به آنها توجه نکند یا فراموششان کند.

انتخاب یک موضوع و هدف آموزشی مناسب

نکته اول و اساسی در طراحی سؤالات تشریحی آن است که از این سؤالات برای ارزیابی موضوعاتی استفاده شود که نیازمند به کارگیری سطوح بالای حیطه شناختی است. بهتر است از سؤالاتی که صرفاً اطلاعات تئوری و محفوظات را می‌سنجند، پرهیز شود و از موضوعاتی استفاده شود که سطوح بالاتر یادگیری را می‌سنجند. در هنگام طراحی سؤالات تشریحی باید بخش مناسبی از محتوای دوره را که با استفاده از سایر آزمون‌های بسته‌پاسخ قابل سنجش نیست، انتخاب کرد. این موضوع از آن جهت حائز اهمیت است که تعداد سؤالاتی را که در یک آزمون تشریحی می‌توان طراحی نمود، در مقایسه با سؤالات بسته‌پاسخ بسیار محدود است. لذا انتخاب بخش مناسبی از محتوای دوره برای ارزیابی به وسیله سؤالات تشریحی از حساسیت بالایی برخوردار است. بهترین روش برای این کار، استفاده از بلوپرینت برای انتخاب ابزار ارزیابی متناسب با هدف و محتوای آموزشی است. در این شرایط با استفاده از فاکتورهای متعدد موجود در بلوپرینت، از قبیل میزان زمان اختصاص یافته برای آموزش هر هدف آموزشی و سطح شناختی مورد انتظار (در تاکسونومی بلوم)، می‌توان مشخص کرد که آیا استفاده از سؤالات تشریحی برای سنجش هدف مورد نظر مناسب است یا خیر. بدیهی است که اگر یک آزمون از انتخاب تاکسونومی مناسب سؤالات برخوردار نباشد، نه تنها نقش اصلی آزمون به عنوان جزء تکمیل کننده حیاتی چرخه آموزش از دست می‌رود، بلکه این امر می‌تواند اثرات منفی بر عملکرد یادگیری فراگیران و سیستم آموزشی داشته باشد.

بنابراین به‌طور خلاصه می‌توان گفت برای طراحی نوع سؤال مورد استفاده، اولین گام توجه به پیامدهای مورد انتظار یادگیری و اهداف آموزشی است. به گونه‌ای که می‌توان با اطمینان گفت در صورتی‌که در هنگام طرح سؤال، اهداف آموزشی در نظر گرفته نشوند، روایی محتوایی آزمون زیر سؤال خواهد رفت. نکته دیگری که باید در هنگام طراحی سؤال در نظر داشت، قابلیت سنجش اهداف آموزشی است. مثال زیر را در نظر بگیرید:

هدف شناختی ضعیف
دانشجو بتواند فرایند تقسیم سلولی را درک کند.
هدف شناختی بهتر
دانشجو بتواند با ارائه یک نمودار نمایشی، گام‌های اصلی در فرایند تقسیم سلولی را با یکدیگر مقایسه و تحلیل کند.

هدف اول بسیار کلی است و سنجش فعل «درک کردن» دشوار و مبهم است. به ویژه آنکه قضاوت در خصوص میزان درک فراگیران از یک موضوع و اختصاص نمره به پاسخ‌های فراگیران، کار بسیار دشواری است. بنابراین توجه به این نکته ضروری است که استفاده از اهداف کلی برای توصیف یک دوره آموزشی یا ارائه راهنما برای فرایند یادگیری مناسب است اما برای سنجش عملکرد فراگیران مفید نخواهد بود. افعال مورد استفاده در اهداف آموزشی، تقریباً همیشه در حکم راهنمایی برای انتخاب ابزار ارزیابی عمل می‌کنند و باید با دقت انتخاب شوند تا نشان دهند که چه فرایندهای فکری و عملکردی باید از سوی فراگیران انجام گیرد تا شواهد ارائه شده معرف یادگیری آنان باشد. به همین دلیل افعال انتخابی تحت عنوان افعال راهنما[1] شناخته می‌شوند (موس و هولدر[2] ۱۹۸۸). قاعدتاً اهمیت این موضوع در سؤالات تشریحی به دلیل ماهیت ذهنی بودن آن بیشتر مورد تاکید قرار می‌گیرد. توجه به این نکته ضروری است که افعال مورد استفاده باید باعث هدایت پاسخ فراگیران به سمت عملکرد مورد ارزیابی شوند. به عنوان مثال تعدادی از افعال به وضوح نشان می‌دهند که فراگیر باید پاسخ را خود ارائه دهد (به عنوان مثال شرح دادن)، به جای آنکه آن را از بین مجموعه‌ای از گزینه‌های انتخاب نماید. برخی دیگر از افعال نشان می‌دهند که برای دستیابی به اهداف مورد انتظار یادگیری، فراگیر باید پاسخ را انتخاب کند (به عنوان مثال تعیین کردن).

ون هوئیج و همکاران [1] ۲۰۰۴

در پژوهشی، سطح شناختی سؤالات تشریحی مربوط به دو دوره آموزشی در دانشکده پزشکی دانشگاه اتریش با استفاده از یک ابزار طبقه‌بندی ساده بر اساس تاکسونومی بلوم ارزیابی شد. در این مطالعه ۸۹ سؤال پاتوفیزیولوژی و ۹۰ سؤال ژنتیک استفاده شد. طبقه‌بندی سؤالات به وسیله اعضای هیات علمی که در فرایند آموزش مشارکت داشتند و تعدادی دیگر از مصححان که در تدریس دوره‌های آموزشی مربوطه حضور نداشتند، صورت گرفت. به تمام مصححان، مواد آموزشی دوره از قبیل سیلابوس‌ها، راهنماهای آموزشی و کتابچه‌های راهنمای مطالعه ارائه شد و از آنان درخواست شد تا پیش از شروع تصحیح سؤالات تشریحی مطالعه نمایند. آنها همچنین سؤالات امتحانی، الگوهای نمره‌دهی، نمونه سؤالاتی برای هر سطح شناختی و فرم نمره‌دهی را نیز دریافت کردند. در فرم نمره‌دهی، مصححان باید سطح شناختی هر سؤال را مشخص می‌کردند. همچنین در مقابل هر سؤال فضایی برای ارائه پیشنهادهای مصححان در خصوص قضاوت‌هاشان در نظر گرفته شده بود. ارزیابان ابتدا با استفاده از معیارهای طبقه‌بندی، سطح شناختی سؤالات را مشخص می‌کردند و سپس از آنان خواسته می‌شد تا مجدداً سطح شناختی سؤالات را تعیین نموده و با نتایج حاصل از ارزیابی اولیه مقایسه نمایند. تنها آن دسته از ارزیابانی که پایایی نمراتشان در دو بار ارزیابی یک سؤال بالا بود در مطالعه شرکت داده شدند. یافته‌های مطالعه نشان داد که سطح توافق بین نمرات دو گروه ارزیابان از ۰/۳۴ تا ۰/۷۷ و ضریب پایایی کاپا آنان از ۰/۱۲ تا ۰/۶۰ متغیر بود. سطح توافق بین ارزیابانی که در فرایند تدریس شرکت داشتند و آنهایی که شرکت نداشتند، در دوره آموزشی ژنتیک و پاتوفیزیولوژی به ترتیب برابر با ۶۵ درصد و ۷۳ درصد بود. محققان در این مطالعه نتیجه گرفتند که معرفی ابزار طبقه‌بندی در فرایند تدریس اعضای هیات علمی و آگاهی آنان از اهمیت توجه به سطح شناختی در فرایند ارزیابی تاثیر مثبت دارد. همچنین آنها در مطالعه خود نشان دادند که تدوین یک ابزار طبقه‌بندی دقیق توافق بین ارزیابان را افزایش می‌دهد

1. Van Hoeij et al

1. Directive verbs
2. Moss & Holder

به صورت کلی، با این که هیچ معیار مشخصی برای اینکه از چه نوع سؤالی باید برای هر هدف آموزشی استفاده نمود وجود ندارد اما نکات کلیدی وجود دارد که باید در این راه به آن‌ها توجه نمود. ارزیابان می‌توانند از طریق پاسخگویی به چند سؤال تشخیص دهند که آیا استفاده از سؤال تشریحی برای ارزیابی پیامد مورد انتظار یادگیری مناسب است یا خیر؟

- آیا مناسب‌تر است که فراگیران عملکرد و فرایندهای فکری خود را از طریق تولید پاسخ ارائه نمایند یا انتخاب پاسخ از میان مجموعه گزینه‌های موجود صورت گیرد؟
- مهارت طراحان سؤال در طراحی سؤالات چندگزینه‌ای مطلوب به چه صورت است؟
- منابع و زمان در دسترس برای تصحیح و نمره‌دهی، تعداد فراگیران و وسعت دانش مورد اندازه‌گیری چگونه است؟

طراحی صورت سؤال

همان‌طور که پیشتر نیز اشاره گردید، هنگام طراحی سؤالات تشریحی باید در نظر داشت که سؤال طراحی شده به گونه‌ای نوشته شود که با اهداف آموزشی منطبق باشد. به‌علاوه در هنگام طراحی صورت سؤالات تشریحی باید از به کاربردن افعال مبهم که منجر به ارائه پاسخ‌های متنوع از سوی فراگیران می‌شود، خودداری نمود. این موضوع به ویژه در بحث تصحیح باعث بروز مشکلات جدی خواهد شد. به عبارت دیگر، استفاده از سؤالات تشریحی ساختارمند به اعضای هیأت علمی کمک می‌کند که از بروز دو مشکل گزافه‌گویی[1] و دشواری نمره‌دهی تا حدودی جلوگیری کنند. به طور کلی، در این مرحله باید به نکات زیر توجه شود:

- **پرهیز از کلی‌گویی و نگارش صورت سؤال با عبارات و کلمات واضح:** استفاده از عبارات کلی و مبهم در سؤالات تشریحی باعث می‌شود که دانشجویان هر طور که خود تمایل دارند پاسخ را ارائه دهند، به این معنا که پاسخ‌های فراگیران ممکن است خارج از هدف مورد ارزیابی باشد یا تنها برای بخشی از آن کافی باشد. بنابراین، طراح سؤال باید سؤال را به گونه‌ای طراحی کند که باعث محدود نمودن پاسخ‌های ارائه شده از سوی فراگیران شود.
- **استفاده از کلماتی مانند «چرا»، «چگونه»، یا «به چه دلیل»:** استفاده از این کلمات منجر به تشویق فراگیران به ترکیب و سازمان‌دهی اطلاعات فرا گرفته به شیوه جدید می‌شود در حالی‌که اگر هنگام طراحی سؤالات تشریحی از کلماتی مانند «چه وقت»، «چه کسی» و «کجا» استفاده شود، تنها دانش فراگیران در سطح یادآوری سنجیده خواهد شد.
- **خودداری از فعل‌هایی مانند «بررسی کنید»، «بحث کنید» و «بیابید»:** بیشتر اوقات اعضای هیأت علمی که از تجربه و تخصص کافی برای سؤالات تشریحی برخوردار نیستند، به اشتباه تصور می‌کنند که در صورت استفاده از افعالی از قبیل «بررسی کنید»، «بحث کنید» و ... می‌توانند سطوح بالاتری از حیطه شناختی را مورد ارزیابی قرار دهند. در حالی‌که در این حالت به وضوح مشخص نیست که از فراگیر خواسته شده است چه کاری انجام دهد. به جای آن بهتر است از افعالی که به طور واضح تکلیف فراگیر را مشخص می‌کنند، استفاده شود مانند «توضیح دهید»، «استفاده کنید»، «تفسیر کنید» و «مقایسه کنید»). در مثال زیر در حالت اول مشخص نیست که فراگیر تاثیر ابتلا به دیابت را بر چه چیز باید توضیح دهد. تاثیر آن را بر سیستم کلیوی، قلب و عروق یا بینایی؟ بنابراین بهتر است سؤال به صورت حالت دوم اصلاح شود.

سؤال ضعیف
تاثیرات بیماری دیابت را بر بدن بحث کنید.
سؤال بهتر
تاثیرات بیماری دیابت را بر سیستم کلیوی توضیح دهید.

1. Bluffing

- **خودداری از عبارت‌هایی نظیر «به نظر شما»، «شما در این خصوص چه فکر می‌کنید»:** این عبارات به فراگیر این امکان را می‌دهند که نظر شخصی خود را در خصوص یک موضوع ارائه نماید. از آنجا که هر یک از افراد ممکن است از منظر خود به سؤال جواب دهد، احتمالاً پاسخ ارائه‌شده اشتباه نخواهد بود. بنابراین استفاده از این افعال تنها در شرایطی توصیه می‌گردد که هدف، سنجش نظرات و ایده‌های فردی فراگیران است نه سنجش دانش افراد.

سؤال ضعیف
به نظر شما اهمیت رعایت اصول اخلاقی در برخورد با بیمار چیست؟
سؤال بهتر
بر اساس اصول مندرج در بیانیه هلینیسکی [1] توضیح دهید چرا رعایت اصول اخلاقی در برخورد با بیمار اهمیت دارد؟

1. Declaration of Helsinki

- **ذکر تمام مواردی که در پاسخ باید آورده شود:** یکی از مشکلاتی که در اغلب اوقات در سؤالات تشریحی به چشم می‌خورد، عدم ذکر موضوعاتی است که فراگیر باید در پاسخ خود لحاظ کند. اگر لازم است فراگیر اقدام خاصی برای پاسخ انجام دهد، مثلاً مثال بزند یا واحد اندازه‌گیری خاصی را ذکر کند، حتماً این موضوع باید در صورت سؤال قید شود. به مثال زیر توجه کنید:

سؤال ضعیف
چرا در اثر مجاورت با فلزات سنگین، تغییرات رنگ پوست رخ می‌دهد؟
سؤال بهتر
با ذکر مثال توضیح دهید چرا در اثر مجاورت با فلزات سنگین رنگ پوست تغییر می‌کند.

- **خودداری از طرح سؤال‌های انتخابی:** ارائه فرصت برای انتخاب سؤالات توسط فراگیران از چندین منظر قابل نقد و تحلیل است. اولاً از آنجا که طراحی سؤالات همگن که از ارزش یکسانی برخوردار باشند کار بسیار دشواری است و اصل سنجش عادلانه خدشه‌دار می‌شود. ثانیاً معمولاً دانشجویان ساعی کلاس در انتخاب خود به سراغ سخت‌ترین سؤالات برای پاسخ می‌روند که این موضوع باعث می‌شود تا نمره پایین‌تری در امتحان کسب کنند و آخر اینکه در مواردی که یکی از اهداف آزمون، شناسایی مشکلات یادگیری فراگیران است، به دلیل انتخابی بودن سؤالات در این حالت امکان تشخیص مشکلات یادگیری فراگیران وجود ندارد. به طور خلاصه، بهتر است از سؤالات انتخابی به دلایل زیر استفاده نشود:
  - وقت دانشجویان ممکن است در نتیجه انتخاب سؤالات به هدر رود.
  - پاسخگویی به تعدادی از سؤالات ممکن است دشوارتر از بقیه باشد. در این حالت مقایسه عملکرد دانشجویان عادلانه نخواهد بود.
  - ارزیابی میزان دانسته‌های فراگیران در خصوص یک موضوع مشخص دشوار خواهد بود که بالتبع این موضوع مانع شناسایی مشکلات یادگیری فراگیران خواهد شد.
  - استفاده از سؤالات انتخابی عمدتاً به ضرر دانشجویان ساعی خواهد شد (سیف ۱۳۹۴).

تصمیم‌گیری در خصوص پاسخ‌های صحیح و تعیین نحوه نمره‌دهی

از آنجا که در این نوع سؤالات نیاز است تا فراگیران خود پاسخ‌ها را تولید و ارائه کنند، احتمالاً پاسخ‌های متنوعی از دانشجویان دریافت خواهد شد. به همین منظور ضروری است که پیش از اجرای آزمون در خصوص نحوه نمره‌دهی به هر سؤال تصمیمات لازم اتخاذ گردد. این موضوع مخصوصاً وقتی تعداد دانشجویان زیاد است یا هنگامی که بیش از یک مصحح وجود دارد، حائز اهمیت است. انجام این کار منجر به بهبود پایایی بین ارزیابان و ثبات نمرات حاصل از ارزیابی خواهد شد. در واقع دلیل عمده مخالفت برخی از صاحب‌نظران برای استفاده از سؤالات تشریحی از بی‌ثباتی نمره‌گذاری آنها ناشی می‌گردد. بنابراین در هنگام نمره‌دهی سؤالات تشریحی باید همواره به نکات زیر توجه داشت:

- **تهیه الگوی پاسخ برای سؤال:** همان‌طور که پیشتر نیز اشاره شد، یکی از انتقاداتی که همواره به سؤالات تشریحی وارد می‌شود، بحث پایین بودن پایایی بین ارزیابان در این سؤالات است. به منظور یکسان‌سازی فرایند ارزیابی سؤالات تشریحی و متعاقب آن افزایش میزان پایایی بین ارزیابان، باید حتماً پیش از اجرای آزمون الگوی پاسخ تهیه شود. این کار نه تنها به اعضای هیأت علمی در فرایند نمره‌دهی کمک خواهد نمود، بلکه با ارائه الگوی پاسخ به دانشجویان بعد از اجرای آزمون، می‌توان به تقویت اثر آموزشی آزمون نیز کمک کرد. ذکر این نکته ضروری است که در بیشتر اوقات به دلیل انتظارات بالای اعضای هیأت علمی از فراگیران، الگوی پاسخی که قبل از برگزاری آزمون طراحی شده است، در عمل نیاز به اصلاح دارد. بنابراین بهتر است پیش از شروع نمره‌دهی، ابتدا چند برگه امتحانی به سرعت مطالعه شوند و سپس در خصوص نیاز به اصلاح بخش‌هایی از کلید پاسخ تهیه شده تصمیمات لازم اتخاذ گردد.
- **تصحیح پاسخ‌ها به صورت سؤال به سؤال:** یکی از مشکلات مربوط به سؤالات تشریحی در فرایند نمره‌دهی، تاثیر نمره اختصاص داده شده به یک سؤال بر نمرات مابقی سؤالات آزمون است. در واقع در صورتی که تصحیح سؤالات آزمون به شکل برگه به برگه صورت گیرد، احتمال اینکه مصحح تحت تاثیر پاسخ‌های قبلی دانشجو قرار گیرد زیاد است. به منظور رفع این مشکل توصیه می‌شود که در هنگام تصحیح برگه‌های امتحانی، پاسخ یک سؤال برای تمامی آزمون شوندگان بدون فاصله زمانی ارزیابی شود و سپس مصحح به سراغ ارزیابی سایر سؤالات برود.
- **تصحیح پاسخ‌های تمام آزمون‌شوندگان به یک سؤال در یک زمان و بدون وقفه:** به این موضوع بارها اشاره شد که از جمله مشکلات مربوط به سؤالات تشریحی، ماهیت ذهنی بودن فرایند ارزیابی آن‌هاست. به گونه‌ای که نمره‌دهی به این سؤالات می‌تواند، به شدت تحت تاثیر شرایط روحی و حتی خستگی ارزیابان قرار گیرد. لذا یکی از توصیه‌هایی که به منظور به حداقل رساندن این مشکل پیشنهاد می‌گردد، تصحیح پاسخ‌های تمام آزمون‌شوندگان به یک سؤال در یک زمان و بدون وقفه زمانی است. انجام این کار با اصول ارزیابی عادلانه نیز منطبق خواهد بود زیرا به این طریق سعی می‌شود تا در حد امکان شرایط برای تمامی شرکت‌کنندگان یکسان در نظر گرفته شود.
- **تعیین بارم سؤال‌ها به صورتی که قابل تقسیم‌بندی باشد:** روش‌های مختلفی برای تصحیح سؤالات تشریحی ارائه شده است که در ادامه به تفصیل به شرح آن‌ها خواهیم پرداخت. در تعدادی از این روش‌ها نیاز است که نمره سؤال بین بخش‌های مختلف پاسخ شکسته و پخش شود از این رو با توجه به ارزش هر سؤال، باید نمره‌ای برای پاسخ‌ها در نظر گرفت که به نوعی قابل تقسیم‌بندی باشد.
- **کدگذاری برگه‌های امتحانی:** یکی دیگر از مشکلات نمره‌دهی سؤالات تشریحی بروز اثر هاله‌ای[1] است. مراد از این اثر آن است که تصور اعضای هیأت علمی از عملکردهای قبلی فراگیران‌شان به شدت بر فرایند نمره‌دهی آنان تاثیرگذار خواهد بود. به منظور جلوگیری از بروز این مشکل، در هنگام نمره‌دهی بهتر است از کد برای شناسایی برگه‌های امتحانی استفاده شود. حتی پیشنهاد می‌شود که پیش از شروع به تصحیح، برگه‌های امتحانی با یکدیگر

1. Hallo effect

مخلوط شوند تا ترتیب تقدم و تاخر آن‌ها بهم بخورد و مشخص نباشد که برگه‌های ابتدایی مربوط به دانشجویان قوی یا ضعیف کلاس است.

- **ذکر اشتباهات فراگیران در برگه پاسخ:** یکی از اصلی‌ترین اهداف آزمون، تقویت میزان یادگیری فراگیران است. ذکر اشتباهات فراگیران در برگه امتحانی منجر به افزایش تاثیر آموزشی آزمون خواهد ش و به این طریق فراگیران در جهت رفع نواقص یادگیری خود بر خواهند آمد.

مرور و ارزیابی سؤالات

چهارمین مرحله از فرایند طراحی سؤالات تشریحی، نقد و مرور سؤالات طراحی‌شده است. برای انجام این کار می‌توان از چک‌لیست ارزیابی سؤالات تشریحی که در جدول ۲-۱۰ آمده است، استفاده کرد. به طور ویژه دو مورد زیر باید با دقت بررسی شوند:

جدول ۲-۱۰: چک لیست ارزیابی سؤالات تشریحی

ردیف	سؤال	
۱	آیا برای طراحی سؤال‌ها از بلوپرینت استفاده شده است؟	بله
۲	آیا سؤال با توجه به یک هدف مهم آموزشی طراحی شده است؟	بله
۳	آیا سؤال تنها آن دسته از هدف‌های آموزشی را شامل می‌شود که با سایر انواع سؤال‌ها به خوبی قابل اندازه‌گیری نیست؟	بله
۴	آیا سطح دانشی سؤال (تاکسونومی بلوم) منطبق با هدف آموزشی است؟	بله
۵	آیا هر سؤال به زمینه مشخصی محدود شده است؟	خیر
۶	آیا سطح دشواری سؤالات متناسب با ویژگی‌های آزمون‌شوندگان است؟	بله
۷	آیا از افعال مناسب در صورت سؤال استفاده کرده‌اید؟	بله
۸	آیا آزمون‌شوندگان از طریق راهنمایی‌های اختصاصی به ارائه پاسخ مورد نظر هدایت شده‌اند؟	بله
۹	آیا تعداد سؤالات طراحی شده متناسب با مدت زمان آزمون است؟	خیر
۱۰	آیا در سؤال‌ها موقعیت‌های تازه‌ای بکار گرفته شده است به گونه‌ای که بتواند فرایندهای فکری آزمون‌شوندگان را بر انگیزانند؟	بله
۱۱	آیا راهنمای نمردهی پاسخ‌ها از قبل تهیه شده است؟	خیر

- **پیش‌بینی پاسخ‌های فراگیران:** باید سعی شود تا از دیدگاه دانشجویان به سؤالات پاسخ داده شود و سپس این موضوع ارزیابی شود که آیا دانشجویان از دانش و مهارت کافی برای پاسخ‌گویی مناسب به سؤالات برخوردار هستند یا خیر. به عبارت دیگر آیا سؤال طراحی شده متناسب با سطح فراگیران است.
- **نقد و ارزیابی الگوی پاسخ:** هنگام بررسی الگوی پاسخ باید به طور پیوسته، میزان انطباق آن را با سؤال مربوطه و پیامد مورد انتظار یادگیری مورد بررسی قرار داد تا در صورت نیاز، تغییرات لازم به منظور بالا بردن هم‌راستایی ساختاری در فرایند ارزیابی صورت گیرد.

معمول‌ترین راه برای ارزیابی کیفیت سؤالات طرح شده این است که سؤالات طراحی‌شده پیش از اجرا، به یکی از همکاران داده شود تا به آن پاسخ دهد. معمولاً دیگران متوجه اشتباه‌هایی می‌شوند که طراح ممکن است در هنگام طراحی سؤال به آنها توجهی نکند یا آنها را فراموش کند. مزیت دیگر انجام این کار این است که با بررسی پاسخ‌های ارائه شده

به‌وسیله همکار، طراح سؤال می‌تواند دریابد که پاسخ به سؤال نیازمند استفاده از فرایند شناختی سطح بالا است یا خیر. همچنین به این طریق می‌توان از مناسب بودن سطح دشواری سؤال اطمینان حاصل کرد و مطمئن شد که سؤالات طرح شده منطبق با پیامدهای مورد انتظار یادگیری هستند. در صورتی که به دلیل دغدغه‌های کاری زیاد همکاران، امکان استفاده از کمک ایشان وجود نداشته باشد، می‌توان از یکی از دانشجویان سال بالاتر در خواست کرد که به سؤالات طرح شده پاسخ دهد و حتی در صورت عدم امکان انجام این کار، می‌توان به صورت پایلوت آزمون را برای دانشجویان این دوره اجرا کرد و سپس بعد از رفع اشکالات موجود، آن‌ها را در بانک سؤالات قرار داد و در دوره‌های بعدی از آن‌ها استفاده نمود. همچنین در صورت وجود کمیته مرور سؤال در دانشگاه، می‌توان از مشاوره‌های تخصصی آنها برای بهبود کیفیت سؤالات طراحی شده استفاده نمود.

تصحیح سؤالات تشریحی

به طور معمول سه روش برای تصحیح سؤالات تشریحی وجود دارد: روش تحلیلی[1]، روش کلی[2] و روش ویژگی‌های اصلی[3] که به ترتیب آنها را شرح می‌دهیم (سیف ۱۳۹۴):

روش تحلیلی

در روش تحلیلی که به آن تصحیح بر مبنای نکات مورد انتظار[4] هم می‌گویند، الگوی پاسخ به چندین بخش تقسیم می‌شود و سپس برای هر بخش، نمره یا امتیاز مشخصی در نظر گرفته می‌شود. از آنجا که در این روش برای هر بخش امتیاز خاصی تعیین می‌شود به‌همین دلیل به آن، روش امتیازبندی[5] نیز اطلاق می‌شود. استفاده از این روش به دلیل مشخص بودن معیارهای ارزیابی تا حدودی مشکل تاثیر ذهنیت و پیش‌فرض‌های ارزیاب را بر فرایند نمره‌دهی کاهش می‌دهد.

روش کلی

اصطلاح نمره‌دهی کلی برای اولین بار در سال ۱۹۶۰ و در پاسخ به لزوم معرفی یک روش معتبر و اقتصادی در سنجش مستقیم مهارت‌های دانشجویان در نوشتن ارائه شد (هانتر و همکاران[6] ۱۹۹۶). از نظر پل دایدریک[7]، در روش تصحیح کلی آزمونگر کل پاسخ دانشجو را مطالعه می‌کند و سپس در مورد کیفیت آن قضاوت می‌نماید. در واقع در این روش آزمونگر بر اساس برداشت کلی خود از پاسخ به آن نمره می‌دهد. در این روش برای هیچ یک از عوامل امتیاز جداگانه‌ای در نظر گرفته نمی‌شود، بلکه کیفیت کلیه عوامل در ارتباط با یکدیگر قضاوت می‌شوند. از آنجا که در این روش، ارزیاب برای نمره‌دهی باید درجه کیفیت پاسخ ارائه شده را تعیین کند، به‌همین دلیل به این روش، درجه‌بندی[8] نیز اطلاق می‌گردد. در این روش، نمره ممکن است به شکل حروف (مانند الف تا د) یا نوعی مقیاس لیکرت (به طور مثال رد، مرزی، متوسط، خوب، عالی) ارائه شود. کاشین[9] در سال ۱۹۸۷ روشی برای تصحیح کلی پیشنهاد کرده است که در آن مصحح کل پاسخ را به سرعت می‌خواند و آن را به قسمت‌هایی با نمرات مختلف طبقه‌بندی می‌کند. سپس هر قسمت از پاسخ را دوباره می‌خواند و مطمئن می‌شود که نمره به صورت صحیح و عادلانه به هر قسمت اختصاص داده شده است.

1. Analytic
2. Holistic
3. Primary traits
4. Point- Scoring
5. Rating
6. Hunter et al.
7. Paul Diederich
8. Grading
9. Cashin

نورسینی و همکاران [1] ۱۹۹۰

در این مطالعه پژوهشگران به بررسی نحوه تصحیح و پایایی یک آزمون تشریحی در خصوص قضاوت بالینی پرداختند. هدف آن‌ها از مطالعه، بررسی پایایی نمرات حاصل از روش نمره‌دهی تحلیلی در مصححان پزشک و غیر پزشک بود. سؤالات تشریحی طراحی شده در مطالعه مقدماتی پایلوت شده و در نهایت ۱۲ سؤال از مجموع آن‌ها انتخاب شد. الگوی پاسخ‌دهی توسط محققان مطالعه طراحی شد و سپس در طول فاز پایلوت اصلاحات لازم بر روی آن صورت گرفت. به آزمون‌شوندگان آموزش‌های لازم برای توجه به مسائل روانشناختی-اجتماعی در کنار به کارگیری دانش پزشکی در ارائه برنامه درمانی ارائه شده بود. در فرایند ارزیابی برای هر یک از پاسخ‌های ارائه شده با توجه به اهمیت آن وزنی در نظر گرفته شده بود اما در چک‌لیست طراحی شده هیچ گونه اطلاعاتی در خصوص قوانین نمره‌دهی و وزن هر پاسخ ارائه نشده بود تا بر عملکرد مصححان تاثیر نگذارد. نمره‌دهی به روش کلی توسط ۱۲ متخصص داخلی در مقیاس لیکرت ۹ تایی صورت گرفت. به منظور آموزش مصححان غیر پزشک (سه نفر)، دوره ۱۴ ساعته برگزار شد که شامل ارائه دستورالعمل‌های کلی نمره‌دهی، مرور سؤالات تشریحی و تمرین نمره‌دهی به هر پاسخ بود. مصححان پزشک (دو نفر) نیز دوره آموزشی هفت ساعته مشابه با مصححان غیر پزشک دریافت کردند. آزمون تشریحی برای ۴۷ دانشجوی رزیدنتی سال اول و ۵۱ کارورز اجرا شد. ضریب تعمیم‌پذیری برای یک آزمون تشریحی شامل ۱۲ سؤال برای یک مصحح غیر پزشک برابر با ۰/۳۶ بود. افزایش تعداد مصححان به ۲۰ نفر، بهبود مختصری در پایایی آزمون ایجاد کرد (ضریب تعمیم‌پذیری برابر با ۰/۴۳). زمانی که تعداد مصححان برای هر سؤال تشریحی به دو نفر افزایش یافت و مدت زمان آزمون نیز طولانی شد، ضریب تعمیم‌پذیری به ۰/۸۰ افزایش یافت. در روش نمره‌دهی کلی در نتیجه افزایش تعداد مصححان از یک نفر به سه نفر، ضریب تعمیم‌پذیری افزایش قابل توجهی داشت (از ۰/۶۳ به ۰/۷۳). همبستگی بین میانگین نمرات مصححان پزشک و غیر پزشک ۰/۸۷ شد. همبستگی بین نمرات در روش نمره‌دهی کلی و نمرات حاصل از روش نمره‌دهی تحلیلی مصححان غیر پزشک ۰/۶۶ بود.

1. Norcini et al

روش ویژگی‌های اصلی

در این روش، ارزیاب بر اساس ویژگی‌های اصلی که از قبل برای ارائه پاسخ درست با توجه به موضوع مورد نظر در ذهن دارد، پاسخ آزمون‌شوندگان را ارزیابی می‌کند. به عنوان مثال در صورتی که ارزیاب در نظر دارد که پاسخ مربوط به ارائه توصیه‌های پزشکی به بیمار را ارزیابی کند، باید به وجود عواملی از قبیل سادگی و قابل فهم بودن موارد بیان شده، عدم استفاده از اصطلاحات پزشکی، کوتاه و موجز بودن آنها و ... توجه نماید. در صورتی که در پاسخ دانشجو این ویژگی‌ها لحاظ شده باشد، در این صورت در مورد نمره اختصاصی داده به آن تصمیم گرفته می‌شود. معمولاً برای انجام این روش یک چک‌لیست طراحی می‌شود که بر اساس تعداد ویژگی‌های ارائه شده توسط فراگیر، نمره او تعیین می‌شود. در این روش همچنین به وجود عواملی از قبیل قدرت بیان، شیوه ارائه مطالب و سازمان‌دهی منطقی پاسخ نمره‌ای اختصاص داده می‌شود.

استفاده از الگوی پاسخ مشخص به‌همراه روبریک نمره‌دهی مانند جدول ۳-۱۰ به ارزیابی منصفانه‌تر فراگیران کمک خواهد نمود.

جدول ۳-۱۰: نمونه روبریک نمره‌دهی برای یک سؤال تشریحی چهار نمره‌ای

نمره	تعریف	اطلاعات اضافی	ساختار نگارش
عالی (۴ نمره)	تعریف درست و واضح اطلاعات	ارائه اطلاعات اضافی برای حمایت از تعریف اصلی	مناسب بودن ساختار جملات و کلمات به کارگرفته شده در ارتباط‌دهی بین مطالب
خوب (۳ نمره)	تعریف درست بخشی از اطلاعات	ارائه اطلاعات اضافی برای درک بهتر معنا و مفهوم مطالب	وجود تعدادی خطای گرامری بدون تاثیر بر مفهوم مطالب ارائه شده
متوسط (۲ نمره)	تعریف ضعیف و محدود اطلاعات	ارائه تعدادی از اطلاعات اضافی نسبتاً مرتبط با مطالب	وجود تعدادی خطای گرامری، نگارشی و املایی
ضعیف (۱ نمره)	پاسخ کاملاً نادرست	ارائه اطلاعات اضافی نامرتبط با مطالب	استفاده نامناسب و محدود از کلمات مرتبط با موضوع، وجود تعداد زیادی خطا گرامری، املایی

هانتر و همکاران ۱۹۹۶

محققان در مطالعه خود به طیف روش‌های نمره‌دهی در سئوالات تشریحی اشاره داشتند. در یک طرف این طیف قضاوت و درک کلی آزمونگر از پاسخ به عنوان روش کلی و در طرف دیگر آن، نمره‌دهی جزء به جزء به عنوان روش تحلیلی اشاره شده است. از نظر محققان پنج شیوه‌ای که برای ارزیابی سئوالات تشریحی دراین طیف می‌توان به کار برد، شامل موارد زیر است:

- روش برداشت کلی: در این روش هر آزمونگر بدون داشتن معیار ارزیابی واضح و از پیش تعیین شده، پاسخ را مطالعه و در خصوص کیفیت آن تصمیم گیری می‌کند.
- روش کلی نمره‌دهی : در این روش به منظور ارتقاء پایایی و ثبات نتایج حاصل از نمره‌دهی، از طریق برگزاری جلسات آموزشی و یا تدوین دستورالعمل آزمون، معیارهای نمره‌دهی کلی بین آزمون گران مختلف یکسان می‌گردد. در حالیکه در روش برداشت کلی، هر یک از آزمونگران معیارهای شخصی را بکارمی‌گیرند.
- روش ویژگی‌های اصلی: در این روش برخلاف روش کلی نمره‌دهی، معیارها به صورت دقیق و واضح برای هر ویژگی مدنظر تعیین می‌شود.
- روش تحلیلی، برخلاف پیش فرض حاکم بر روش کلی که کل را فراتر از جمع اجزا می‌داند، در روش تحلیلی جمع اجزا تعیین کننده کل است، به‌همین دلیل در این روش ارزیاب ابتدا نمره بخش‌های مختلف پاسخ را تعیین و سپس نمره کلی را بر اساس مشخص می‌نماید.
- روش جزئی: این روش شامل شمارش و یا توجه به وجود یا عدم وجود نکات مورد نظر در پاسخ است.

خطاهای نمره‌دهی سؤالات تشریحی

به دلیل ماهیت ذهنی بودن فرایند تصحیح و نمره‌دهی سؤالات تشریحی، احتمال بروز مجموعه‌ای از خطاها در فرایند تصحیح این سؤالات وجود دارد که در بیشتر اوقات به طور سهوی و بدون آگاهی مصححان رخ می‌دهند. بدون شک این خطاها می‌توانند بر نتیجه آزمونی که از دانشجویان گرفته می‌شود، تاثیرگذار باشند زیرا در بسیاری از موارد باعث تداخل در نمره فراگیران می‌شوند (هاپکینز و همکاران[1] ۱۹۹۸، نیتکو[2] ۲۰۰۱).

- **اثر هاله‌ای[3]:** در این حالت دیدگاه و نظر مصحح در خصوص فراگیران می‌تواند بر روی نمرات تاثیرگذار باشد. در صورتی‌که مصحح دیدگاه مثبتی نسبت به یک فراگیر داشته باشد، نمرات وی تمایل به بالا رفتن دارد. به‌همین دلیل تا حد امکان، برگه‌های امتحانی باید به طور ناشناس تصحیح شوند تا در حد امکان از بروز اثر هاله‌ای جلوگیری شود.
- **اثر مکانیکی[4]:** نمراتی که مصححان می‌دهند در بیشتر موارد علاوه بر محتوای ارائه شده تحت تاثیر دست‌خط، علایم نگارشی، و حجم پاسخ ارائه شده نیز قرار می‌گیرد. در این حالت دانشجویانی که دست‌خط خوبی ندارند، نمرات پایین‌تری را کسب می‌کنند در حالی‌که ممکن است به خوبی به پاسخ مورد نظر اشاره کرده باشند.
- **اثر ترتیبی[5]:** منظور از اثرات ترتیبی این است که سؤالاتی که در ابتدا تصحیح می‌شوند از نظر مصحح مطلوب‌تر خواهند بود و نمره بالاتری به آنها اختصاص خواهد یافت، درحالی‌که به سؤالاتی که در انتهای جلسه نمره‌دهی ارزیابی می‌شوند به دلیل بالارفتن انتظارات مصححان نمره پایین‌تری اختصاص داده می‌شود.
- **انتقال تاثیر یک سؤال به سؤال دیگر[6]:** در این حالت، دیدگاه مصحح درباره چگونگی پاسخ به سؤالات ابتدایی توسط یک فراگیر بر روی ارزیابی پاسخ سؤالاتی که متعاقب آن‌ها ارائه شده است تاثیر گذار است. به عبارت دیگر در صورتی که مصحح تشخیص بدهد که اولین پاسخ دانشجو خوب است و به آن نمره بالای بدهد، احتمال دارد تحت تاثیر سؤال اول به سؤالات بعدی همان فراگیر نیز نمره بالایی بدهد. به‌همین دلیل به منظور جلوگیری از بروز این پدیده توصیه می‌شود که پاسخ‌های تمام آزمون‌شوندگان به یک سؤال در یک زمان و بدون وقفه زمانی تصحیح شود و سپس سراغ تصحیح سؤال دوم برای تمام آزمون شوندگان رفت.

1. Hopkins et al
2. Nitko
3. Hallo effect
4. Mechanics effect
5. Order effect
6. Item-to-item Carryover effect

- **انتقال تاثیر یک آزمون به آزمون دیگر[1]:** تصحیح یک آزمون تشریحی به شدت تحت تاثیر آزمون‌هایی است که به فاصله زمانی کوتاهی قبل از آن اجرا شده است. در صورتی‌که دانشجو در آزمون تشریحی قبلی نمره‌ای پایین‌تر از حد انتظار دریافت کرده باشد، آزمون متعاقب آن مطلوب‌تر ارزیابی خواهد شد.

چیس ۱۹۸۶

در این مطالعه تاثیر جنسیت، نژاد، انتظارات ارزیابان و کیفیت دست‌خط آزمون شوندگان در ارتباط با یکدیگر، هنگام تصحیح سؤالات تشریحی مورد بررسی قرار گرفت. در این مطالعه به ۸۰ ارزیاب یک سؤال تشریحی به‌همراه یک فلش‌کارت حاوی اطلاعات دموگرافیک آزمون‌شوندگان (سیاه وست یا سفید پوست، مرد یا زن) ارائه شد. پاسخ‌های ارائه شده یا با دست‌خط خوب یا بد نوشته شده بود. نتایج حاصل از مطالعه نشان‌دهنده تعاملات پیچیده بین متغیرهای جنسیت، دست‌خط و انتظارات ارزیابان در هر نژاد بود.

سودمندی سؤالات تشریحی

روایی سؤالات تشریحی

زمان‌بر بودن پاسخ‌دهی به این سؤالات و دشواری نمره‌دهی آن‌ها باعث شده است که تعداد محدودی از آن‌ها را بتوان در هر آزمون ارزیابی نمود که این موضوع منجر به بروز مشکل ویژگی محتوا و در نتیجه کاهش روایی محتوایی آزمون می‌شود. به عنوان مثال، یک آزمون متشکل از ۸۰ سؤال چندگزینه‌ای نسبت به آزمونی که شامل چهار یا پنج سؤال تشریحی است، پوشش بهتری از محتوای دوره را می‌تواند فراهم سازد. به‌همین دلیل در مواردی که آزمون از تعداد محدودی سؤال تشریحی تشکیل شده است، به ندرت احتمال دارد که بتواند از روایی محتوایی مطلوبی نیز برخوردار باشد. برای حل این مشکل می‌توان آن را به چندین سؤال کوته‌پاسخ تبدیل نمود تا حجم بیشتری از محتوای آموزشی را پوشش داد. همچنین می‌توان از طریق انجام چندین آزمون تشریحی در طول دوره آموزشی، پوشش بهتری از محتوا داشت. به‌علاوه از آنجا که در سؤالات تشریحی، فراگیران خود باید پاسخ‌های محتمل موجود را ارائه نمایند، این دسته از سؤالات در مقایسه با سؤالات بسته‌پاسخ از روایی صوری بالاتری برخوردارند (آلن[2] ۲۰۰۴).

دی و همکاران ۱۹۹۰

این نویسندگان در مطالعه خود به بررسی روایی آزمون تشریحی قضاوت بالینی دستیاران پرداختند. بر اساس تعریف بورد بیماری‌های داخلی آمریکا و مروری بر متون، کمیته ارزیابی ابتدا ۹ معیار برای قضاوت بالینی خوب را تعیین کرد که شامل فاکتورهایی از قبیل توجه به خطرات تشخیص‌های افتراقی و برنامه درمانی، توانایی تصمیم‌گیری در شرایط غیرمنتظره، توجه به عوامل روانشناختی-اجتماعی در زمان تعیین برنامه درمانی بیمار بود. بعد از تدوین بلوپرینت آزمون که شامل اجزاء قضاوت بالینی و محتوای آموزشی مورد نظر بود، موضوعات مورد نظر به شکل نمونه‌گیری منظم در هر یک از حیطه‌های قضاوت بالینی مشخص شدند و سپس موضوعات انتخابی به صورت مورد بالینی ارائه شدند. از دانشجو خواسته شد تا با توجه به مطالب ارائه شده در صورت سؤال به عنوان مثال گزینه‌های درمانی و یا فاکتورهای مهم در اتخاذ یک برنامه درمانی را تعیین کند. سؤالات تشریحی طراحی شده در مطالعه مقدماتی پایلوت شد و در نهایت ۱۲ سؤال از مجموع آن‌ها انتخاب شدند به گونه‌ای که سؤالات انتخابی بتوانند طیف قضاوت بالینی را مورد سنجش قرار دهند.
در مرحله اجرا، ۱۲ سؤال تشریحی در مدت زمان ۳ ساعت و ۸۵ سؤال چندگزینه‌ای برای ۴۷ دانشجوی رزیدنتی سال اول و ۵۱ کارورز از ۱۲ برنامه‌های آموزشی دستیاری در فیلادلفیا و پنسیلوانیا و برنامه‌های آموزشی مراکز پزشکی در دانویل، پنسیلوانیا، اوهایو، و دانشگاه کنتوکی در لکسینگتون و کنتاکی اجرا شد. سؤالات تشریحی با استفاده از دو روش نمره‌دهی ارزیابی شدند: روش کلی و تحلیلی. نمره‌دهی کلی سؤالات تشریحی توسط ۱۲ نفر از متخصصین داخلی صورت می‌گرفت که برای هر سؤال دو نمره کلی تعیین می‌شد که نمرات کلی بر اساس مقیاس لیکرت ۹ تایی تعیین می‌شدند. نمره‌دهی تحلیلی نیز توسط سه عضو هیات علمی که از دانش پزشکی هیچ گونه اطلاعی نداشتند با استفاده از یک چک لیست طراحی شده برای هر سؤال صورت می‌گرفت. همچنین به منظور تعیین ارتباط و همبستگی نتایج حاصل از آزمون تشریحی با سایر روش‌های ارزیابی توانمندی، داده‌های حاصل از ارزیابی عملکردی گردآوردی شد. نتایج حاصل از مطالعه نشان‌دهنده همبستگی نسبتا کم نمرات آزمون تشریحی با سؤالات چندگزینه‌ای بود. اما زمانی که مشکل پایایی نمره‌دهی سؤالات تشریحی به روش نمره‌دهی تحلیلی تصحیح شد این همبستگی افزایش یافت (۰/۸۳ برای نمره-دهی به روش تحلیلی و ۰/۷۶ برای نمره‌دهی به روش کلی). میزان همبستگی بین نمرات آزمون‌های تشریحی و آزمون‌های عملکردی علی‌رغم آنکه کم بود اما این همبستگی معنادار بود. در این مطالعه تناسب محتوای آزمون (روایی محتوایی) به وسیله تعیین محورهای قضاوت بالینی و تعیین بلوپرینت صورت گرفت. به طور خلاصه مشخص شد که آزمون‌های تشریحی می‌توانند به طریقی طراحی شوند که حیطه‌هایی از قضاوت بالینی را پوشش دهند.

1. Test-to-test carryover effect
2. Allen

پایایی سؤالات تشریحی

از جمله انتقاداتی که همواره به سؤالات تشریحی وارد می‌شود، پایایی پایین نمره‌دهی آن‌ها است. طرح این موضوع به شکل رسمی از سال ۱۹۱۶ در پی افزودن یک آزمون تشریحی به آزمون ورودی کالج[1] در انگلیس آغاز شد اما در عین حال مطالعات صورت گرفته در سال ۱۸۸۰ نیز نشان می‌دهد که همواره آزمون‌های تشریحی از لحاظ پایایی پایین مورد انتقاد قرار گرفته‌اند (چیس ۱۹۶۸). به دنبال افزایش نگرانی در خصوص پایایی آزمون‌های تشریحی، سه محقق از بورد کالج انگلیس با بررسی اطلاعات حاصل از امتحانات متعدد بورد در سال ۱۹۴۵ گزارشی ارائه نمودند که نشان‌دهنده پایایی بسیار پایین آزمون‌های تشریحی بود. نتایج حاصل از مطالعات متعدد بعدی نیز نشان‌دهنده پایین بودن پایایی نمرات در سؤالات تشریحی مشابه بود (بال[2] ۱۹۵۷، هلینگورث[3] ۱۹۸۱).

بال ۱۹۵۷

بال در مطالعه خود به این نتیجه رسید که همبستگی بین نمراتی که ارزیابان مختلف به مجموعه مشابهی از سؤالات تشریحی داده بودند، پایین است و حتی در آزمون‌هایی که از الگوهای نمره‌دهی یکسانی به منظور تصحیح سؤالات فراگیران استفاده شده بود، دامنه کاملاً متغیری از رد شدن تا کسب نمره کامل به یک پاسخ مشاهده شده بود. او همچنین در مطالعه خود دریافت که نمرات ارائه شده توسط ارزیابان مشابه در زمان‌های مختلف نیز در اغلب اوقات همبستگی پایینی دارند.

فرنچ و کارلتون[1] ۱۹۶۱

این پژوهشگران از ۵۳ آزمونگر شرکت‌کننده در مطالعه خواستند تا ۳۰۰ سؤال تشریحی را بر اساس ۹ آیتم دسته‌بندی نمایند. سپس با استفاده از تحلیل عاملی، محققین، عملکرد مصححان را در دسته بندی سؤالات تشریحی بر اساس پنج ویژگی (دیدگاه‌ها، استدلال، روش، تمایلات و مکانیسم) به پنج دسته کلی تقسیم نمودند. نتایج حاصل از مطالعه نشان داد که میانگین همبستگی بین این پنج عامل در بین مصححان مختلف ۰/۳۱ بود. این پژوهشگران نتیجه گرفتند که این همبستگی پایین غیرقابل قبول می‌باشد.

1. French & Carlton

پایایی آزمون، در ارتباط مستقیم با تعداد سؤالات آزمون و مدت زمانی است که به آن‌ها اختصاص داده می‌شود. بنابراین یکی دیگر از عواملی که بر پایایی سؤالات تشریحی تاثیرگذار است، زمان‌بر بودن پاسخگویی به این سؤالات است که بالتبع نسبت به سؤالات چندگزینه‌ای تعداد کمتری از این سؤالات را می‌توان در یک آزمون استفاده نمود. این موضوع به نوبه خود باعث کاهش پایایی آزمون خواهد شد (شوورث و ون‌درولوتن[4] ۲۰۰۴). از نظر ون‌درولوتن، عامل موثر دیگر بر پایایی آزمون‌های تشریحی، مدت زمان اختصاص داده شده به کل آزمون است (ون‌درولوتن ۲۰۰۰). با توجه به آنکه پاسخگویی به سؤالات تشریحی انرژی زیادی از فراگیران می‌برد، امکان اجرای آزمون‌های تشریحی طولانی با تعداد زیادی سؤال وجود ندارد که این موضوع منجر به پایایی پایین‌تر این آزمون‌ها در مقایسه با سؤالات چندگزینه‌ای خواهد شد.

سیمز[1] ۱۹۳۱

سیمز در بررسی خود دو فرم موازی از آزمون تشریحی را به ۸۰ نفر از دانشجویانی که در دانشگاه آلاباما دوره آموزشی روانشناسی را می‌گذراند، ارائه نمود. هر فرم آزمون شامل ۱۰ سؤال تشریحی بود که به شکل تصادفی از مجموعه ۲۰ سؤال آماده شده انتخاب شده بود. نتایج حاصل از آزمون‌های تشریحی با آزمون بسته‌پاسخ شامل ۴۰ سؤال درست-نادرست و ۳۴ سؤال جورکردنی مورد مقایسه قرار گرفت. سؤالات تشریحی توسط سه تا شش ارزیاب تصحیح شده بود. ضریب پایایی بین مصححین در آزمون‌های تشریحی ۰/۷۲ بود که در حدود ۰/۱، ضریب پایایی پایین تر از آزمون‌های بسته‌پاسخ داشت. همبستگی بین نمرات آزمون‌های تشریحی و بسته‌پاسخ ۰/۷۰ بود.

1. Sims

1. College Entrance Examination
2. Bull
3. Hollingworth
4. Schuwirth & Van der Vleuten

بریجمن [1] ۱۹۹۲

بر اساس نتایج حاصل از این مطالعه پیشنهاد شده است که اگرچه در سؤالات چندگزینه‌ای به دلیل تاثیر عامل حدس زدن کمتر می‌توان بر پایه نتایج حاصل از یک سؤال، عملکرد فراگیر را در سؤال دیگر پیش‌بینی نمود اما از آنجا که این سؤالات به زمان کمتری برای پاسخ‌دهی و همچنین نمره‌دهی نیاز دارند، در یک آزمون تعداد بیشتری از آنان را می‌توان مطرح کرد. بنابراین این آزمون‌ها از پایایی بالاتری در مقایسه با آزمون‌هایی که تعداد کمتری از سؤالات بازپاسخ را شامل می‌شوند برخوردار هستند.

1. Bridgeman

هیچ آزمونی نمی‌تواند به طور کامل پایا باشد (گرینبرگ[1] ۱۹۹۲). به این دلیل که منابع خطای متعددی در هر موقعیت ارزیابی تاثیرگذار است که از آن جمله به تغییر در عملکرد آزمون‌شوندگان تحت تاثیر عوامل مختلف (بیماری یا خستگی ناشی از کمبود خواب و ...)، شرایط اجرای آزمون (نور ناکافی، فضای فیزیکی نامناسب و ...) و تفاوت در عملکرد ارزیابان به هنگام تصحیح سؤالات (ساده یا سخت‌گیر بودن و ...) اشاره کرده است. از نظر گرینبرگ به منظور افزایش پایایی باید بیشترین تمرکز بر مورد آخر باشد؛ شاید به این دلیل که کنترل بیشتری می‌توان بر آن داشت. از این رو در اکثر مواقع تنها توافق بین ارزیابان به عنوان شاخصی از پایایی در آزمون‌های تشریحی محاسبه می‌شود. هر چند از نظر گرینبرگ هیچگاه نمی‌توان به پایایی ایده‌ال در آزمون های تشریحی دست یافت؛ با این حال، همواره باید سعی نمود که با کنترل عوامل مداخله‌گر به بالاترین میزان پایایی در این آزمون‌ها دست یافت.

مقبولیت سؤالات تشریحی

مطالعات نشان می‌دهند که سؤالات تشریحی معمولاً از مقبولیت پایینی در بین فراگیران برخوردار هستند که دلایل متعددی برای این امر ذکر شده است. از جمله این دلایل می‌توان به زمان‌بر بودن پاسخ‌دهی به این سؤالات، تاثیر دست‌خط بر نمره کسب شده و ذهنی بودن فرایند نمره‌دهی اشاره نمود. البته در برخی از مطالعات، یافته‌های متناقضی در این خصوص ارائه شده است. پارمنتر[2] در مطالعه خود به این نتیجه دست یافت که فراگیران عموماً به سؤالات چندگزینه‌ای تمایل بیشتری دارند اما زمانی که از آمادگی بیشتری در آزمون برخوردار هستند، گرایش آنان به سؤالات تشریحی افزایش می‌یابد (پارمنتر ۲۰۰۹).

پارامنتر ۲۰۰۹

پارمنتر مطالعه‌ای با هدف بررسی تمایل فراگیران به ارزیابی بوسیله سؤالات تشریحی و چندگزینه‌ای انجام داد. داده‌های این مطالعه با استفاده از یک ابزار پرسشنامه‌ای جمع‌آوری شد. ۸۱ نفر از دانشجویان مقطع کارشناسی مدیریت در این مطالعه شرکت داده شدند. از فراگیران خواسته شده بود که به ۱۱ آیتم پاسخ دهند. دو آیتم ابتدایی در خصوص ترجیح کلی آنان نسبت به سؤالات تشریحی و چندگزینه‌ای بود و نُه آیتم بعدی در ارتباط با بررسی منطق پشت تمایلات آنان به انتخاب فرمت آزمون مورد نظر اختصاص داشت. فراگیران به هر یک از آیتم‌های پرسشنامه از کاملاً مخالف تا کاملاً موافق پاسخ دادند. در انتهای پرسشنامه نیز دو سؤال باز طرح شده بود که یکی از آنها در مورد اینکه چگونه دانشجویان برای یک آزمون فرضی تشریحی یا چندگزینه‌ای آماده می‌شوند و دیگری در مورد اینکه در صورت آمادگی کامل در آزمون کدام شکل از سؤالات را ترجیح می‌دهند. همچنین به منظور آزاد گذاشتن دانشجویان در انتخاب نوع سؤال مورد نظر، از آنان در خواست شده بود تا پاسخ‌های خود را در قالب آزمون‌های چندگزینه‌ای، آزمون‌های تشریحی و آزمون‌های ترکیبی (چندگزینه‌ای و تشریحی) ارائه نمایند.
نتایج حاصل از مطالعه نشان داد که به طور کلی از نظر دانشجویان، پاسخ‌دهی به سؤالات چندگزینه‌ای آسانتر است. همچنین یافته‌ها نشان داد که دانشجویان احساس می‌کنند که سؤالات تشریحی سنجش عادلانه‌تری از عملکردشان فراهم می‌سازند و از این رو از روایی بالاتری برخوردارند. به طور خلاصه نتایج این مطالعه مشخص کرد که دانشجویان به ارزیابی به وسیله سؤالات چندگزینه‌ای تمایل دارند و این تمایل به میزان زیادی، در ارتباط با اعتقاد آن‌ها به آسانی این نوع از سؤالات است که زمانی که دانشجویان از آمادگی کافی برای امتحان برخوردار نیستند، این گرایش در آنان تقویت می‌یابد.

1. Greenberg
2. Parmenter

علی‌رغم پایین بودن مقبولیت سؤالات تشریحی در بین فراگیران، این سؤالات مقبولیت بالایی در بین اعضای هیأت علمی دارند. استقبال زیاد اعضای هیأت علمی به استفاده از سؤالات تشریحی در مقایسه با سایر روش‌های ارزیابی تا حدودی مربوط به این است که این نوع سؤالات باعث به چالش کشیدن فراگیران در ارائه پاسخ به جای انتخاب آن از میان تعدادی گزینه می‌شود. به‌علاوه شماری از اعضای هیأت علمی نیز به دلیل آنکه این سؤالات امکان ارزیابی توانایی استدلال، تحلیل، ترکیب و ارزشیابی را فراهم می‌کنند به استفاده از این سؤالات در امتحانات خود روی می‌آورند. در این میان، ساده بودن فرایند طراحی این سؤالات در بین اعضای هیأت علمی، بر نگرش آن‌ها بی‌تاثیر نبوده و بر افزایش مقبولیت این سؤالات می‌افزاید. البته در کنار این این موضوع، باید هزینه و وقت لازم برای تصحیح سؤالات نیز در نظر گرفته شود.

ترکیلماز و همکاران [1] ۲۰۰۸

این مطالعه با هدف بررسی دیدگاه‌های مدرسان زبان ترکی برای ارزیابی و اندازه‌گیری، استراتژی‌های طرح سؤال و استفاده از سؤالات تشریحی برای ارزیابی دانشجویان انجام گرفت. این مطالعه بر روی ۳۸ مدرس زبان ترکی که به شکل تصادفی انتخاب شده بودند انجام گرفت. به منظور جمع‌آوری اطلاعات از پرسشنامه ۵۰ سؤالی استفاده شد و در نهایت با استفاده از آلفای کرونباخ پایایی درونی پرسشنامه ۰/۷۶ تعیین شد. یافته‌های حاصل از مطالعه نشان داد که تقریباً تمامی مدرسان ترجیح می‌دهند که از سؤالات تشریحی به منظور ارزیابی فراگیرانشان استفاده کنند. آسانی طراحی، ارزشیابی و کاربرد این سؤالات به عنوان اصلی‌ترین دلایلی بود که به‌وسیله شرکت کنندگان ذکر شده بود.

1. Turkyilmaz et al.

هزینه و قابلیت اجرای سؤالات تشریحی

نمره‌دهی سؤالات تشریحی نیازمند صرف زمان و نیروی انسانی بسیار است. بنابراین برخلاف آنکه طراحی این سؤالات در مقایسه با برخی سؤالات بسته‌پاسخ آسان‌تر است، اما معمولاً به هزینه بیشتری برای تصحیح برگه‌های امتحانی نیاز است. این موضوع قابلیت اجرای این سؤالات را در مواجه با موقعیت‌هایی که تعداد فراگیران زیاد است و یا بخش وسیعی از محتوای آموزشی باید مورد ارزیابی قرار گیرد، زیر سؤال می‌برد.

تاثیر آموزشی سؤالات تشریحی

رویکرد مطالعه و یادگیری یکی از عوامل مؤثر در پیشرفت تحصیلی فراگیران است که تحت تأثیر عوامل مختلفی از جمله نوع سنجش عملکرد فراگیران قرار دارد. مطالعات مختلف نشان می‌دهد که استفاده از سؤالات تشریحی بر روش مطالعه فراگیران تاثیر مثبت داشته و باعث تشویق آنان به درک عمیق مطالب می‌شود (اسکولر و پروسر ۱۹۹۴، اسکولر ۱۹۹۸، گیبلز[1] ۲۰۰۵). بنابراین در صورتی که فراگیران از ابتدای دوره آموزشی بدانند که در انتهای دوره با استفاده از این نوع سؤالات سنجش خواهند شد، احتمالاً در طول دوره به شکل معنی‌دارتری مطالب را فرا خواهند گرفت و به دنبال درک روابط میان موضوعات خواهند بود. واضح است که در این میان استفاده از استراتژی‌های مفید دیگر از قبیل بکارگیری کتابچه‌های راهنمای مطالعه[2]، ارائه الگوی پاسخ و اجرای کوئیزهای متعدد می‌تواند نقش بسیار مهمی در افزایش آگاهی فراگیران از سطح دانش مطلوب ایجاد نماید (ون‌درولوتن و دریسن[3] ۲۰۰۰). به‌علاوه، از آنجا که در سؤالات تشریحی این فرصت برای مصحح فراهم است که علت نادرست بودن پاسخ‌های ارائه شده توسط فراگیر را بیان نماید، این موضوع خود می‌تواند تاثیر آموزشی زیادی برای فراگیر به‌همراه داشته باشد.

1. Gijbels
2. Study guides
3. Driessen

شکورنیا و همکاران ۱۳۹۲

این پژوهش نیمه‌تجربی با هدف بررسی تأثیر آزمون‌های چندگزینه‌ای و تشریحی بر رویکرد مطالعه و راهبردهای آماده شدن برای امتحان روی ۱۲۴ نفر از دانشجویان دانشکده پرستاری و مامایی دانشگاه علوم پزشکی جندی شاپور اهواز در ترم اول سال ۹۱-۹۰ انجام شد. نمونه‌ها به طور تصادفی به دو گروه دانشجویان مشمول سنجش با سؤالات چندگزینه‌ای یا تشریحی تقسیم شدند. ابزار گردآوری اطلاعات دو پرسشنامه رویکردهای مطالعه (۳۲ سؤال) و راهبردهای آماده شدن برای امتحان (۲۸ سؤال) بود که روایی و پایایی آنها بررسی و تأیید شد. دانشجویان پرسشنامه‌ها را در دو نوبت، یک بار در اولین جلسه کلاس و بار دوم در پایان ترم تکمیل کردند. برای تجزیه و تحلیل داده‌ها از آزمون Paired T test استفاده شد. میانگین نمره رویکرد عمقی مطالعه دانشجویان، مشمول روش سنجش تشریحی در ابتدای ترم ۳/۷۰±۰/۳۷ و انتهای ترم ۳/۷۵±۰/۳۳ بود که تفاوت معنادار آماری داشت (T=۲/۷۳, P=۰/۰۰۸). مقایسه میانگین نمرات رویکردهای مطالعه دانشجویان به تفکیک جنسیت نشان داد که بین نمره رویکرد عمقی مطالعه دانشجویان دختر در روش سنجش تشریحی در ابتدا (۳/۶۶±۰/۳۷) و انتهای ترم (۳/۷۳±۰/۳۴) تفاوت معنادار وجود دارد (T=۳/۵۰, P =۰/۰۰۱). بین میانگین نمرات رویکردهای سطحی مطالعه دانشجویان پسر در روش سنجش چندگزینه‌ای در ابتدا (۳/۳۴±۰/۴۶) و انتهای ترم (۳/۶۳±۰/۶۰) نیز تفاوت معنادار آماری مشاهده شد (T=۲/۳, P=۰/۰۵). نتایج پژوهش نشان داد که استفاده از سؤالات تشریحی دانشجویان را به سمت رویکرد عمقی مطالعه سوق می‌دهد. پژوهشگران در پایان توصیه کردند که استفاده بیشتر از سؤالات تشریحی در فرایند ارزیابی به تقویت رویکردهای عمقی مطالعه در فراگیران کمک می‌نماید.

باورهای نادرست در مورد سؤالات تشریحی

ارزیابی ذاتی سطوح بالای حیطه شناختی

یکی از تصورات نادرست رایج در مورد سؤالات تشریحی آن است که این سؤالات ذاتاً توانایی سنجش سطوح بالای حیطه شناختی را دارند. در حالی‌که بر خلاف این تصور اشتباه، در صورت طراحی نامناسب، این سؤالات تنها به ارزیابی سطوح پایین شناختی خواهند پرداخت. به‌علاوه، در صورتی که ارزیاب تنها به دانسته‌ها و حقایق ارائه شده توسط فراگیران نمره دهد و به شیوه و ساختار ترکیب مطالب توجهی نداشته باشد، در این حالت نیز سؤال طراحی شده موثر نخواهد بود. بنابراین هیچ‌گاه نمی‌توان ادعا نمود که استفاده از شکل سؤالات تشریحی، تضمینی برای ارزیابی سطوح بالای حیطه شناختی است، بلکه آنچه که در این میان نقش کلیدی دارد، انتخاب محتوای مورد ارزیابی و طراحی مناسب سؤال است. برای درک بهتر موضوع به مثال‌های زیر توجه کنید.

نمونه سؤال تشریحی ضعیف

محاسن و معایب دریچه‌های بیولوژیک را شرح دهید.

سؤال بهتر

با توجه با محاسن و معایب دریچه های بیولوژیکی و مصنوعی، در یک بیمار با نقص پلاکت استفاده از چه نوع دریچه‌ای توصیه می‌شود؟

طراحی آسان سؤالات تشریحی

با وجود آنکه طراحی سؤالات تشریحی در مقایسه با سؤالات چندگزینه‌ای به دلیل عدم نیاز به طراحی گزینه‌های انحرافی آسان‌تر است اما این بدان معنی نیست که طراحی سؤال تشریحی خوب (منطبق با معیارهای استالینکر) کاری آسانی است. برای کسب اطلاعات بیشتر در این مورد به قسمت ساختار سؤالات تشریحی مراجعه کنید.

کاهش احتمال حدس زدن

یک باور عمومی وجود دارد مبنی بر اینکه سؤالات چندگزینه‌ی امکان پاسخ حدسی را برای دانشجو فراهم می‌کنند و موجب افزایش نمره وی می‌شوند در حالی که امکان ندارد دانشجو بتواند سؤال تشریحی را با حدس پاسخ دهد. در خصوص حدس در سؤال چندگزینه‌ای به صورت مفصل در بخش دوم کتاب بحث شد اما در مورد سؤال تشریحی باید گفت که

این سؤالات باعث تغییر شکل حدس زدن و نه حذف آن می‌شوند. استفاده از سؤالات تشریحی باعث تشویق فراگیران به گزافه‌گویی می‌شود که در واقع شکلی از حدس زدن برای ارائه پاسخ است (رینر[1] ۲۰۰۲). تعدادی از فراگیران با استفاده از روش‌های مختلف از قبیل استفاده از جملات مبهم و نامفهوم، حشوگویی و ... سعی می‌کنند تا ارزش پاسخ ارائه شده را در نظر ارزیاب بیشتر جلوه دهند تا به این طریق بتوانند نمره بالاتری را در آزمون کسب کنند. در واقع، گزافه‌گویی شامل مجموعه استراتژی‌هایی است که دانشجویان به کار می‌گیرند تا بتوانند از طریق آن به نحوی میزان نمره خود را در آزمون بالا ببرند. گزافه‌گویی زمانی بیشترین تاثیر را بر نتایج آزمون به‌همراه خواهد داشت که ارزیاب، الگوی پاسخی برای تصحیح سؤالات بازپاسخ تهیه نکرده باشد (گرانلاند[2] ۱۹۹۸). در این شرایط فراگیر سعی می‌کند با حدس زدن پاسخ‌های احتمالی و تشریح آن در قالب جملات مبهم و پیچیده، آن‌ها را به گونه‌ای ارائه نماید که در نظر مصحح پاسخ درست تلقی شود.

تلاش بیشتر برای آمادگی در آزمون

با اینکه برخی پژوهش‌ها نشان می‌دهند فراگیران برای آزمون‌های تشریحی در مقایسه با سؤالات چندگزینه‌ای آمادگی بیشتری کسب می‌کنند، در مقابل شمار دیگری از مطالعات، یافته‌های متناقضی را در این خصوص ارائه نموده‌اند. کروک[3] از طریق مرور گسترده بر مستندات موجود و مطالعه بر روی این موضوع به این نتیجه رسید که انتظارات فراگیران از سطح شناخت و محتوای یادگیری بیشتر تحت تاثیر مهارت‌های مطالعه آنان است تا شکل و نوع آزمونی که از آنان به عمل می‌آید. از نقطه نظر کروک، فراگیران بیشتر بر اساس انتظارات اعضای هیأت علمی خود (استدلال و تحلیل مطالب، وسعت و عمق یادگیری موضوعات) برای امتحان آماده می‌شوند تا بر اساس نوع آزمونی که از آنها به عمل می‌آید (کروک ۱۹۸۸).

سؤالات رایج در مورد سؤالات تشریحی

استفاده از کدام یک از روش‌های تصحیح در سؤالات تشریحی بهتر است؟

در این خصوص گفته می‌شود که هر چند روش تصحیح «کلی» آسان‌تر و سریع‌تر از روش «تحلیلی» است اما انجام قضاوت کلی در این روش برای برخی از آزمونگران یک فعالیت ذهنی بسیار انرژی‌بر است (سیف ۱۳۹۴).

به‌علاوه در صورتی که آزمون‌شوندگان علت دریافت نمره خود را جویا شوند، در روش کلی، نه تنها هیچ توجیه روشنی برای متقاعد کردن آن‌ها نمی‌توان ارائه داد، بلکه درک اینکه کدام بخش از پاسخ آنها نادرست بوده است نیز برای آن‌ها امکان‌پذیر نیست.

مسأله دیگری که همواره به عنوان یک انتقاد به روش کلی وارد می‌شود این است که در صورت ارزیابی مجدد پاسخ‌ها توسط مصححان مختلف و در زمان‌های گوناگون نمی‌توان انتظار داشت که پایایی نمرات حفظ گردد. پایین بودن پایایی نمرات مصححان در روش کلی را ناشی از تاثیر سه عامل ذکر کرده‌اند (چارنی[4] به نقل از کافمن[5] ۱۹۷۱):

- ممکن است مصححان در کاربرد استاندارد کلی نمره‌دهی متفاوت از یکدیگر عمل کنند. یک مصحح ممکن است فرد آسان‌گیری باشد و در مجموع نمرات بالایی به پاسخ‌ها بدهد در حالی‌که دیگری به دلیل سخت‌گیر بودن نمرات پایین‌تری برای همان پاسخ‌ها در نظر بگیرد.
- ممکن است دامنه توزیع نمرات مصححان، متفاوت از یکدیگر باشد. نمرات تعیین شده توسط تعدادی از مصححان ممکن است بر یک دامنه بسیار کوچکی متمرکز شود. این در حالی است که سایرین ممکن است دامنه نمرات‌شان گسترده‌تر باشد.

1. Reiner
2. Gronlund
3. Crook
4. Charney
5. Coffman

- ممکن است مصححان معیارهای متفاوتی را در نمره‌دهی به کار برند.

با توجه به موارد ذکر شده برخی از مطالعات استفاده از روش تحلیلی را به دلیل عدم وجود مشکلات فوق توصیه می‌کنند.

دی و همکاران ۱۹۹۰

این پژوهشگران در مطالعه خود به این نتیجه رسیدند که در تصحیح آزمون‌ها با روش نمره‌دهی تحلیلی، تنها ۶ درصد واریانس نمرات کل آزمون به سؤالات تشریحی اختصاص یافت در حالی‌که در روش نمره‌دهی کلی این میزان به حدود ۵ درصد کاهش می‌یابد. این موضوع می‌تواند مطرح‌کننده این حقیقت باشد که تصحیح سؤالات تشریحی به روش کلی، اطلاعات مختصری در خصوص تغییرات عملکرد آزمون‌شوندگان فراهم می‌سازد.

نورسینی ۱۹۹۰

مطالعه انجام شده به وسیله بورد بیماری‌ها داخلی آمریکا نشان داد که با وضع معیارهای نمره‌دهی مشخص و استاندارد از طریق روش نمره‌دهی تحلیلی، ضریب تعمیم‌پذیری نتایج یک آزمون تشریحی ۱۲ سؤالی در حدود ۰/۳۶ بود. بر طبق این نتایج برآورد شد که به منظور دستیابی به سطح قابل قبولی از ضریب تعمیم‌پذیری (حدود ۰/۸۰) به تعداد ۷۲ سؤال تشریحی با مدت زمان تقریبی ۱۸ ساعت آزمون نیاز خواهد بود. با استفاده از روش تصحیح کلی، ضریب تعمیم‌پذیری بهبود یافت اما همچنان به حدود ۲۲ سؤال با مدت زمان تقریبی ۵/۵ ساعت نیاز خواهد بود تا ضریب تعمیم پذیری به حد قابل قبول ۰/۸۰ برسد.

توصیه دیگر این است که با وجود آنکه از هر دو روش تحلیلی و کلی می‌توان در آزمون‌های هنجاری و معیاری استفاده کرد اما روش تحلیلی برای آزمون‌های معیاری و روش کلی برای آزمون‌های هنجاری بهتر است (سیف ۱۳۹۴). به‌علاوه، در خصوص استفاده از روش کلی برای آزمون‌های هنجاری پیشنهاد می‌شود که به منظور ایجاد یک دید کلی در مورد کیفیت پاسخ‌های ارائه شده توسط آزمون‌شوندگان ابتدا باید مجموعه‌ای از پاسخ‌های ارائه شده را مطالعه نمود و سپس درباره کیفیت تک‌تک سؤالات قضاوت کرد.

هانتر و همکاران (۱۹۹۶) نیز معتقدند که استفاده از روش تحلیلی در دوره‌های آموزشی محتوا–محور[1] پیشنهاد می‌شود. در مقابل، به دنبال اصلاحات برنامه‌های درسی در تاکید به ارزیابی مهارت‌های سطوح بالای شناختی دانشجویان در فرایند استدلال و حل مساله، استفاده از روش کلی به دلیلی فراهم‌سازی فرصتی جهت ارزیابی محتوای ادغام‌یافته و ترکیبی توصیه می‌شود.

علی‌رغم یافته‌های متناقض موجود، ویرسما و جونز[2] در مطالعه خود بهترین شیوه تصحیح را ترکیبی از دو روش فوق معرفی نموده‌اند. به طوری که ابتدا سعی شود با مطالعه تمام پاسخ‌ها یک دید کلی در خصوص کیفیت پاسخ‌های ارائه شده کسب شود و سپس با استفاده از روش تحلیلی جزئیات سؤالات را ارزیابی نمود. ویل و هادسون[3] (۱۹۸۳) نیز بر اساس نتایج حاصل از مطالعه خود پیشنهاد دادند که پایایی قابل‌قبول نمره‌دهی (۰/۷۹)، در صورت استفاده از سه روش نمره‌دهی کلی، تحلیلی و ویژگی‌های اصلی به صورت همزمان حاصل می‌شود. کلماز[4] (۱۹۸۲) نیز معتقد است که استفاده از روش ترکیبی نمره‌دهی کلی و تحلیلی به پایایی بالاتر بین ارزیابان می‌انجامد.

این اعتقاد وجود دارد که استفاده از روش ترکیبی باعث رفع کاستی‌های مربوط به هر یک از دو روش می‌شود اما قطعاً توجه به این نکته ضروری است که انجام این کار به زمان و انرژی بیشتری نیاز خواهد داشت که از نظر اجرایی ممکن است امکان‌پذیر نباشد به ویژه زمانی که تعداد فراگیران و سؤالات آزمون زیاد است.

1. Content-driven courses
2. Vircima & Jones
3. Veal and Hudson
4. Quellmalz

دی و همکاران ۱۹۹۰

در این مطالعه، روایی آزمون تشریحی در ارزیابی قضاوت بالینی دستیاران بررسی شد. ۱۲ سؤال تشریحی در مدت زمان ۳ ساعت و ۸۵ سؤال چندگزینه‌ای برای ۴۷ دستیار سال اول و ۵۱ کارورز از ۱۲ برنامه آموزشی دستیاری اجرا شد. سؤالات تشریحی با استفاده از دو روش کلی و تحلیلی تصحیح شدند. تصحیح کلی سؤالات توسط ۱۲ نفر از متخصصین داخلی صورت گرفت. بدین ترتیب که برای هر سؤال دو نمره کلی بر اساس مقیاس لیکرت ۹ تایی تعیین شد. تصحیح تحلیلی نیز توسط سه نفر از اعضای هیات علمی با استفاده از یک چک‌لیست صورت گرفت. بررسی همبستگی بین نمرات حاصل از آزمون تشریحی و چندگزینه‌ای نشان داد زمانی که نمره‌دهی سؤالات تشریحی به روش نمره‌دهی تحلیلی انجام شد، همبستگی افزایش یافت (۰/۸۳ برای نمره‌دهی به روش تحلیلی و ۰/۷۶ برای تصحیح به روش کلی).

تاثیر عوامل مداخله‌گر در نمره‌دهی سؤالات تشریحی به چه صورت است؟

یکی از سؤالاتی که سال‌ها ذهن ارزیابان و طراحان سؤال را به خود مشغول کرده است، این است که در نمره‌دهی سؤالات تشریحی از تاثیر چه عوامل مداخله‌گری باید اجتناب نمود. به صورت کلی فرض می‌شود که هر امری در آزمون‌های تشریحی که توجه ارزیاب را از ارزیابی صحیح محتوای سؤال دور نماید، اثرات بالقوه‌ای روی نمرات داده شده به فراگیران خواهد داشت (چیس ۱۹۶۸). در مطالعات صورت گرفته توسط چیس (۱۹۷۹) نشان داده شد که کیفیت دست‌خط آزمون شوندگان به طور مشخص بر روی نمرات سؤالات تشریحی تاثیرگذار است. او در مطالعه خود از این پیش‌فرض پشتیبانی کرد که وجود شرایطی از قبیل عدم رعایت نکات گرامری منجر می‌شود تا ارزیابان به ساختار فیزیکی پاسخ ارائه شده توجه بیشتری نماید که این موضوع قطعاً از میزان توجه آنان به محتوای پاسخ ارائه شده خواهد کاست و در نهایت باعث می‌شود تا این دسته از آزمون‌شوندگان نمرات پایین‌تری را در مقایسه با افرادی که پاسخ‌های آنان فارغ از هر گونه خطای گرامری است کسب نمایند. سیمز[1] برای کاهش احتمال بروز این حالت پیشنهاد داد از یک الگوی پاسخ استفاده شود تا توجه ارزیاب به سمت محتوای پاسخ ارائه شده متمرکز شود و اثرات سایر عوامل مداخله‌گر به این طریق به حداقل ممکن کاهش یابد. به صورت کلی، مطالعات تجربی در طی سال‌های مختلف نشان می‌دهد که تغییرات زیادی در زمینه تاثیر عوامل مداخله گر در تصحیح سؤالات تشریحی صورت گرفته است. در گذشته به ویژه تا پیش از جنگ جهانی دوم، اعضای هیأت علمی باید تاکید بیشتری بر روی دست‌خط آزمون‌شوندگان می‌داشتند و این عامل تعمداً در فرایند نمره‌دهی آنان دخالت داده می‌شد (چیس ۱۹۶۸). امروزه این موضوع از اهمیت مشابهی در برنامه‌های درسی برخوردار نیست و به ارزیابان در سطح وسیع توصیه می‌شود تا آن را به عنوان یک عنصر در نمره‌دهی سؤالات در نظر نگیرند. همچنین ارزیابان تشویق می‌شوند که به سایر عوامل تاثیرگذار در نمره‌دهی سؤالات تشریحی که نشان‌دهنده بخشی از عملکرد معیار نیست، توجهی نداشته باشند. از این رو با نگاهی به مطالعات مختلف مشخص می‌شود که به موازات تغییرات صورت گرفته در برنامه‌های درسی و گسترش دانش مربوط به بهبود شیوه‌های نمره‌دهی، عوامل موثر بر آزمون‌های تشریحی و ارتباط‌شان با نمرات آزمون شوندگان تغییر کرده است.

آیا شیوه تصحیح سؤالات تشریحی گسترده‌پاسخ با محدودپاسخ متفاوت است؟

یکی از عواملی که در انتخاب روش تصحیح باید لحاظ کرد، نوع سؤال تشریحی است. به طور کلی توصیه می‌شود که پیش از شروع فرایند تصحیح سؤالات تشریحی محدودپاسخ، یک الگوی پاسخ طراحی گردد. در این حالت می‌توان از دو استراتژی برای تهیه الگوی پاسخ استفاده کرد. در روش اول می‌توان یک الگوی پاسخ کلی را تهیه کرد و سپس نمره مشخصی را به اجزا و ویژگی‌های مختلف آن تخصیص داد. در روش دوم می‌توان یک الگوی پاسخ عالی[2]، یک الگوی

1. Sims
2. Perfect

پاسخ خوب[1]، یک الگوی پاسخ قابل قبول[2] و در نهایت یک الگوی پاسخ غیر قابل قبول[3] تهیه کرد و سپس بر اساس آن سؤالات را ارزیابی نمود. البته روش اول از قابلیت اجرای بالاتری برخوردار است به ویژه زمانی که تعداد سؤالات آزمون زیاد است.

در سؤالات تشریحی گسترده‌پاسخ، طراحی الگوی پاسخ دقیق و ساختارمند چندان مناسب نیست چرا که هدف اصلی از این سؤالات عمدتاً شناسایی قدرت نوآوری و خلاقیت فراگیران در ارائه پاسخ است. بنابراین استفاده از یک الگوی پاسخ مشخص منجر به سرکوب بروز خلاقیت و عدم توجه به ارزیابی این مهم است. در این نوع سؤالات توصیه می‌شود که یک نظام نمره‌دهی طراحی شود که شامل چندین معیار (معمولاً بین ۳ تا ۵ معیار) است. این معیارها بر اساس سطوح مختلف مهارت و تخصص فراگیران توصیف می‌شوند و سپس برای هر یک از این سطوح امتیازی در نظر گرفته می‌شود (استیگنز[4] ۱۹۹۷).

آیا باید به دانشجویان حق انتخاب سؤال را از بین مجموعه‌ای از سؤالات تشریحی داد؟

همان‌طور که در مبحث راهنمای عملی طراحی سؤالات تشریحی نیز گفته شد، به منظور اجتناب از اتلاف وقت دانشجویان برای انتخاب سؤالات، عادلانه بودن مقایسه عملکرد فراگیران با یکدیگر و شناسایی دقیق مشکلات یادگیری فراگیران بهتر است از طرح سؤالات انتخابی در آزمون ممانعت کرد. همچنین از آنجا که طراحی سؤالات همسنگ از نظر درجه دشواری کار تقریباً غیرممکنی است، معمولاً فراگیران به پاسخگویی سؤالاتی روی می‌آورند که آسان‌تر هستند. لذا در نهایت انجام این کار منجر به کاهش روایی آزمون خواهد شد. می‌توان گفت که دادن حق انتخاب به آزمون شوندگان در سؤالات تشریحی، مشکل نمونه‌گیری را به دلیل گرایش دانشجویان به پاسخگویی به سؤالاتی که بیشتر می‌دانند، حادتر می‌کند. لاکل[5] و همکاران در مطالعه خود به این موضوع اشاره کردند که دادن حق انتخاب به فراگیران برای پاسخگویی به مجموعه‌ای از سؤالات تشریحی باعث می‌شود که نتوان برآورد دقیقی از میزان یادگیری آنان به عمل آورد. آنان معتقد هستند که انجام این کار باعث می‌شود که فراگیران از پاسخ دادن به سؤالاتی که به خوبی فرا نگرفته‌اند، ممانعت نمایند و بدین ترتیب قاعدتاً نمی‌توان به شناسایی مشکلات یادگیری آنان پرداخت.

در واقع تعدادی از اعضای هیأت علمی زمانی به این راهکار روی می‌آورند که فراگیران به آن‌ها در خصوص دشواری نامعقول آزمون اعتراض می‌کنند و آنان به این طریق سعی می‌کنند تا اعتراض دانشجویان را برطرف کنند. اما نکته کلیدی در این مواقع آن است که با شناسایی دقیق اعتراضات دانشجویان، می‌توان به رفع نواقص موجود و طراحی سؤالات بهتر روی آورد تا اینکه به طرح آزمون‌های فاقد اعتبار پرداخت.

1. Good
2. Satisfactory
3. Unsatisfactory
4. Stiggins
5. Lukhele

منابع

1. Allen, Mary J. Assessing academic programs. Boston: Anker Publishing, 2004.
2. Alwardy NM. Assessment methods in undergraduate medical education. QU Medical Journal 2010; 10, 2:203-9
3. Boulet JR, Friedman Ben-David M, Ziv A, Burdick WP, Gary NE. use of holistic scoring for post-encounter written exercises. In D. Melnick (ed.), Proceedings of the Eighth Ottawa Conference of Medical Education and Assessment 2000
4. Bridgeman BA. comparison of quantitative questions in open-ended and multiple choice formats Journal of Educational Measurement 1992;29(3), 253-271
5. Bull GM. An examination of the final examination in medicine. The Lancet 1956; 268(6939): 368-372.
6. Charney D. The validity of using holistic scoring to evaluate writing: A critical overview. Research in the Teaching of English 1984; 65-81.
7. Chase CT. The impact of some obvious variables on essay test scores. Journal of educational measurement 1968; 5(4): 315-318.
8. Chase CT. Essay test scores and reading difficulty. Journal of Educational Measurement 1983; 20(3): 293-297
9. Chase CT. Essay test scoring: interaction of relevant variables. Journal of Educational Measurement 1986; 23(1): 33-41
10. Chase, Clinton I. Essay test scoring: Interaction of relevant variables. Journal of Educational Measurement 1986; 33-41.
11. Chase C. Essay test scoring: Expectancy and handwriting quality. Psychology: A Journal of Human Behavior 1990; 28(1): 38-41.
12. Cashin WE. Improving essay tests. Idea Paper No. 17. Manhattan 1987; KS: Center for Faculty Development and Evaluation, Kansas State University.
13. Day SC, Norcini J, Disernes D, Cebul R, Schwartz S, Beck LH, Webster GD, Schnabel TG, Elstein A. The validity of an essay test of clinical judgment. Academic Medicine 1990; 65(9): 39-40.
14. Paul DB, French JW, Carlton ST. Factors in judgments of writing ability. ETS Research Bulletin Series 1961; 1961(2): i-93.
15. Driessen E, van der Vleuten C. Matching student assessment to problem-based learning: Lessons from experience in a law faculty. Studies in Continuing Education 2000; 22(2): 235-248
16. Dunn L, Morgan C, Reilly M, Parry SH. The student assessment handbook. Routledge Flamer publisher, 2004
17. Gijbels D. The relationship between students' approaches to learning and the assessment of

learning outcomes. European Journal of Psychology of Education 2005 20(4): 327-341.

18. Greenberg K. Validity and reliability issues in the direct assessment of writing. WPA: Writing program administration 1992; 16(1-2): 7-22.
19. Gronlund NE. Assessment of student achievement. Allyn & Bacon Publishing, Longwood Division, 160 Gould Street, Needham Heights, MA 02194-2310; tele, 1998.
20. Hollingsworth HL. Specialized vocational test sand methods. School and Society 1981; China's Examination Hell (New Haven: Yale University Press)
21. Hopkins KD, Stanley JC, Hopkins BR. Constructing and using essay tests. Educational and psychological measurement and evaluation 1990; 193-223.
22. Hunter, Darryl M., Richard M. Jones, and Bikkar S. Randhawa. The use of holistic versus analytic scoring for large-scale assessment of writing. The Canadian Journal of Program Evaluation 1996l 11(2): 61-85.
23. Kazemi A, Ehsanpour S. Item analysis of core theortical course exams for midwifery students in Isfahan University of Medical sceinces. Iranian journal of Medical Eduaction 2010; 10(5): 643-50.
24. King JL. The effects of gender bias and errors in essay grading. Educational Research Quarterly 1998; 22: 13–25
25. Lipton A, Huxham GJ. Comparison of multiple-choice and essay testing in preclinical physiology. Medical Education 1970; 4(3): 228–238.
26. Lukhele R, Thissen D, Wainer H. On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests. Journal of Educational Measurement 1994; 31(3): 234-250
27. MCC guideline: Guide to Writing Exam Questions for Entry into Questionbank. 2011. Available at:http://www.med.uottawa.ca/Students/MD/assets/documents/CDM_Guidelines e.pdf
28. Mohammadi H, Javadian Y, Nikbakhsh N, Javanian M, Jahanian Z. Effects of self assessment of the examination questions by educational groupe on the quality of final exams. J Babol University of Medical Sciences 2010; 12(suppl1): 65-69.
29. Moss, Andrew, and Carol Holder. Improving Student Writing: A Guidebook for Faculty in All Disciplines. Kendall/Hunt Publishing Co., 2460 Kerper Blvd., Dubuque, IA 52001, 1988.
30. Nieswandt M, Bellomo K. Written Extended-Response Questions as Classroom Assessment Tools for Meaningful Understanding of Evolutionary Theory. Journal of research in science teaching 2009; 46,(3): 333–356
31. Nitko AJ. Educational assessment of students. Prentice-Hall, Inc., PO Box 11071, Des Moines, IA 50336-1071, 2001.
32. Norcini J, Disernes D, Day SC, Cebul R, Schwartz S, Beck LH, Webster GD, Schnabel TG, Elstein A. The scoring and reproducibility of an essay test of clinical judgment. Academic Medicine 1990; 65(9): 41-42.

33. Nunnink L, Venkatesh B, Krishnan A, Vidhani K, Udy A. prospective comparison between written examination and either simulation-based or oral viva examination of intensive care trainees' procedural skills. Anaesth Intensive Care 2010; 38: 876-882
34. Parmenter DA. Essay versus multiple-choice: student preferences and the underlying rationale with implications for test construction. Academy of Educational Leadership Journal 2009; 13(2): 1831-1847.
35. Powers DE, Fowles ME, Farnum M, Ramsey P. Will they think less of my handwritten essay if others word process theirs? Effects on essay scores of intermingling handwritten and word-processed essays. Journal of Educational Measurement 1994; 31(3): 220-233.
36. Purdie N, Hattie J, Douglas G. Student conceptions of learning ahte their use of self-regulated learning strategies: A cross-cultural comparison. Journal of Educational Psychology 1996; 88(1): 87-100.
37. Quellmalz ES. Designing writing assessments: Balancing fairness, utility and cost. 1982. Los Angeles: Centre for the Study of Evaluation, University of California.
38. Rassaian N, Nakhaei S, Sadeghi Ghandehari N. Camparison of three exam techniques in medical students: multiple choices, true-false and short answer questions. J of Hakim 2001; 5(4): 271-80.
39. Reiner CM, Bothell TW, Sudweeks RR, Wood B. Preparing effective essay questions. 2002; New Forums Press.
40. Rushton P, Eggett D. Comparison of Written and Oral Examinations in a Baccalaureate Medical-Surgical Nursing Course. Journal of Professional Nursing 2003; 19,(3): 142-148
41. Schuwirth LWT, van der Vleuten CP. ABC of learning and teaching in medicine: Written assessment. BMJ 2003; 326:643–5
42. Schuwirth LWT, van der Vleuten CP. ABC of learning and teaching in medicine: Written assessment. BMJ 2003; Schuwirth LW, van der Vleuten CP. Different written assessment methods: what can be said about their strengths and weaknesses? Medical Education 2004; 38:974–9
43. Scouller K. The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay Higher Education 1998; 35: 453-472.
44. Scouller KM. Prosser M. Students' experiences in studying for multiple choice question examinations. Studies in Higher Education 1994; 19(3): 267-279.
45. Sims VM. The objectivity, reliability and validity of an essay examination graded by rating. Journal of Educational Research 1931; 24: 216-223.
46. Skogedal O, Lauvas P. classification of written examinations in dental education in Norway according to educational aims. Journal of dental research 1976; 84: 438-442
47. Stalnaker, John M. The essay type of examination. Educational measurement 1951; 495-530.
48. Stecher BM, Rahn ML, Ruby A, Naomi Alt M, Robyn A. Using Alternative Assessments in Voca-

tional Education. 1997. Rand publishing

49. Stiggins RJ. Student-centered classroom assessment (2nd ed.) 1997; Upper Saddle River, NJ: Merrill.
50. Turkyilmaz M. The opinions of teachers about the usage of essay examinations as a measurement tool in language and expression course. Ahi Evran niversitesi Kirsehir Egitim Fakültesi Dergisi (KEFAD) 2008; 9, 3: (1-14)
51. Van Hoeij M, Haarhuis J, Wierstra R, Van Beukelen P. Developing a classification tool based on bloom's taxonomy to assess the cognitive level of short essay questions. JVME 2004; 31(3): 261-267.
52. Veal LR, Hudson SA. Direct and indirect measures for large-scale evaluation of writing. Research in the Teaching of English 1983;17(3), 290–296.
53. Venhoeven BH, Hamers J, Scherpbier A, Hoogenboom R, Van der Vleten CP. The effect on reliability of adding a separate written assessment component to an objective structured clinical examination. Medical Education 2000; 34: 525–529
54. Verma M, Chhatwala J, Singha T. Reliability of Essay Type Questions: effect of structuring. Assessment in Education Principles Policy and Practice 1997; 4:265-270.
55. Vleuten CP, van Luijk SJ, Beckers HJ. A written test as an alternative to performance testing. Medical Education 1989; 23: 97–107

۵۶. رساییان ن، نخعی س، صادقی قندهاری ن. مقایسع روش های آزمون های چندگزینه ای، صحیح غلط و کوته‌پاسخ در دانشجویان پزشکی، مجله ایرانی آموزش در علوم پزشکی ۱۳۸۱؛ ۵ (۴): ۲۷۱–۲۷۸.

۵۷. سیف ع.ا. اندازه گیری،سنجش و ارزشیابی آموزشی، نشر دوران، ۱۳۹۴، صفحه ۶۹۴

۵۸. شکورنیا ع، علیخانی ه، نجار ش، کمیلی ثانی ح، الهام پور ح. تاثیر دو روش آزمون تشریحی و چندگزینه‌ای بر رویکردهای مطالعه و راهبردهای آماده شدن برای مطالعه. مجله ایرانی آموزش در علوم پزشکی ۱۳۹۲؛ ۱۳ (۴): ۳۰۶–۳۱۸

فصل ۱۱

سؤالات تشریحی تغییر یافته

ساختار سؤالات تشریحی تغییریافته

سؤالات تشریحی تغییریافته انواعی از سؤالات بازپاسخ هستند که امکان ارزیابی توانایی سطوح مختلف یادگیری از فراخوانی مطالب گرفته تا ارزشیابی و حل مسأله را در جایی که استفاده از بیماران شبیه‌سازی شده یا واقعی مقدور نباشد، به صورت کتبی فراهم می‌سازند. از این سؤالات برای اولین بار در انگلیس به عنوان بخشی از امتحانات تایید صلاحیت پزشکان عمومی استفاده شد (رابینوویتز[1] ۱۹۸۷). در برخی از منابع، این نوع سؤالات ترکیبی از سؤالات تشریحی و سؤالات چندگزینه‌ای در نظر گرفته می‌شوند (پالمر و دویت[2] ۲۰۰۷).

سؤال تشریحی تغییریافته بر اساس یک سناریو یا مورد بالینی مطرح می‌شود که معمولاً مرحله به مرحله معرفی می‌گردد و در هر مرحله از دانشجو خواسته می‌شود به سؤال یا سؤالاتی که از جنبه‌های مختلف به مورد بالینی پرداخته‌اند، پاسخ دهد. دفترچه سؤالات اغلب شامل چندین صفحه و معمولاً بین ۶ تا ۱۰ صفحه است. در صفحه اول، راهنمایی‌های لازم در خصوص آزمون از قبیل تعداد سؤالات، زمان آزمون و نکات مهم ارائه می‌شود. در صفحه دوم، یک مورد بالینی به شکل سناریوی کوتاه بیان می‌شود و متعاقب آن تعدادی سؤال مرتبط مطرح می‌گردد. در صفحه‌ی بعد بخش دیگری از مشکل بالینی ارائه می‌شود و سؤالات دیگری در رابطه با آن از فراگیر پرسیده می‌شود.

در این آزمون‌ها امکان برگشت به سؤالات ابتدایی و تغییر پاسخ‌های ارائه شده به سؤالات قبلی وجود ندارد زیرا ممکن است با پیشرفت سناریو و مطرح شدن ابعاد جدیدی از مشکل، پاسخ قسمت‌های قبلی در سؤالات بعدی ارائه شود. به همین دلیل از فراگیران خواسته می‌شود که در ارائه پاسخ‌های خود دقت کافی داشته باشند. مسأله مهم در این سؤالات، احتمال بروز خطای تجمعی[3] است. خطای تجمعی زمانی رخ می‌دهد که پاسخ ارائه شده به سؤال بعدی به پاسخ‌های قبلی آزمون‌شوندگان مرتبط باشد. در این شرایط چنانچه دانشجو به سؤالات ابتدایی پاسخ نادرست دهد، پاسخ به سؤالات بعدی او نیز تحت تاثیر قرار خواهد گرفت. به منظور جلوگیری از بروز این خطا در سؤالات تشریحی تغییریافته می‌توان قبل از طرح سؤال بعدی، پاسخ صحیح سؤال قبلی را به فراگیر ارائه داد.

پاسخ‌های چنین سؤالی مطلقاً صحیح یا مطلقاً غلط نیستند. به عبارت دیگر پاسخ‌هایی که دانشجویان تولید می‌کنند در طیفی از رنگ‌های خاکستری قرار می‌گیرند. همچنین، نمی‌توان آنها را صرفاً به یک گروه و رشته خاص نسبت داد. بنابراین هنگام نوشتن پاسخ، دانشجو باید مجموعه‌ای از اطلاعات خود را که از طرق گوناگون به دست آورده است، با هم تلفیق کند. در برخی از موارد، دانشجویان خیلی راحت با این نوع سؤال کنار نمی‌آیند چون احساس می‌کنند اطلاعات کافی برای

1. Rabinowits
2. Palmer & Devitt
3. Cumulative error

حل مسأله در اختیار آنها قرار داده نشده است. باید توجه داشت که در اکثر مسائل واقعی پزشکی، تصمیم‌گیری در شرایط عدم قطعیت رخ می‌دهد و پزشک مجبور است بدون در اختیار داشتن اطلاعات کافی و لازم، تصمیم‌گیری کند. از این رو، سؤال تشریحی تغییریافته می‌تواند موقعیت مناسبی برای ارزیابی نحوه برآورد «احتمالات» و به کار بردن اطلاعات قبلی در شرایطی مشابه شرایط واقعی فراهم نماید.

به طور کلی، هدف از این آزمون‌ها قرار دادن دانشجو در موقعیتی شبیه موقعیت واقعی و مواجه کردن او با مسائلی است که زمینه‌ساز تصمیم‌گیری‌های آتی اوست. محتوایی که توسط سؤال تشریحی تغییریافته می‌تواند مورد ارزیابی قرار گیرد، شامل یادآوری محفوظات، کاربرد اطلاعات، نحوه استدلال و همچنین جنبه‌های نگرشی و اخلاقی مسائل است. بنابراین استفاده از این نوع سؤال، تنها به سنجش حیطه شناختی محدود نمی‌شود، بلکه سایر موارد از قبیل موضوعات اخلاقی، نگرشی و ... را به این ترتیب می‌توان سنجید. از آنجا که با طرح یک سناریو، سؤالات از جنبه‌های مختلف مطرح می‌شوند، طراحی سؤال معمولاً توسط مجموعه‌ای از متخصصان حوزه‌های مختلف در یک پانل انجام می‌شود. همچنین پاسخ صحیح سؤال نیز به وسیله جمعی از استادان تعیین می‌شود. بنابراین یکی از موارد شایع استفاده از این روش، ارزیابی دانشجو در دوره‌های آموزشی ادغام‌یافته[1] یا دروس بین رشته‌ای[2] است. باید توجه داشت که طراحی سؤالات تشریحی تغییریافته بسیار وقت‌گیر است و طراحی آن نیاز به تخصص و تجربه کافی دارد.

گام‌های طراحی سؤال تشریحی تغییریافته

خلاصه مراحل طراحی سؤال تشریحی تغییریافته در جدول ۱-۱۱ نشان داده شده است.

جدول ۱-۱۱: خلاصه مراحل طراحی یک سؤال تشریحی تغییریافته

ردیف	عنوان مرحله	توضیح
۱	انتخاب یک موضوع و هدف آموزشی مناسب	موضوعات قسمتی از اطلاعات و دانش داوطلب هستند که قرار است سؤال آن را بسنجد. این موضوعات در ارتباط با توان‌مندی‌هایی است که قرار است در داوطلبان مورد ارزیابی قرار گیرد.
۲	انتخاب مورد بالینی مناسب	اساسی‌ترین مرحله در طراحی سؤال تشریحی تغییریافته، یافتن سناریوی مناسب به شکلی است که حوزه کاری فراگیر در آینده را پوشش دهد.
۳	نگارش پایه سؤال	پایه سؤالات تشریحی تغییریافته باید به شکلی نگارش یابد که اطلاعات به یک باره در اختیار دانشجو قرار ندهد. بلکه باید به شکل تدریجی و مرحله به مرحله اطلاعات در اختیار فراگیر قرار گیرد.
۴	تصمیم‌گیری در مورد پاسخ‌های صحیح و نحوه نمره‌دهی	در این مرحله باید در مورد پاسخ‌های احتمالی و اینکه چه مواردی به عنوان پاسخ صحیح پذیرفته می‌شوند یا نمی‌شوند باید تصمیمات لازم اتخاذ شود. در خصوص نحوه نمره‌دهی به قسمت‌ها و آیتم‌های مختلف یک سؤال، نیز باید توافق نظر حاصل شود.
۵	مرور و ارزیابی سؤالات	از چک لیست ارزیابی سؤالات تشریحی تغییریافته استفاده کنید. از همکاران خود بخواهید یک بار به سؤال پاسخ دهند .معمولاً دیگران متوجه اشتباه‌هایی می‌شوند که طراح سؤال ممکن است در هنگام طراحی سؤال به آنها توجه نکند یا فراموششان کند.

1. Integrated course
2. Interdisciplinary

انتخاب موضوع (یا موضوعات) مناسب

اساسی‌ترین مرحله در طراحی سؤال تشریحی تغییریافته که موفقیت سؤال را تا حد زیادی تضمین می‌کند، یافتن سناریوی مناسب به شکلی است که حوزه کاری فراگیر را پوشش دهد. برای این کار ابتدا باید بر اساس بلوپرینت، موضوع‌هایی که قرار است مورد ارزیابی قرار گیرند، مشخص شوند. در واقع می‌توان گفت که یکی از اساسی‌ترین مراحل در تدوین سؤالات آزمون، تصمیم در خصوص ماهیت و سطح توانمندی مورد انتظار جهت ارزیابی است. به عنوان مثال، «تشخیص، درمان و مسائل اخلاقی در بیمار مبتلا به سرطان» می‌تواند موضوع مناسبی باشد. هدف سؤال می‌تواند بیش از یک موضوع باشد و حتی به چند رشته مربوط شود مانند «آناتومی و رادیولوژی در شکستگی لگن».

انتخاب مورد بالینی مناسب

مرحله بعدی نوشتن سناریویی است که جنبه‌های موردنظر را پوشش دهد. همان‌طور که گفته شد، انتخاب مورد بالینی به نحوی که قابلیت پرداختن از جنبه‌های مختلف داشته باشد و در عین حال تقریباً مطابق شرایط واقعی باشد و از طرفی با سطح دانشجویان متناسب باشد، از اهمیت بالایی برخوردار است.

نگارش پایه سؤال

هرچند سؤال تشریحی تغییریافته به صورت مجموعه‌ای از چند سؤال کوته‌پاسخ مرتبط با یک پایه نیز طراحی می‌شود اما در شکل معمول آن، همه اطلاعات به یک باره در اختیار دانشجو قرار داده نمی‌شود. به صورت کلی، تعداد سؤالات یک آزمون که از سؤالات تشریحی تغییریافته تشکیل شده، ۱۰ سؤال و زمان مورد نیاز حدود یک ساعت در نظر گرفته می‌شود که البته به سطح دشواری سؤالات نیز بستگی دارد. در شکل ۱-۱۱ توالی طراحی سؤال تشریحی تغییریافته نمایش داده شده است. آغاز پایه سؤال معمولاً به صورت معرفی فرد مراجعه کننده است:

خانم ۲۷ ساله‌ای که دو فرزند دارد، به پزشک عمومی مراجعه کرده تا برای کودک کوچکترش که سرماخوردگی جزیی دارد، دارو بگیرد. هنگامی که از در خارج می‌شود، به شما می‌گوید: «اگر سرتان خیلی شلوغ نیست، خواستم بگویم که در این دو ماه خیلی احساس خستگی می‌کنم».

۱. از نظر شما به عنوان پزشک عمومی، او سعی داشته به چه چیزی اشاره کند (۰/۵ نمره)؟
۲. شما چه جوابی به او می‌دهید (۰/۵ نمره)؟

دانشجو با خواندن این جملات با کلیات موقعیت آشنا می‌شود. در ادامه، بقیه سناریو در اختیار او قرار می‌گیرد و وضعیت بیشتر روشن می‌شود. در اینجا سؤلات بیشتری مطرح می‌شوند:

این خانم می‌گوید که سال‌هاست آبریزش بینی دارد و طی دو ماه گذشته سردردی داشته که با التهاب بینی بدتر می‌شده است.
۳. در این مرحله چه تشخیص افتراقی‌هایی برای او مطرح است؟ سه مورد را نام ببرید (۰/۷۵ نمره).
۴. برای هر یک از تشخیص‌ها، سه علامت یا نشانه نام ببرید (۰/۷۵ نمره).

شکل ۱-۱۱: الگوریتم نمایشی از توالی کلی یک سؤال تشریحی تغییریافته که نشان‌دهنده نحوه ارائه تدریجی اطلاعات در سناریو و مورد بالینی است.

همین‌طور که مورد بالینی بیشتر معرفی می‌شود و اطلاعات بیشتری در اختیار دانشجو قرار می‌گیرد، مشخص می‌شود که بیمار مبتلا به تومور مغزی است. با این حال، دانشجویی که در پاسخ به سؤال سوم، رینیت، سینوزیت و سردرد تنشی را به عنوان تشخیص افتراقی ذکر کرده است، نمره بالاتری می‌گیرد نسبت به دانشجویی که تومور مغزی را در لیست تشخیص‌های افتراقی خود آورده است زیرا با اطلاعاتی که تا همین جا به دانشجو داده شده بود، رسیدن به تشخیص تومور، نشان دهنده استدلال ضعیف است.

به علاوه، همان‌طور که پیشتر گفته شد، جنبه‌های نگرشی و اخلاقی و رفتاری را نیز می‌توان با این سؤال سنجید. به عنوان مثال ممکن است طراح بخواهد سؤال را این طور پیش ببرد:

دو هفته بعد بیمار با همسرش مراجعه می‌کند و می‌گوید که سردردش بدتر شده است. بیمار معمولاً همراه همسرش مراجعه نمی‌کرده است.
۵. فکر می‌کنید دو دلیلی که احتمالاً برای این کار وی وجود داشته و عملکرد شما را تحت تاثیر قرار می‌دهد، چه بوده است (۱ نمره)؟

همان‌طور که ملاحظه می‌کنید، این سؤال مهارت‌های شناختی و ذهنی دانشجو را در حیطه رفتارهای انسان بررسی می‌کند. از جمله اینکه برای هر موردی ممکن است بیش از یک توضیح و دلیل وجود داشته باشد و اینکه در یک ارتباط سه نفره، نظرات و عقاید دیگران نیز باید لحاظ شود.

در ادامه، طراح از دانشجو می‌خواهد تا یافته‌های شرح حال و معاینه را تحلیل کند:

در معاینه چشم بیمار متوجه می‌شوید که احتمالاً مشکل بیمار علت ارگانیک دارد.
۶. سه یافته‌ای که احتمالاً در معاینه چشم بیمار پیدا کرده‌اید و با تشخیص‌های افتراقی شما مطابقت دارند، چه هستند (۰/۷۵ نمره)؟

سپس، بر اساس شرح حال و معاینه بیمار اطلاعات بیشتری در اختیار دانشجو قرار داد می‌شود که مشخص می‌کند بیمار مشکل بینایی به شکل دوبینی دارد. بسته به سطح دانشجو و سطحی که برای دشواری سؤال در نظر گرفته شده است، می‌توان سناریو را برای دریافت مشاوره پزشکی و تصمیم‌گیری برای بستری کردن بیمار ادامه داد و در نهایت، با طرح مسائلی که متعاقب مرگ غیرمنتظره بیمار برای همراهان و پزشکان ایجاد می‌شود، آن را پیچیده‌تر کرد. نکته‌ای که باید در نظر داشت، این است که حتماً برای هر سؤال فضای مناسب و کافی جهت نوشتن پاسخ در نظر گرفته شود.

در زیر نمونه دیگری از سؤالات تشریحی تغییریافته ارائه شده است.

نمونه‌ای از سؤال تشریحی تغییر یافته

شیرخوار ۱۳ ماهه توسط مادرش به درمانگاه آورده شده است. نگرانی اصلی مادر، سرفه کودک است. او می‌گوید که بچه تا ۶ ماهگی کاملاً سالم بوده ولی از آن موقع به طور مداوم مریض بوده است. این طور که مادر ذکر می‌کند کودک تاکنون ۵ بار به اوتیت مدیا، دو بار سینوزیت و یک بار برونشیت مبتلا شده است. همچنین با تشخیص پنومونی پنوموکوکی در یازده ماهگی در بیمارستان بستری شده است. مادر به شدت نگران عفونت‌های مکرر و مصرف زیاد آنتی بیوتیک است و می‌گوید دختر ۴ ساله‌اش هرگز این طور نبوده است.
۱. چهار مکانیسم دفاعی طبیعی سیستم تنفس را نام ببرید و به طور خلاصه توضیح دهید که هر یک چگونه از عفونت جلوگیری می‌کنند (۲ نمره).
۲. مشکلات کودک بعد از شش ماهگی شروع شده است. علت آن را در چه می‌دانید (۰/۵ نمره)؟
شما فکر می‌کنید که مشکل کودک، نقص سیستم ایمنی است.
۳. سه کاری که می‌توانید در این لحظه برای بررسی بیشتر انجام دهید، نام ببرید و توضیح دهید علت اینکه آنها را انتخاب کردید، چه بوده است (۱/۵ نمره)؟
شما در حال انجام بررسی‌های تشخیصی هستید که از وجود نقص سیستم ایمنی در کودک مطمئن شوید. در معاینه درمی‌یابید که یافته‌ها منطبق با پنومونی برونشیال هستند.
۴. دو مکانسیمی که منجر به مقاومت باکتری نسبت به آنتی بیوتیک می‌شود، توضیح دهید (۱ نمره).
یک هفته بعد و با مصرف کامل داروهای تجویز شده، حال کودک خوب می‌شود. ولی نتایج آزمایش‌ها نشان دهنده سطح بسیار پایین تمام انواع ایمونوگلوبولین هاست.
۵. دو نقصی که می‌توانند موجب این امر شوند، توضیح دهید (۱ نمره).
درمان کودک را با انفوزیون وریدی ایمونوگلوبولین (IVIg) آغاز می‌کنید.
۶. چه نوع ایمنی از طریق این درمان در کودک ایجاد می‌شود (۰/۵ نمره)؟
۷. یک نکته مثبت و یک نکته منفی در مورد این نحوه درمان نام ببرید (۱ نمره).

تصمیم گیری در مورد پاسخ‌های صحیح و تعیین نحوه نمره‌دهی

در این مرحله باید پاسخ‌های احتمالی را پیش‌بینی کرد و تصمیم گرفت چه مواردی به عنوان پاسخ صحیح پذیرفته می‌شوند و چه مواردی به عنوان پاسخ صحیح پذیرفته نخواهند شد. در خصوص نحوه نمره‌دهی به قسمت‌ها و آیتم‌های مختلف یک سؤال، تصمیم‌گیری و توافق‌نظر باید بین طراحان سؤال حاصل گردد. همچنین در این مرحله باید مشخص

گردد که آیا با ذکر موارد نادرست از دانشجو نمره کم خواهد شد یا خیر.

در نهایت باید یک برگه راهنمای تصحیح مانند مثال زیر تهیه شود و در آن مواردی که توافق شده به عنوان پاسخ صحیح پذیرفته شود یا نشود، مشخص گردد. گاهی بیش از یک کلمه و عبارت مدنظر است. گاهی لازم است برای مصحح توضیح داده شود و تشریح گردد که چه جوابی از دانشجو مورد قبول است. توجه کنید که تدوین راهنمای تصحیح در کنار طراحی یک پایه و سناریوی مناسب، مواردی هستند که یک سؤال تشریحی تغییریافته مطلوب را تشکیل می‌دهند.

راهنمای تصحیح سؤالات تشریحی تغییریافته

ذکر این موارد صحیح است (هرکدام چند نمره).
ذکر این موارد مورد قبول است.
ذکر این موارد نادرست است.

در هنگام تدوین الگوی پاسخ باید توجه شود که عینی کردن و جزیی کردن بیش از حد پاسخ، لزوماً به معنای افزایش پایایی آزمون نیست و نه تنها به بهبود کیفیت سؤال کمک نمی‌کند، می‌تواند مشکل‌ساز نیز باشد.

راهنمایی‌های لازم باید به دانشجو نیز ارائه شود. ابتدای دفترچه آزمون، باید برای دانشجویان توضیح داده شود که قسمت‌های مختلف سؤال را به ترتیب پاسخ دهند. دانشجویان باید توجه کنند که خواندن زودتر قسمت‌های بعدی سؤال به آنها در یافتن پاسخ کمکی نخواهد کرد. همچنین نمره‌بندی قسمت‌های مختلف سؤال و زمان کل آزمون را باید برای ایشان مشخص کرد.

برای تصحیح برگه‌ها، نام و مشخصات دانشجو نباید آشکار باشد. هر دانشجو با یک شماره مشخص می‌شود. نکته مهم دیگری که در زمان تصحیح باید به آن توجه شود این است که احتمال دارد مصحح با پاسخ‌هایی مواجه شود که قبلاً پیش‌بینی نشده است. در این گونه موارد او می‌تواند این موارد را یادداشت کند تا مجدداً در خصوص نمره‌دهی آنان توسط پنل طراحان سؤال تصمیم‌گیری شود.

برای نمره‌دهی سؤالات تشریحی تغییریافته معمولاً از یک یا دو روش استفاده می‌شود که این موضوع بستگی به این دارد که هدف از آزمون ارزیابی عملکرد فراگیران در مقایسه با یکدیگر (آزمون هنجاری) یا مقایسه عملکرد آن‌ها با معیار از پیش تعیین شده (آزمون معیاری) است. در صورتی که آزمون از نوع هنجاری باشد، استفاده از یک مقیاس نمره‌دهی گسترده از صفر تا ۱۰ برای هر سؤال پیشنهاد می‌شود. در حالی‌که در آزمون‌های معیاری، مقیاس نمره‌دهی محدود صفر و یک توصیه می‌شود.

مرور و ارزیابی سؤالات طراحی شده

برای بهبود کیفیت سؤالات می‌توان توسط چک‌لیستی که در جدول ۸ آمده است، به نقد و ارزیابی سؤال پرداخت.

همان‌طور که پیشتر نیز اشاره شد برای بهبود کیفیت سؤالات طراحی شده می‌توان آن‌ها را در اختیار تعدادی از همکاران که در فرایند طراحی سؤالات مشارکت نداشتند قرار داد و از آنها درخواست نمود که به سؤالات پاسخ دهند. به این طریق می‌توان مواردی را که برای ایشان مبهم بوده است، در متن سؤال اصلاح و بازبینی کرد.

سودمندی سؤالات تشریحی تغییریافته

روایی سؤالات تشریحی تغییریافته

اکثر محققان در مطالعات مختلف معتقد هستند که این سؤالات از روایی محتوایی بالای برخوردارند زیرا قادر هستند اطلاعات حاصل از شرح حال‌گیری و معاینه فیزیکی را به منظور ارائه یک فرضیه مناسب در ارتباط با مشکل بالینی ترکیب کنند (فلتی ۱۹۸۰، نیوبل و همکاران[1] ۱۹۸۱). روایی سازه این نوع سؤالات نیز به طرق گوناگون در مطالعات مختلف مورد بررسی قرار گرفته است. نتایج حاصل از مطالعه نیوبل و همکاران (۱۹۷۹ و ۱۹۸۱) نشان داد که نمرات حاصل از اجرای آزمون‌های تشریحی تغییریافته با افزایش مقطع تحصیلی دانشجویان افزایش می‌یابد (به عنوان مثال دستیاران در آزمون عملکرد بهتری از کارورزان خواهند داشت و کارورزان عملکرد بهتری نسبت به کارآموزان) که نشانه روایی سازه است. نتایج حاصل از تحلیل آزمون‌های تشریحی تغییریافته در دانشگاه نیوکاسل نیز نشان داد که پاسخگویی به سؤالاتی که نیازمند مهارت حل مسأله و استدلال بالینی بودند، در دانشجویان سال دوم پزشکی بیشتر از دانشجویان سال اول است. سایر مطالعات به ارزیابی روایی پیش‌بینی این نوع سؤالات با آزمون‌های بالینی و سایر انواع آزمون‌های کتبی نیز پرداخته‌اند (نیوبل و همکاران ۱۹۷۹ و ۱۹۸۱). جهت کسب اطلاعات بیشتر در این مورد به مبحث مسائل چالشی سؤالات تشریحی تغییریافته مراجعه کنید.

والرستد و همکاران[1] ۲۰۱۱

این مطالعه در دوره طب داخلی برنامه پزشکی دانشگاه گوتنبرگ انجام شد. در این پژوهش، دانشجویان در انتهای دوره آموزشی در یک امتحان برای ارزیابی دانش نظری شرکت می کردند. ۴۹ دانشجو در آزمونی که شامل ۱۵ سؤال کوته‌پاسخ و ۵ سؤال تشریحی تغییریافته بود شرکت کردند. تمامی سؤالات قبلاً توسط سایر دانشگاه‌های سوئد طراحی شده بود و روایی آن‌ها نیز محاسبه شده بود. همبستگی بین سؤالات تشریحی تغییریافته و سؤالات کوته‌پاسخ ۰/۵۹ بود ($P<0/001$). درصد پاسخ‌های درست فراگیران در هر دو نوع سؤالات با یکدیگر تفاوت معناداری نداشت. محققان بر اساس نتایج حاصل از مطالعه پیشنهاد کردند که بین سؤالات تشریحی تغییریافته و سؤالات کوته‌پاسخ از حیث ارزیابی دانش شناختی تفاوتی وجود ندارد.

1. Wallerstedt et al.

پایایی سؤالات تشریحی تغییریافته

مطالعات متعددی به ارزیابی پایایی آزمون‌های تشریحی تغییریافته پرداخته‌اند (فلتی ۱۹۸۰، نیوبل و همکاران ۱۹۸۱، فلتی و گیلیز[2] ۱۹۸۲). اولین مطالعه در این زمینه در کالج سلطنتی پزشکان عمومی نشان داد که این نوع از سؤالات در مقایسه با سؤالات تشریحی مزایایی دارند، از جمله آنکه پایایی بین ارزیابان در این سؤالات بیشتر از سؤالات تشریحی است. بعدها این موضوع با نتایج حاصل از مطالعه نیوبل و همکاران (۱۹۸۱) که ضریب پایایی ۰/۹۵ گزارش کردند نیز مورد تایید قرار گرفت.

فلتی ۱۹۸۰

فلتی در مطالعه خود از سؤالات تشریحی تغییریافته به منظور ارزیابی مهارت‌های حل مسأله و استدلال بالینی دانشجویان سال اول و دوم دوره پزشکی عمومی دانشگاه نیوکاسل استفاده نمود. در این مطالعه ضریب پایایی آزمون برابر با ۰/۹۱ تخمین زده شد. روایی سؤالات نیز بر اساس مدلی از مهارت حل مسأله پزشکی و تاکسونومی‌های مهارت شناختی تعیین شد.

1. Newble et al.
2. Feletti & Gillies

استراتفورد و پیرس‌فن [1] ۱۹۸۰

این محققان در مطالعه خود به بررسی روایی و پایایی آزمون‌های تشریحی تغییریافته پرداختند. روایی صوری مطالعه از طریق بررسی سؤالات توسط گروهی از متخصصین بالینی ارزیابی شد. همچنین برای بررسی روایی ملاکی آزمون نتایج حاصل از آزمون در دو گروه متخصصین بالینی و دانشجویان مقایسه شد. در این مطالعه، یک آزمون ۱۸ سؤاله طراحی شد و سپس بر روی گروهی از متخصصان بالینی پایلوت گردید. بعد از انجام پایلوت، شش سؤال انتخابی در قالب آزمونی برای ۲۵ نفر از دانشجویان فیزیوتراپی اجرا گردید. این آزمون بلافاصله پیش از دومین راند کارآموزی ارتوپدی اجرا شد. شاخص پایایی آزمون (ضریب آلفا) برابر با ۰/۳۹ محاسبه شد. به علاوه روایی آزمون نیز از طریق بررسی همبستگی بین نمرات آزمون تشریحی تغییریافته با نمرات آزمون چند گزینه‌ای حاوی ۵۰ سؤال و نمرات استدلال بالینی فراگیران مورد سنجش قرار گرفت.

محققان بر اساس نتایج حاصل از مطالعه مطرح کردند که ضریب پایایی به دست آمده در این مطالعه با سایر منابع موجود تفاوت چشمگیری دارد. همچنین نتایج این مطالعه نشان‌دهنده همبستگی بالای نمرات این آزمون با نمرات استدلال بالینی فراگیران در بخش‌های بالینی بود. بر اساس نتایج حاصل محققان توصیه کردند که سؤال تشریحی تغییریافته می‌تواند یک ابزار ارزیابی کتبی مفید در ارزیابی استدلال بالینی باشد.

1. Pierce-Fenn & Stratford

تاثیر آموزشی سؤالات تشریحی تغییریافته

تاثیر استفاده از سؤالات تشریحی تغییریافته بر میزان یادگیری فراگیران در مطالعات تجربی مختلف مورد بررسی قرار گرفته است. به عنوان مثال، نشان داده شده است که نمرات فراگیران با اجرای مستمر آزمون‌های تشریحی تغییریافته بهبود می‌یابد به این طریق که دانشجویان سال‌های بالاتر توانستند عملکرد بهتری در آزمون‌ها در مقایسه با دانشجویان سال‌های پایین‌تر کسب کنند (نیوبل و همکاران ۱۹۸۱).

هزینه و قابلیت اجرای سؤالات تشریحی تغییریافته

به دلیل آشنایی کمتر استادان و ماهیت چند رشته‌ای بودن سؤال، طراحی سؤالات تشریحی تغییریافته نیازمند صرف زمان و نیروی انسانی بسیار بیشتری است. این موضوع قابلیت اجرای این سؤالات را در مواجهه با موقعیت‌هایی که بخش وسیعی از محتوای آموزشی باید مورد ارزیابی قرار گیرد، قدری دشوار می‌کند. با این حال، از طریق تهیه بانک سؤالات، در سال‌های متمادی از سؤالات مناسب طراحی شده با حفظ اصول امنیتی می‌توان استفاده نمود.

مقبولیت سؤالات تشریحی تغییریافته

طراحی سؤالات تشریحی تغییریافته در قالب سناریو و مورد بالینی باعث گردیده است که فراگیران از این سؤالات استقبال خوبی داشته باشند. هر چند به دلیل آشنایی کمتر با این سؤالات توصیه می‌شود که از آن‌ها در قالب ارزیابی تکوینی در طی فرایند آموزش استفاده شود. به این ترتیب می‌توان انتظار داشت که در امتحان پایان دوره اضطراب و استرس ناشی از مواجه با شکل جدیدی از آزمون تا حدودی کاهش یابد. از نقطه نظر استادان نیز در صورت ارائه آموزش‌های مناسب در جهت طراحی صحیح این شکل از سؤالات، تمایل به استفاده از این سؤالات افزایش خواهد یافت.

سؤالات رایج در مورد سؤالات تشریحی تغییریافته

آیا سؤالات تشریحی تغییریافته به ارزیابی سطوح بالای حیطه شناختی می‌پردازند؟

یکی از انتقادهای جدی که به سؤالات تشریحی تغییریافته وارد می‌گردد، این است که این سؤالات علی‌رغم ادعای مطرح شده، توانایی چندانی در ارزیابی سطوح بالای حیطه شناختی ندارند. از این رو، محققان در مطالعات گوناگون سعی کرده‌اند تا به بررسی این موضوع بپردازند.

والارستد و همکاران در مطالعه خود دریافتند که بین نتایج حاصل از آزمون‌های کوته‌پاسخ و سؤالات تشریحی

تغییریافته همبستگی معناداری وجود دارد. از دیدگاه نویسندگان این مقاله این موضوع چندان غیرعادی نیست زیرا می‌تواند تاییدکننده گزارش‌های موجود باشد که بر اساس آنها تعداد بسیار زیادی از سؤالات تشریحی تغییریافته به گونه‌ای طراحی می‌شوند که تنها به ارزیابی سطوح پایین حیطه شناختی می‌پردازند. در واقع آنها را می‌توان تنها نوع تکامل یافته‌تری از سؤالات کوته‌پاسخ در نظر گرفت و به این ترتیب بدیهی است که همبستگی بین آنها خوب به دست خواهد آمد.

ایروین و بامبر[1] نیز در مطالعه خود به بررسی سطح شناختی سؤالات تشریحی تغییریافته به کار رفته در آزمون نهایی پزشکی در دانشگاه کوئینز[2] پرداختند. در این مطالعه محققان با استفاده از تاکسونومی بلوم و سطوح شناختی بوک‌والتر[3] تک‌تک سؤالات را تحلیل و مقایسه کردند. محققان از نمرات حاصل از هر سؤال به عنوان یک عامل برای پیش‌بینی هرگونه ارتباط زیربنایی بین سؤالات استفاده کردند. نتایج حاصل از مطالعه، نشان‌دهنده توزیع نابرابر[4] حیطه شناختی سؤالات بین سال‌های ۱۹۷۸ تا ۱۹۸۰ بود که تاحدی منعکس‌کننده تنوع موجود در مسائل ارائه شده و مشکل مربوط به تصحیح تعداد زیاد برگه‌های امتحانی به صورت دستی بود. مقایسه سؤالات تشریحی تغییریافته سال ۱۹۸۰ در ارزیابی دانش شناختی دانشجویان با سؤالات سال ۱۹۷۸ نیز نشان داد که این سؤالات بیشتر سطوح تحلیل، ترکیب و ارزشیابی را مورد سنجش قرار داده بودند. طبق نتایج حاصل از مطالعه ایروین و بامبر مشخص گردید که سؤالات تشریحی تغییریافته به صورت بالقوه توانایی سنجش سطوح مختلف تاکسونومی بلوم و بوک‌والتر را دارند.

پالمر و دویت [1] ۲۰۰۷

پژوهشگران در این مطالعه به بررسی سؤالات چند گزینه‌ای و سؤالات تشریحی تغییریافته به کاررفته در آزمون پایان دوره بالینی پزشکی عمومی پرداختند. دانشجویان سال چهارم دوره بالینی دانشگاه آدلاید به صورت معمول، یک آزمون کتبی را به عنوان بخشی از ارزیابی کلی عملکرد در پایان هفته نهم بخش جراحی دریافت می کردند. در این مطالعه، یک آزمون مشابه شامل ۵۰ سؤال چندگزینه‌ای و سه سؤال تشریحی تغییریافته در ابتدای بخش طراحی و به فراگیران ارائه شد. سؤالات چندگزینه‌ای شامل یک پاسخ درست و چهار گزینه انحرافی بود که بر اساس راهنمای عملی طراحی سؤالات چندگزینه‌ای طراحی شده بودند. به‌علاوه نتایج حاصل از دو آزمون تشریحی تغییریافته (هر کدام شامل ۱۵ سؤال) در دانشگاه آدلاید مورد تحلیل قرار گرفت. در مجموع ۳۳ سؤال تشریحی تغییریافته مورد تحلیل قرار گرفت. سؤالات تشریحی تغییریافته حداقل توسط ۱۲ طراح سؤال به صورت مجزا از یکدیگر بر اساس استانداردهای متدولوژی مورد بازبینی قرار گرفت. روایی سازه و سطح شناختی مورد ارزیابی توسط هر کدام از سؤالات چندگزینه‌ای و سؤالات تشریحی تغییریافته نیز توسط دو ارزیاب تحلیل شد. در نهایت ارزیابان در جلسه‌ای، در خصوص تحلیل شخصی‌شان از سؤالات بحث و تبادل‌نظر کردند و نمره نهایی برای هر سؤال چند گزینه‌ای و سؤال تشریحی تغییریافته در نظر گرفتند. ضریب توافق نظر بین ارزیابان بوسیله آزمون کاپا مورد بررسی قرار گرفت. پایایی بین دو ارزیاب و نمره نهایی سؤالات تشریحی تغییریافته برابر ۰/۷ و ۰/۸ و سؤالات چند گزینه‌ای برابر با ۰/۷ و ۰/۸ بود. نتایج حاصل از مطالعه نیز نشان داد که بیش از ۵۰ درصد سؤالات تشریحی تغییریافته به ارزیابی سطوح پایین حیطه شناختی می پردازند و این عدد در سؤالات چند گزینه‌ای نیز مشابه بود. نتیجه بسیار جالب توجه اینکه سؤالات چندگزینه‌ای در مقایسه با سؤالات تشریحی تغییریافته به دفعات بیشتری به ارزیابی سطوح بالاتر حیطه شناختی پرداختند. محققان در این مطالعه پیشنهاد دادند در صورتی که سؤالات تشریحی تغییریافته نتواند به شکل مناسب به ارزیابی سطوح بالای شناختی بپردازد، سؤالات چندگزینه‌ای که خوب طراحی شده باشند، می‌توانند جایگزین مناسبی برای سؤالات تشریحی تغییریافته باشند.

1. Palmer & Devitt

ظفرخان و الجراله [2] ۲۰۱۱

این مطالعه با هدف بررسی مقایسه توانایی سؤالات چندگزینه‌ای و تشریحی تغییریافته در ارزیابی سطوح مختلف حیطه شناختی انجام شد. در این مطالعه ۵۰ سؤال تشریحی تغییریافته و ۵۰ سؤال چندگزینه‌ای، به صورت تصادفی از بانک سؤالات امتحانات دانشجویان سال چهارم دوره طب داخلی دانشگاه کاظم در پاکستان استخراج شد. اثربخشی سؤالات به‌وسیله دو ارزیاب بررسی شد. از ارزیابان خواسته شده بود تا سطح شناختی سؤالات را بر اساس تاکسونومی بلوم مشخص نمایند. آنالیز نتایج حاصل از مطالعه نشان داد که تنها ۴۰ درصد سؤالات تشریحی تغییریافته و ۶۰ درصد سؤالات چندگزینه‌ای به ارزیابی اهداف شناختی سطح سوم (حل مسأله) می‌پرداختند و مابقی سؤالات صرفاً سطح یادآوری و درک را می‌سنجیدند. هیچ تفاوت معناداری بین سؤالات تشریحی تغییریافته و چندگزینه‌ای در ارتباط با نوع سؤالات (یادآوری، درک و حل مسأله) مشاهده نشد (x^2 =۵/۳, p=۰/۰۷). ضریب توافق بین ارزیابان در سؤالات چندگزینه‌ای بالا (kappa=0.609; SE=0.093; CI=95% 0.426 – 0.792) و در سؤالات تشریحی تغییریافته پایین‌تر *(kappa=0.195; SE =0.073; CI=95% 0.052 – 0.338)* بود.

2. Zafar Khan & Aljarallah

1. Irwin & Bamber
2. Queen's University
3. Bookwalter
4. Uneven distribution

آیا سؤال تشریحی تغییریافته پیش‌بینی‌کننده عملکرد بالینی دانشجویان است؟

یکی از سؤالاتی که ذهن طراحان سؤال و دست‌اندرکاران آموزش را به خود مشغول کرده است، این است که نتایج حاصل از سؤالات تشریحی تغییریافته تا چه حد می‌تواند پیش‌بینی کننده عملکرد آتی فراگیران در بالین باشد؟ ایده ابتدایی طراحی این سؤالات بر این تصور استوار بود که از این سؤالات می‌توان برای ارزیابی مهارت‌های حل مسأله و استدلال بالینی فراگیران استفاده نمود. هرچند که در بررسی‌های نظری صورت گرفته به این موضوع اشاره شده بود اما مطالعات تجربی به نتایج متناقضی در این زمینه رسیدند.

ایروین و بامبر در مطالعه خود یک ارتباط متوسط بین ۰/۳۶ الی ۰/۴۳ را بین نمرات حاصل از آزمون‌های تشریحی تغییریافته و آزمون‌های بالینی برگزار شده یافتند. نیوبل و همکاران در مطالعه خود به دلیل پایایی پایین امتحانات بالینی، امکان بررسی ارتباط بین نمرات حاصل از این آزمون‌ها را با آزمون‌های تشریحی تغییریافته نداشتند. در مطالعه استراتفورد و پیرس‌فن (۱۹۸۰) نیز ضریب همبستگی بین سؤالات تشریحی تغییریافته و سؤالات چندگزینه‌ای ۰/۳۹ گزارش شده است که مطرح‌کننده روایی ملاکی پایین این آزمون‌ها بود.

رابینوویتز ۱۹۸۹

گروه پزشکی خانواده در دانشکده پزشکی جفرسون از سال ۱۹۷۶ از سؤالات تشریحی تغییریافته در امتحان پایان دوره دانشجویان پزشکی سال سوم مقطع کارآموزی استفاده می‌کرد. به منظور مقایسه سؤالات تشریحی تغییریافته با سؤالات چندگزینه‌ای مورد استفاده در امتحان پایان دوره سال پنجم دانشجویان مقطع کارآموزی، ارتباط نمرات ۲۱۷۴ فارغ التحصیل (۱۹۸۵-۱۹۷۶) با نمرات آزمون بورد ملی ارزیابان پزشکی، ارزیابی عملکرد بالینی دانشجویان سال سوم کارآموزی و ارزیابی کلی عملکرد آنان در چهار محور اصلی فارغ‌التحصیلی مورد بررسی قرار گرفت.

نتایج حاصل از مطالعه نشان داد که نمرات حاصل از آزمون چندگزینه‌ای بخش داخلی مکرراً دارای بالاترین همبستگی با نمرات حاصل از آزمون بورد ملی ارزیابان پزشکی و ارزیابی کلی دانش پزشکی فارغ التحصیلان است. نمرات حاصل از آزمون تشریحی تغییریافته در دورهٔ آموزشی پزشک خانواده دارای کمترین ارتباط با نمرات حاصل از این آزمون‌ها بود. با این وجود نمرات دوره آموزشی پزشک خانواده مکرراً همبستگی بالایی را با ارزیابی عملکرد بالینی دانشجویان در سال سوم و عملکرد فارغ‌التحصیلان در محور جمع‌آوری اطلاعات، قضاوت بالینی و نگرش حرفه‌ای نشان داد. نویسندگان این مقاله پیشنهاد دادند که بر اساس نتایج حاصل از مطالعه، سؤالات تشریحی تغییریافته ممکن است اطلاعات مهم و متفاوتی را در خصوص ارزیابی عملکرد دانشجویان پزشکی فراهم سازد.

منابع

1. Feletti GI. Reliability and validity studies on modified essay questions.Medical Education 1980; 55:933–941.
2. Feletti GI, Engel CE. The modified essay question for testing problemsolving skills.Med J Australia 1980; 1:79–80.
3. Feletti GI, Smith EKM. Modified essay questions: Are they worth the effort? Med Educ 1986; 20(2):126–132.
4. Irwin WG, Bamber JH.The cognitive structure of the modified essay question. Medical Education 1982; 16(6):326-31.
5. Knox JDE. How to use modified essay questions. Medical Teacher 1980; 2:20–24
6. Lim EC, Seet RC, Oh VM, Chia BL, Aw M, Quak SH, et al. Computerbased testing of the modified essay question: the Singapore experience.Medical Teacher 2007; 29: 261-8.
7. Lockie C, McAleer S, Mulholland H, Neighbour R, Tombleson P. Modified essay question (MEQ) paper: Perestroika. Occasional paper (Royal College of General Practitioners) 1990; 46:18–22.
8. Newble Dl, Baxter A, Elmslie RG.A comparison of multiple-choice tests and free response tests in examinations of clinical competence. Medical Education 1979; 13:263-268.
9. Newble DI, Hoare J, Elmslie RG.The validity and reliability of a new examination of the clinical competence of medical students. Medical Education 1981; 15:46–52
10. Norman G. Reliability and construct validity of some cognitive measures of clinical reasoning. Teaching and Learning Medicine 1989: 1(4):194–199
11. Palmer EJ, Devitt PG. Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? BMC Medical Education 2007; 7(49):1-7.
12. Palmer EJ, Duggan P, Devitt PG, Russellr. The modified essay question: Its exit from the exit examination? MedicalTeacher 2010; 32: e300–e307
13. Rabinowitz HK. Expansion of the modified essay Questions into an Audiovisual format. Medical Education 1985; 60: 883-85.
14. Rabinowitz HK. The modified essay question: an evaluation of its use in a family medicine clerkship. . Medical Education 1987; 21(2):114-8.
15. Rabinowitz HK, Hojat M. A comparison of the modified essay question and multiple choice question formats: their relationship to clinical performance. J of Family Medicine 1989; 21(5):364-7.
16. Schuwirth LWT, van der Vleuten CP. ABC of learning and teaching in medicine: Written assessment. BMJ 2003; 326:643–5
17. Schuwirth LWT, van der Vleuten CP. Written Assessments. In: Dent J, Harden R, Eds. New York: Elsevier Churchill Livingstone 2005; 311–22

18. Stratford P, Pierce-Fenn H. Modified essay question. Physical Therapy1985 ;7)65):1075–1079.
19. The Board of Censors of the Royal College of General Practitioners.The modified essay question. Proceedings of the Royal College of General Practitioners 1971; 21:373-376
20. Vleuten CP, van Luijk SJ, Beckers HJ. A written test as an alternative to performance testing. Medical Education 1989; 23: 97–107
21. Wallerstedt S, Erickson G, Wallerstedt SM. Short Answer Questions or Modified Essay Questions—More than a Technical Issue. International Journal of Clinical Medicine 2012; 3:28-30
22. Wass V, Van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. Lancet 2001; 357(9260):945–949.
23. Zafar Khan M. Badr Muhammad Aljarallah.Evaluation of Modified Essay Questions (MEQ) and Multiple Choice Questions (MCQ) as a tool for assessing the Cognitive Skills of Undergraduate Medical Students. International Journal of health sciences 2011; 5(1): 39–43

فصل ۱۲

سؤالات کوته پاسخ

ساختار سؤالات کوته‌پاسخ

در سؤالات کوته‌پاسخ همان‌طور که از نامشان پیداست از دانشجو خواسته می‌شود که در حد یک کلمه یا عبارت کوتاه، پاسخ خود را ارائه دهد. پاسخ ارائه شده برای سؤال کوته‌پاسخ که نوعی سؤال بازپاسخ و نیمه‌ساختاریافته هستند، آنقدر مختصر است که این آزمون‌ها را به کلی از آزمون‌های تشریحی مجزا می‌سازد و به آزمون‌های بسته‌پاسخ نزدیک می‌کند.

این سؤالات از آن جهت که نیازمند ارائه پاسخ‌های دقیق و مشخص هستند، شبیه به سؤالات بسته‌پاسخ هستند اما وجه تمایز این سؤالات در مقایسه با سؤالات بسته‌پاسخ، در تولید پاسخ به وسیله خود فراگیران به جای انتخاب آن از بین گزینه‌های ارائه شده است. بنابراین، می‌توان گفت سؤالات کوته‌پاسخ تنها نوع سؤال بازپاسخ عینی هستند که فراگیر، خود، پاسخ را تهیه می‌کند و می‌توان آنها را حدفاصل مابین سؤالات تشریحی و سؤالات چندگزینه‌ای در نظر گرفت. به گونه‌ای که در سؤالات کوته‌پاسخ، عموماً توانایی بازخوانی[1] آزمون‌شوندگان ارزیابی می‌شود، در حالی‌که در آزمون‌های بسته‌پاسخ معمولاً توانایی بازشناسی[2] یا تشخیص سنجیده می‌شود. بنابراین، این سؤالات به منظور رفع محدودیت‌های مربوط به انواع سؤالات تشریحی و بسته‌پاسخ از قبیل سؤالات چندگزینه‌ای به وجود آمدند.

به طور خلاصه، آزمون کوته‌پاسخ از تعدادی سؤال مختصر تشکیل می‌شود که پاسخ‌های آنها ممکن است در قالب یک کلمه، عبارت یا جمله تنظیم گردد. در این سؤالات معمولاً فضای در نظر گرفته شده برای پاسخ و وزن نمره سؤال، تعیین‌کننده حجم پاسخ ارائه شده است. از آنجا که سعی می‌شود سنجش سطوح بالای حیطه شناختی (درک و کاربرد مطالب) در سؤالات کوته‌پاسخ مدنظر قرار گیرد، تنه این سؤالات ممکن است به شکل یک سناریوی بالینی مطرح شود و به‌دنبال آن چندین سؤال کوتاه مرتبط ارائه گردد. این سؤالات ممکن است به ارزیابی ویژگی‌های کلیدی بیماری‌ها، مکانیسم عمل، عوارض جانبی و ... بپردازند. مشخصات و ویژگی‌های کلی سؤالات کوته‌پاسخ در جدول ۱-۱۲ خلاصه شده است.

جدول ۱-۱۲: کاربردهای سؤالات کوته‌پاسخ

کاربرد	بله	تا حدودی	خیر
امکان ارزیابی دانش شناختی	❊		
امکان ارزیابی دانش عملکردی		❊	
امکان ارائه بازخورد		❊	
مناسب برای ارزیابی تعداد زیاد فراگیران	❊		

1. Recall
2. Recognition

انواع سؤالات کوته‌پاسخ

به طور کلی دو نوع طبقه‌بندی برای سؤالات کوته‌پاسخ ارائه شده است: دسته‌بندی بر اساس سطح دانش مورد ارزیابی و دسته‌بندی بر اساس شکل سؤالات طراحی شده. در ادامه به شرح هر یک از زیر طبقات دسته‌بندی‌های ارائه شده خواهیم پرداخت.

دسته‌بندی سؤالات کوتاه‌پاسخ بر اساس موضوع مورد ارزیابی

در واقع اولین گام در طراحی سؤالات کوته‌پاسخ تصمیم در خصوص این موضوع است که چه سطحی از دانش قرار است به وسیله این سؤالات مورد ارزیابی قرار گیرد. پاسخ به این سؤالات نیازمند قرار دادن آن‌ها در دو دسته کلی است:

- **سؤالات دانشی[1]:** تعدادی از سؤالات کوته‌پاسخ تنها به ارزیابی یادآوری حقایق و سطوح پایین حیطه شناختی می‌پردازند. این سؤالات معمولاً برای سنجش اصول پایه و حقایق علمی مناسب هستند. نکته حائز اهمیت در خصوص این سؤالات آن است که استفاده بیش از حد از این سؤالات در آزمون، فراگیران را به یادگیری طوطی‌وار و سطحی مطالب سوق خواهد داد که با هدف اصلی آموزش در تضاد است. بنابراین از این نوع سؤالات تنها در شرایطی باید استفاده شود که هدف از آزمون، سنجش اصول پایه یک مطلب علمی است. نمونه افعالی که برای طراحی این دسته سؤالات استفاده می‌شود شامل: «نام ببرید»، «بیان کنید»، «شناسایی کنید»، «فهرست کنید» و «تعریف کنید» است.

نمونه سؤال کوته‌پاسخ دانشی
سینوس مایل پریکاردی در خلف کدام حفره قلبی قرار دارد؟

نمونه سؤال کوته‌پاسخ دانشی
درکدام ناحیه عقده لنفاوی، تجمع لنفوسیت‌های T مشاهده می‌شود؟

- **سؤالات تفسیری[2]:** این سؤالات عمدتاً توانایی کاربرد مطالب فراگرفته شده را در موقعیت‌های جدید و در مواجهه با مشکلات و مسائل موجود مورد سنجش قرار می‌دهند. همان‌طور که در بالا نیز اشاره شد، طراحی سؤالات کوته‌پاسخی که به ارزیابی سطوح بالاتر حیطه شناختی (درک و کاربرد اطلاعات) بپردازند کار بسیار دشواری است؛ به همین دلیل اغلب سؤالات طرح شده به خصوص توسط استادان کم تجربه‌تر تنها به سنجش سطح یادآوری حقایق می‌پردازند. بنابراین طراحی این نوع سؤالات کوته‌پاسخ نیازمند کسب تجربه و مهارت کافی است. نمونه افعال مورد استفاده برای طراحی این نوع سؤالات شامل «مقایسه کنید»، «تشخیص دهید»، «بحث کنید»، «دلیل آن را ارائه دهید»، «محاسبه کنید» و «توصیف کنید» است. نکته حائز اهمیت این است که معمولاً در آزمون‌ها سعی می‌شود تا ترکیبی از دو نوع سؤال دانشی و تفسیری استفاده شود. در این حالت اغلب اوقات قسمت اول سؤال یک سؤال دانشی و قسمت دوم تفسیری خواهد بود. بدین ترتیب آزمون می‌تواند نمونه بهتری از سؤالات مربوط به سطوح مختلف حیطه شناختی را در بر گیرد.

1. Factual questions
2. Interpretive questions

نمونه سؤال کوته‌پاسخ تفسیری

بیمار خانم ۳۰ ساله‌ای ساکن یکی از روستاهای شمال است که به علت تنگی‌نفس فعالیتی به خصوص بعد از کشاورزی، توسط پزشکی در خانه بهداشت ویزیت می‌گردد. پزشک حاذق به دنبال گرفتن شرح حال و انجام معاینه فیزیکی و سمع سوفل در ناحیه قدام و سمت چپ قفسه صدری به نارسایی دریچه میترال و نارسایی قلب به دنبال آن مشکوک شده و ایشان را به یک متخصص قلب ارجاع می‌کند که تشخیص قطعی می‌گردد.

۱. در این فرد سوفل (صدای غیرطبیعی) در کدام دوره قلبی شنیده می‌شود؟
۲. پیش‌بینی می‌کنید برون‌ده قلبی و حجم ضربهای در این فرد چه تغییری کند؟
۳. چه عاملی در این فرد میتواند منجر به ایجاد ادم ریه شود؟

دسته‌بندی سؤالات کوتاه‌پاسخ بر اساس شکل سؤال

یکی از طبقه‌بندی‌های شایع سؤالات کوته‌پاسخ که در منابع مختلف ارائه شده است، دسته‌بندی بر اساس شکل و فرمت سؤال است. در این حالت سؤالات کوته‌پاسخ به چهار دسته کلی طبقه‌بندی می‌شوند:

- **سؤال پرسشی[1]:** در این شکل، تنه سؤال به صورت یک عبارت پرسشی نوشته می‌شود و فراگیران نیز مجموعه‌ای از جملات، عبارات و حتی پاراگراف را در پاسخ مستقیم به سؤال ارائه می‌دهند. از این نوع از سؤالات کوته‌پاسخ، به خاطر آنکه طراحی و پاسخ‌دهی به آن‌ها آسان‌تر از سایر انواع سؤالات کوته‌پاسخ است، بیشتر استفاده می‌شود.

نمونه سؤال کوته‌پاسخ پرسشی

چه ویژگی‌هایی در مایع مغزی نخاعی باید وجود داشته باشد تا بتوان مننژیت ویروسی را از نوع باکتریال افتراق داد؟

سؤال کامل‌کردنی[2] یا تکمیلی: در این سؤال، پرسش به صورت یک جمله ناقص نوشته می‌شود که در آن یک جای خالی وجود دارد و فراگیران باید کلمه، عبارت، عدد یا علامت درست را در جای خالی بنویسند. درک منظور و ارائه پاسخ درست در این نوع سؤالات برای فراگیران تا حدودی دشوارتر از نوع قبلی است. بنابراین در هنگام طراحی این سؤالات باید تمام تلاش خود را به کار برد که سؤالات طراحی شده فاقد هر گونه ابهام از سوی آزمون‌شوندگان باشد. سؤالات نوع تکمیلی ممکن است به صورت دیاگرام، عکس، نقشه مفهومی و جدول نیز طرح گردند.

نمونه سؤال کوته‌پاسخ کامل‌کردنی

با توجه به ارتباط موارد نامبرده، نام یک کرم دیگر را ذکر کنید.
انتروبیوس ورمیکولاریس، آسکاریس لومبریکوئیدس، .....

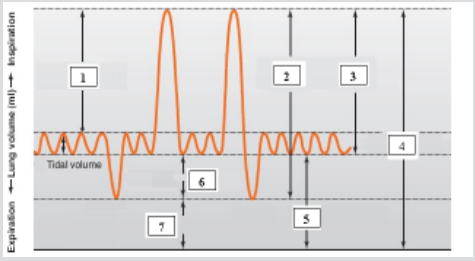
نمونه سؤال کوته‌پاسخ کامل‌کردنی

در هفته دوم جنینی، حفرات آمنیون و کیسه زرده به طور موقت توسط ..... به یکدیگر متصل می‌شوند.

1. Question
2. Completion

نمونه سؤال کوته‌پاسخ کامل‌کردنی

جوان سی ساله ای با سابقه تنگی نفس و خس خس سینه مراجعه نموده و پس از معاینات بالینی احتمال ابتلا به آسم آلرژیک برای وی مطرح می شود. برای وی اسپیرومتری انجام گردید. با توجه به منحنی اسپیرومتری که در زیر آمده است. در مقابل هر یک از شماره های ارائه شده، حجم ها و ظرفیت های ریوی را ذکر کنید.



شماره ۱

شماره۲

شماره۳

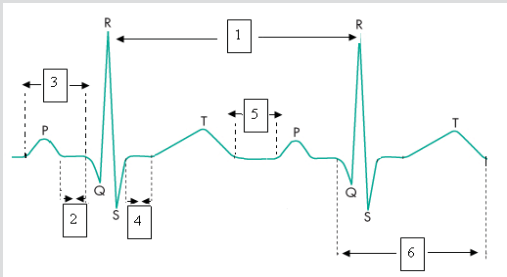
شماره۴

شماره۵

شماره۶

نمونه سؤال کوته‌پاسخ کامل‌کردنی

آقای ۵۰ ساله‌ای به علت درد قفسه سینه مراجعه نموده است. به گفته وی، درد با انجام فعالیت بیشتر می‌شود. در معاینات انجام شده برای وی تشخیص احتمالی ایسکمی قلبی مطرح شده که برای بررسی بیشتر درخواست الکتروکاردیوگرافی می‌شود. با توجه به الکتروکاردیوگرام ثبت شده، در مقابل هر یک از عبارات ارائه شده شماره درست را ذکر کنید.



مدت زمان تقریبی انقباض بطن

برای تعیین تعداد ضربان قلب

مرحله استراحت قلبی

فاصله بین شروع تحریک الکتریکی دهلیزها و بطن‌ها

مدت زمان تقریبی انقباض دهلیزها

- **سؤال هدایتی[1]:** در این نوع از سؤالات از فراگیران خواسته می‌شود تا کاری را انجام دهند از قبیل فهرست کردن یا نام بردن ویژگی‌ها، اجزاء و عملکردهای یک موضوع. جای خالی ممکن است در این سؤالات ارائه شود یا خیر. در این سؤالات فراگیران باید بخش‌های مختلف یک موضوع را به خاطر بیاورند تا بتوانند به سؤال پاسخ صحیح بدهند.

نمونه سؤال کوته‌پاسخ هدایتی

پروتئین‌های اصلی سازنده فیلامان‌های مسوول انقباضات عضلانی را نام ببرید.

- **سؤال تشخیصی[2]:** در آزمون‌های کوته‌پاسخ تشخیصی، سؤال به صورت یک موضوع یا مطلب ارائه می‌شود که فراگیران باید پاسخ‌های مربوط به فهرست کلمات، عبارات، اعداد یا علائم ارائه شده را در مقابل هر کدام بنویسند. این سؤالات از طریق دسته‌بندی سؤالات مرتبط باهم در قالب یک سؤال کلی باعث کاهش زمان خواندن سؤال برای فراگیران می‌شود. به‌علاوه به این طریق می‌توان حجم زیادی از محتوای دوره آموزشی را با استفاده از این سؤالات در محدوده زمان کوتاه ارزیابی نمود.

نمونه سؤال کوته‌پاسخ تشخیصی

در مقابل هر بیماری، نام ویروس یا باکتری ایجادکننده آن را بنویسید.

سل	دیفتری .....
کزاز	سفلیس .....
هرپس	تب مالت .....

استراتفورد ۱۹۸۸

هدف از این مطالعه، مقایسه کارآمدی دو نوع از سؤالات کوته‌پاسخ بود. کارآمدی در این مطالعه به عنوان پایایی در هر واحد از زمان آزمون در نظر گرفته شد. ۲۰ دانشجوی فیزیوتراپی سال دوم در دوره آموزشی ارتوپدی، در یک آزمون ۱۲ سؤالی از دو نوع سؤالات تشخیصی و کامل‌کردنی شرکت داده شدند. پیامدهای مورد اندازه‌گیری در این مطالعه، نمرات فراگیران و زمان مورد نیاز برای پاسخ به هر سؤال بود. نتایج حاصل از مطالعه نشان‌دهنده پایایی بالاتر (در واحد زمان) سؤالات نوع تشخیصی نسبت به نوع کامل‌کردنی است. همچنین سؤالات نوع تشخیصی سریع‌تر تصحیح شده و ضریب توافق بالاتری نسبت به سؤالات کامل کردنی داشتند. نمرات دانشجویان در سؤالات نوع تشخیصی بالاتر از سؤالات نوع کامل کردنی بود. هر چند این تفاوت از نظر آماری معنادار نبود.

ضرورت و کاربرد سؤالات کوته‌پاسخ

در آموزش علوم پزشکی از سؤالات تشریحی کوته‌پاسخ عمدتاً با هدف ارزیابی توانایی تفسیر تصاویر و نمودارها، حل مسأله و نیز کاربرد مطالب آموخته شده در موقعیت‌های جدید استفاده می‌شود. در حقیقت کاربرد اصلی این سؤالات زمانی است که لازم باشد به منظور ارزیابی سطوح بالای حیطه شناختی، فراگیران با موارد عینی‌تر مواجه شوند نه آنکه صرفاً سطح دانسته‌های‌شان مورد ارزیابی قرار گیرد. از مزایای دیگر این نوع سؤال این است که طراحی آن خیلی پیچیده نیست و همچنین، از آنجا که این سؤالات پاسخ مشخصی دارند، نحوه تصحیح برگه و نمره‌دهی آنها نیز مشکل زیادی ایجاد نمی‌کند. اما چون نمره‌دهی این سؤالات ممکن است به آسانی از نوع عینی خارج و به نوع ذهنی متمایل گردد، بهتر است در طراحی این نوع سؤالات از سؤال دانشی به جای نوع تفسیری استفاده شود. معمولاً این نوع سؤال وقتی به کار می‌رود که:

1. Directional
2. Identification

- هدف، ارزیابی توانایی یادآوری اطلاعات به جای بازشناسایی آن است.
- هدف، تقویت میزان یادگیری فراگیران در طول فرایند تدریس باشد.
- می‌خواهیم شانس ارائه پاسخ‌های تصادفی و حدسی توسط فراگیر کاهش یابد.
- طراح سؤال در تهیه گزینه‌های مناسب برای آزمون چندگزینه‌ای از تجربه و مهارت کافی برخوردار نباشد.

رسائیان ۱۳۸۴

در این مطالعه، بلوک تنفس و ارزیابی آن در دانشجویان اولین دوره اصلاحات دوره پزشکی عمومی در دانشگاه علوم پزشکی شهید بهشتی بررسی شد. در ترم دوم سال تحصیلی ۸۴-۸۳، آموزش بلوک تنفس به اولین دوره دانشجویان پزشکی که مشمول طرح اصلاحات آموزش پزشکی بوده‌اند انجام گردید. سؤالات امتحان، تحلیلی و از نوع چهارگزینه‌ای و کوته‌پاسخ بوده که در گروه تالیف درسنامه تنفس به تایید رسید. سؤالات شامل ۵۵ درصد چهارگزینه‌ای با یک پاسخ مورد نظر، ۳۷ درصد چهارگزینه‌ای با بیش از یک پاسخ درست و ۸ درصد کوته‌پاسخ بود. نتایج امتحان گویای قبولی کلیه دانشجویان با میانگین و انحراف معیار ۱/۸۹±۱۳/۵ بود. بررسی سؤالات امتحانی طبق تعاریف موجود نشان داد که ۵۷ درصد سؤالات از نظر سطح دشواری آسان بوده است و تنها دو سؤال (سه درصد) سخت بوده است. از نظر ضریب تمیز فقط ۴۸ درصد سؤالات مطلوب شناخته شدند.

کرین و همکاران [1] ۲۰۰۷

کرین در کتاب خود با عنوان «استفاده از سؤالات کوته‌پاسخ بالینی در مقطع تخصصی دندانپزشکی» به شرح مزایا و معایب استفاده از سؤالات کوته‌پاسخ در ارزیابی فراگیران پرداخته است. این کتاب در قالب دو بخش تنظیم شده است که بخش اول شامل ۷۳ سناریو و مورد بیمارمحور است که برای هر سناریو سه تا شش سؤال کوته‌پاسخ به منظور کمک به شرح و توصیف موقعیت بالینی ارائه شده است. حیطه و دامنه سناریوهای ارائه شده بسیار وسیع بوده و سعی شده است تا از رشته‌ها مختلف دندانپزشکی از قبیل پزشکی دهان، جراحی فک و دهان و رادیولوژی مثال‌هایی آورده شود. پاسخگویی به سؤالات نیازمند استفاده از علوم پایه و دانش پزشکی مرتبط با موضوع است. سؤالات طرح شده نه تنها به ارزیابی تشخیص افتراقی موارد بالینی می‌پردازد، بلکه همچنین در خصوص نحوه مدیریت، برنامه‌ریزی برای درمان و مراقبت‌های پیگیری نیز سؤالاتی طرح گردیده است. بخش دوم کتاب شامل ۵۲ مورد بالینی در خصوص دندانپزشکی ترمیمی است. محور اصلی سؤالات طرح شده در این بخش روی نحوه مواجهه و حل مشکلات مختلف دندانی است.

1. Crean et al

گام‌های طراحی سؤالات کوتاه‌پاسخ

در ادامه به شرح مراحل طراحی سؤالات کوته‌پاسخ با ذکر مثال‌های مرتبط خواهیم پرداخت. در جدول ۱۰ خلاصه این مراحل ذکر شده است.

انتخاب یک موضوع و هدف آموزشی مناسب

همان‌طور که گفته شد از سؤالات کوته‌پاسخ معمولاً برای به خاطر آوردن یک نام یا اطلاعات محدود استفاده می‌شود؛ به گونه‌ای که یک جواب مشخص در حد یک کلمه یا عبارت کوتاه وجود داشته باشد. این امر می‌تواند موجب شود که با طرح سؤالات سطحی، دانشجو به یادگیری محفوظات گرایش پیدا کند. بنابراین باید همواره تلاش نمود تا موضوعی برای طرح سؤال انتخاب شود که توان درک و به کارگیری اطلاعات را مورد سنجش قرار دهد نه به یادآوری صرف حقایق بپردازد. به عنوان مثال، در صورتی که هدف کلی آموزش، به صورت «... را بداند» باشد، احتیاج به اهداف عینی‌تری که قابل اندازه‌گیری باشند وجود خواهد داشت. به همین منظور در هنگام طراحی سؤالات کوته‌پاسخ معمولاً از افعالی مانند «بر شمارد»، «نام ببرد»، «فهرست کند»، «توضیح دهد»، «بیان کند» و «تعریف کند» استفاده می‌شود. این گونه اهداف معمولاً برای طراحی سؤالات کوته‌پاسخ مناسب هستند.

بنابراین به طور خلاصه می‌توان گفت برای طراحی نوع سؤال مورد استفاده، اولین گام توجه به پیامدهای مورد انتظار یادگیری و اهداف آموزشی است. به گونه‌ای که می‌توان با اطمینان گفت در صورتی که در هنگام طرح سؤال، اهداف

آموزشی در نظر گرفته نشوند، روایی محتوایی آزمون زیر سؤال خواهد رفت. مثال زیر نمونه‌ای از عدم توجه به اهداف آموزشی در هنگام طراحی سؤال است. همان‌طور که ملاحظه می‌کنید، هدف اول بسیار کلی است و سنجش فعل درک کردن دشوار و مبهم است. به ویژه قضاوت در خصوص میزان درک فراگیران از یک موضوع و اختصاص نمره به پاسخ‌های فراگیران، کار بسیار دشواری است. بنابراین توجه به این نکته ضروری است که هرچند استفاده از اهداف کلی برای توصیف یک دوره آموزشی یا ارائه راهنمایی برای فرایند یادگیری مناسب است، اما در تعیین ابزار ارزیابی، کمک چندانی نمی‌کند.

هدف ضعیف
دانشجو باید عوامل فیزیولوژیک موثر بر خشکی فضای بین آلوئولی را بداند.
هدف بهتر
دانشجو بتواند عوامل فیزیولوژیک موثر بر خشکی فضای بین آلوئولی را نام ببرد.

از طرف دیگر توجه به این نکته ضروری است که برای ارزیابی سطوح بسیار بالای شناختی مانند تحلیل و ترکیب اطلاعات معمولاً از سؤالات کوته‌پاسخ استفاده نمی‌شود. در این مرحله باید در خصوص بخشی از محتوا و یا هدف آموزشی که قصد ارزیابی آن وجود دارد، تصمیم‌گیری شود و سپس سطح دانشی یا مهارت مورد ارزیابی (یادآوری، درک، کاربرد و یا ترکیبی از آنها) تعیین گردد. بهترین روش برای انتخاب ابزار ارزیابی متناسب با هدف و محتوای آموزشی، استفاده از بلوپرینت آزمون است. از آنجا که آزمون‌های کوتاه‌پاسخ در مقایسه با آزمون‌های بسته‌پاسخ تعداد سؤالات کمتری را در یک آزمون مورد سنجش قرار می‌دهند، لذا در هنگام نوشتن و انتخاب سؤالات باید دقت عمل بیشتری نمود. به این منظور باید سعی گردد که هر سؤال با توجه به یک هدف دقیق آموزشی نوشته شود.

همچنین هر سؤال باید یک هدف مهم یادگیری را شامل شود. از جمله مشکلات مربوط به آزمون‌های کوته‌پاسخ، استفاده از آن‌ها در سنجش هدف‌های سطوح پایین حیطه شناختی است. این موضوع منجر به تشویق فراگیران به یادگیری مطالب بی اهمیت و جزئی می‌شود. بنابراین در هنگام نوشتن این سؤالات سعی شود که اهداف مهم آموزشی مورد سنجش قرار گیرند.

سؤال ضعیف
انواع جهش‌ها را در توالی کدکننده (coding sequence) نام ببرید.
سؤال بهتر
در صورت بروز چه نوع جهشی در توالی کدکننده (coding sequence) یک ژن، تغییری در محصول پروتئینی آن ایجاد نمی‌شود؟

نگارش صورت سؤال

به طور کلی مراحل نگارش صورت سؤال کوته‌پاسخ به شکل سؤال، سطح پیامدهای مورد انتظار و محتوای مورد ارزیابی بستگی دارد. با این وجود مجموعه‌ای از اصول پایه‌ای وجود دارد که در هنگام طراحی این سؤالات باید به خاطر داشت. در ادامه به شرح دقیق این نکات به‌همراه ارائه مثال‌های کاربردی خواهیم پرداخت.

نکته قابل توجه در هنگام طراحی سؤالات کوته‌پاسخ آن است که متن سؤال باید واضح، شفاف و در برگیرنده اطلاعات لازم و اختصاصی باشد. یکی از مشکلات سؤالات کوته‌پاسخ این است که فراگیر دقیقاً همان پاسخی را که مد نظر طراح بوده

است، ننوشته است اما در عین حال نمی‌توان گفت که پاسخ او کاملاً اشتباه است. این مشکل معمولاً زمانی رخ می‌دهد که در صورت سؤال ابهام وجود داشته باشد برای رفع این مشکل باید تا جایی که امکان دارد صورت سؤال به صورت اختصاصی و شفاف نوشته شود. در واقع سؤال باید به گونه‌ای نوشته شود که باعث سردرگمی دانشجو نشود.

نمونه‌ای از سؤال کوته‌پاسخ ضعیف

در هنگام لقاح، غشای اسپرم چه سرنوشتی پیدا می‌کند؟

در مثال فوق، پاسخ مد نظر طراح سؤال، «قطعه قطعه می‌شود»، است. در صورتی که فراگیر به سؤال پاسخ «از بین می‌رود» یا «وارد نمی‌شود» بدهد، اگرچه به پاسخ مورد نظر اشاره نکرده است، با این حال نمی‌توان نتیجه گرفت که پاسخ ارائه شده نادرست است. بنابراین همواره باید به این نکته توجه داشت که صورت سؤالات باید به قدری واضح و خالی از ابهام باشد که از ارائه چنین پاسخ‌های دو پهلویی جلوگیری شود. به مثال دیگری که در ادامه ارائه شده است، توجه کنید.

نمونه‌ای از سؤال کوته‌پاسخ ضعیف

در شرایط مختلف فیزیولوژیک، میزان عبور آب[1] از غشای سلول‌ها از طریق کانال‌های آب تغییر می‌کند. این عمل با تغییر ..... کانال های آب رخ می‌دهد.

1. Aquaporins

در مثال فوق پاسخ صحیح «تعداد» است اما اگر دانشجو «تراکم» یا «مقدار» را در پاسخ بنویسد، نمی‌توان در خصوص درستی یا نادرستی پاسخ ارائه شده تصمیم گرفت. بنابراین توجه به این نکته ضروری است که در هنگام طرح سؤالات کوته‌پاسخ باید تا حد امکان صورت سؤال شفاف باشد تا منجر به سردرگمی فراگیران در ارائه پاسخ نشود.

طراحی سؤالاتی که فاقد هر گونه ابهام باشد، کار دشواری است. یکی از راه‌های اطلاع از این مشکل زمانی است که در جلسه آزمون، دانشجویان خواهان توضیح بیشتر در مورد سؤال از مراقبان هستند. در این شرایط به احتمال زیاد متن سؤال به اندازه کافی رسا و گویا نیست. بنابراین پایه سؤال باید حاوی تمام اطلاعاتی باشد که برای پاسخ به سؤال ضروری بوده و در نهایت به یک سؤال مشخص و واضح ختم شود.

شکل ضعیف سؤال کوته‌پاسخ

تغییراتی را که در نتیجه مواجه با نور شدید در چشم اتفاق می‌افتد، توضیح دهید.

شکل بهتر سؤال کوته‌پاسخ

در زمان برخورد نور شدید به چشم:
۱. مردمک‌ها دچار چه تغییری می‌شوند؟
۲. در عدسی چه اتفاقی می‌افتد؟
۳. توضیح دهید که نقش عضلات چشم در ایجاد تغییرات فوق چیست و چگونه منجر به این تغییرات می‌شوند؟

در سؤال فوق، در حالت اول، صورت سؤال به شکل کلی ارائه شده است و از این حیث پاسخ‌های متفاوتی از فراگیران مشاهده می‌شود که ممکن است تمام آن‌ها به طریقی صحیح باشد. در حالی که در حالت دوم، به دلیل محدود کردن صورت سؤال احتمال دریافت پاسخ‌های متنوع از سوی آزمون‌شوندگان کمتر خواهد بود.

به‌علاوه همان‌طور که پیشتر نیز به آن اشاره کردیم، پایه سؤال بهتر است یک مورد بالینی باشد. انجام این کار باعث می‌شود که سؤالات طراحی شده به شکل موثرتری به ارزیابی سطوح بالای حیطه شناختی (درک و کاربرد مطالب) بپردازند.

به نمونه مثال‌هایی که در زیر آمده است توجه کنید.

نمونه‌ای از سؤال کوته‌پاسخ با سناریو

شما به عنوان پزشک عمومی در اورژانس حضور دارید. خانم ۳۲ ساله بعد از دو هفته تعطیلات با شکایت از سردرد و قرمزی چشم به شما مراجعه می‌کند. او در هفته ۳۲ بارداری است. در بررسی تاریخچه پزشکی مشخص می‌گردد که ایشان از تورم پاها و انگشتان و افزایش وزن ناگهانی در هفته گذشته رنج می‌برد.

۱. محتمل‌ترین تشخیص برای وی چیست؟
۲. در معاینه فیزیکی به چه مواردی باید توجه داشته باشید؟ (به سه مورد اشاره کنید).
۳. چه تست‌های تشخیصی دیگری به عنوان پزشک عمومی می‌توانید به منظور تایید یا رد تشخیص‌های احتمالی خود درخواست کنید؟

نمونه‌ای از سؤال کوته‌پاسخ با سناریو

شما به عنوان متخصص در بخش کودکان کار می‌کنید. از شما درخواست می‌شود که نوزاد یک ماهه‌ای را به علت افزایش تعداد تنفس و رتراکسیون زیر دنده‌ای بررسی نمایید. نوزاد ۳۵ هفته و از طریق سزارین الکتیو متولد شده است و مادر وی ۲۹ ساله است. اندیکاسیون سزارین، فشارخون غیرقابل کنترل بوده است. مادر در طول دوره قبل از تولد نوزاد به طور منظم تحت کنترل بوده و دیابت بارداری و پره‌اکلامپسی داشته است.

۱. محتمل ترین تشخیص برای نوزاد چیست؟ (دو مورد نام ببرید).
۲. برای هر یک از تشخیص ها یک مکانیزم پاتوفیزیولوژیک اصلی لیست نمایید؟
۳. در این وضعیت بررسی‌های اولیه‌ای که باید انجام شوند، چه هستند؟ (سه مورد نام ببرید).

سناریوی بالینی اختصاص به ارزیابی علوم بالینی ندارد و از آن از برای سنجش دانش علوم پایه فراگیر نیز می‌توان استفاده کرد:

نمونه‌ای از سؤال کوته‌پاسخ با سناریو برای ارزیابی کاربرد علوم پایه

بیماری با شکایت کاهش قدرت عضلات پا و دست به درمانگاه مراجعه کرده است. او می‌گوید که ۵ روز پیش به دنبال خوردن غذا در رستوران، حالت تهوع، استفراغ و اسهال داشته که به تدریج بهتر شده بود. اما دوباره اسهالش شروع شده و همچنین از دیشب متوجه شده که نمی‌تواند خوب راه برود. از امروز صبح احساس ضعف در عضلات بازو نیز پیدا کرده است. پزشک درمانگاه بعد از انجام معاینات، با شک به سندرم گیلن‌باره او را برای بررسی بیشتر در بخش نورولوژی بستری می‌کند. اگر برای مشاهده زیر میکروسکوپ، از انتهای اعصاب محیطی درگیر، نمونه‌ای تهیه شود، تجمع ماکروفاژ و اکسون‌های دمیلینه شده دیده می‌شود.
در جریان این بیماری، علت تجمع ماکروفاژها در محل آسیب چیست؟

در مورد سؤال فوق، چنانچه دانشجو به هر یک از موارد تعیین شده (شامل پاکسازی منطقه از قطعات میلین، فاگوسیت میلین یا برداشت میلین تخریب شده) اشاره کند، نمره کامل را دریافت خواهد کرد.

البته همواره باید به این نکته توجه کرد که در زمان استفاده از سناریو و موارد بالینی در تنه سؤال، در ادامه بهتر است چندین سؤال مرتبط با سناریو طرح گردد و از طرح یک سؤال خودداری شود. زیرا علاوه بر آنکه یکی از اهداف استفاده از سناریو در تنه سؤال، سنجش سطوح بالاتر حیطه شناختی است، از این طریق می‌توان تعداد سؤالات بیشتری را در یک جلسه آزمون ارزیابی نمود که بالتبع انجام این کار به بهبود پایایی و بالتبع روایی آزمون کمک بیشتری خواهد نمود. بدین ترتیب سؤال بالا تنها به عنوان مثال به شکل تک سؤالی ارائه شده است و در عمل، حتماً در ادامه یک سناریو چندین سؤال مرتبط مطرح می‌گردد.

سؤال را باید به گونه‌ای پرسید که یک جواب کوتاه (یک یا چند کلمه یا یک عبارت) داشته باشد. حتی می‌توان تعداد کلمات مورد نظر را مشخص کرد. مثلاً انتهای سؤال نوشت: «سه مورد را ذکر کنید». این کار در بسیاری از موارد دانشجویان را از سردرگمی در حجم پاسخ مورد نظر نجات می‌دهد.

نمونه‌ای از سؤال کوته‌پاسخ

علت تفاوت در میزان بقای نوزادان با ناهنجاریهای آمفالوسل و گاستروشیزی در چیست؟ (ذکر دو دلیل کافی است).

به علاوه تا حد امکان باید تلاش کرد که صورت سؤال مثبت باشد. به این دلیل که تحقیقات نشان داده است که سؤالات مثبت در مقایسه با شکل منفی آنها توانایی بیشتری در ارزیابی اهداف مهم آموزشی دارند. در صورتی که طرح سؤال به شکل منفی اجتناب‌ناپذیر است، حتماً سعی کنید عبارت منفی را از طریق خط کشیدن یا برجسته کردن مشخص کنید تا توجه فراگیر کاملاً جلب شود.

نمونه‌ای از سؤال کوته‌پاسخ منفی

پلاسمالوژن یک فسفولیپید غشاء سلول‌های عصبی و ماهیچه‌ای بوده و دارای یک عامل شاخص است که در دیگر فسفولیپیدها وجود ندارد. این عامل شاخص چیست؟

به صورت کلی در هنگام نوشتن صورت سؤال همواره نکات زیر را باید در ذهن داشت:

- **نگارش صورت سؤال به صورت کاملاً روشن:** همان‌طور که در بالا اشاره شد، یکی از مشکلات مربوط به سؤالات کوته‌پاسخ این است که سؤالات طراحی شده علی‌رغم آنکه از نقطه‌نظر استادان یک پاسخ منحصر به فرد دارد، با این حال آزمون‌شوندگان پاسخ‌های متعددی به آن می‌دهند. به منظور رفع این مشکل، سؤالات باید به نحوی طراحی شوند که پاسخ‌های مشخصی برای آنها وجود داشته باشد. انجام این کار همچنین باعث می‌شود که در هنگام تصحیح برگه‌های امتحانی مشکلات کمتری ایجاد شود.
- **عدم کپی سؤال عیناً از روی مطالب کتاب:** نقل مستقیم سؤالات کوته‌پاسخ از کتاب منجر به تشویق فراگیران به یادگیری طوطی‌وار مطالب می‌شود. به منظور جلوگیری از بروز چنین مشکلی سعی کنید در هنگام طراحی سؤالات کوته‌پاسخ سؤالات را به شکل کاربردی طرح نمایید.
- **طراحی صورت سؤال به شکل سناریو:** یکی از مواردی که برای سنجش کاربرد دانش، توصیه می‌شود استفاده از یک سناریو در پایه سؤال است. با این رویکرد، علاوه بر اینکه تاکسونومی سؤال افزایش می‌یابد و سطوح شناختی بالاتر مورد ارزیابی قرار می‌گیرد، سؤال برای داوطلب از نظر بالینی مرتبط‌تر با محیط کار واقعی به نظر می‌رسد و دارای روایی صوری بیشتری است. نکته حائز اهمیت این است که:

تا حد امکان سناریوی سؤال، مرتبط با وظایف حرفه‌ای که دانشجویان در آینده برعهده خواهند داشت، طراحی شود و در ادامه سعی شود سؤالات مستقیماً به سناریوی طراحی شده ارتباط داده شوند.

باید حتماً از این موضوع اطمینان حاصل کرد که بدون مطالعه سناریو، سؤالات قابل پاسخگویی نباشند.

محتوای هر سؤال باید مستقل از محتوای سؤالات دیگر باشد.

باید دقت نمود که پاسخ دادن به یک سؤال نباید پیش‌شرط پاسخ به سؤال دیگر بوده یا حتی در پاسخ دادن به سؤال بعد آزمون‌دهنده را راهنمایی نماید. چرا که پاسخ غلط به سؤال اول الزاماً پاسخ ناصحیح به سؤال دوم را در پی دارد و این با اصول ارزیابی عادلانه در تضاد است.

- **ذکر واحد مقیاس و مشخص کردن میزان دقتی که در محاسبات باید رعایت شود:** اغلب آزمون‌شوندگان در سؤالات محاسباتی با این مسأله روبرو خواهند شد که دقت انجام محاسبات به چه میزان باید باشد. به منظور ایجاد هماهنگی بین پاسخ‌های ارائه شده، در صورت سؤال میزان دقت پاسخ مورد انتظار باید مشخص گردد. در صورت عدم توجه به این موضوع ممکن است در هنگام تصحیح این سؤالات نتوان در خصوص قبولی یا رد پاسخ ارائه شده تصمیم‌گیری کرد. به علاوه باید در سؤال قید گردد که آیا ذکر روند انجام محاسبات و نوشتن فرمول هم مورد نیاز

است یا تنها ارائه پاسخ نهایی کفایت می‌کند.

- **توضیح به آزمون‌شوندگان در خصوص شکل ارائه پاسخ:** این فرصت باید برای آزمون‌شوندگان فراهم شود تا آشنایی لازم را در خصوص نحوه ارائه پاسخ‌ها پیدا کنند. شرح این موضوع به ارائه پاسخ‌های یک دست و هماهنگ آنها و متعاقب آن نمره‌دهی آسان‌تر سؤالات کمک زیادی خواهد نمود.
- **قرار ندادن تعداد زیاد جای خالی در سؤال‌های تکمیلی:** انجام این کار منجر به پیچیدگی غیرمنطقی سؤال و ابهام در درک منظور آن خواهد شد که می‌تواند مستقیماً بر پاسخ‌های ارائه شده تاثیر بگذارد. به این منظور در هنگام طراحی سؤالات تکمیلی باید سعی شود که تعداد زیادی جای خالی در یک سؤال منظور نشود.
- **قرار دادن جای خالی سؤال‌های تکمیلی در قسمت پایانی:** هنگام طراحی سؤالات تکمیلی بهتر است که جای خالی در انتهای سؤال باشد. این کار منجر به درک بهتر سؤال از سوی دانشجو خواهد شد. نوشتن سؤالات تکمیلی که در آنها جای خالی در اوایل سؤال قرار دارد، منجر به ابهام می‌شود.
- **در نظر گرفتن جای خالی سؤال‌های تکمیلی به صورت یک اندازه:** اغلب استادان در هنگام طراحی سؤالات نوع تکمیلی دچار این اشتباه می‌شوند که با توجه به حجم پاسخ مورد نظر هر سؤال جای خالی برای آن در نظر می‌گیرند. این مسأله از آن جهت قابل توجه است که در صورت کشف این موضوع توسط آزمون شوندگان، ممکن است آن‌ها را به حدس زدن پاسخ ترغیب نماید.
- **استفاده از سؤال‌های پرسشی و هدایتی به جای سؤال‌های تکمیلی:** به دلیل طراحی آسان‌تر سؤالات پرسشی و هدایتی و ابهام کمتر آنها در مقایسه با سؤالات تکمیلی در آزمون‌های کوته‌پاسخ تا حد امکان بهتر است سعی شود که از این نوع از سؤالات کوته‌پاسخ استفاده شود.
- **اجتناب از سرنخ‌های دستوری و ساختاری:** وجود سرنخ‌های دستوری و ساختاری منجر به راهنمایی دانشجویان و کمک به آن‌ها در حدس زدن پاسخ درست خواهد شد لذا باید از آن اجتناب نمود.

تصمیم‌گیری در خصوص پاسخ‌های صحیح و نحوه نمره‌دهی

از آنجا که در سؤال کوته‌پاسخ خود فراگیر باید جواب را بسازد، احتمالاً پاسخ‌های متنوعی از دانشجویان مختلف دریافت خواهد شد که اگرچه کاملاً صحیح نیستند، به درجاتی واجد پاسخ درست نیز هستند. به همین منظور ضروری است که نحوه نمره‌دهی به هر سؤال از قبل مشخص گردد. این موضوع مخصوصاً وقتی حائز اهمیت است که تعداد دانشجویان زیاد است یا بیش از یک مصحح وجود دارد. گاهی چند جواب صحیح وجود دارد بنابراین پاسخ‌های احتمالی صحیح باید از قبل مشخص شوند تا هنگام تخصیص نمره مشکل ایجاد نشود.

به عنوان نمونه در مثال زیر، چنانچه دانشجو به هر یک از دو مورد هپاتوگلوبین و LDL اشاره کند، نمره کامل را دریافت خواهد کرد.

نمونه‌ای از سؤال کوته‌پاسخ

آقای ۲۴ ساله بدون سابقه بیماری خاص به علت زردی مراجعه کرده است. تست‌های عملکرد کبدی نرمال هستند به جز بیلی روبین که ۴۵ µmol/L است. بیلی روبینوری ندارد. شما به این فکر می‌کنید که آیا ممکن است این مساله ناشی از همولیز باشد. یک تست بیوشیمیایی نام ببرید که برای رد تشخیص همولیز به شما کمک می‌کند (۰/۵ نمره).

به مثال دیگری در این مورد توجه کنید:

نمونه‌ای از سؤال کوته‌پاسخ

علت تفاوت در میزان بقای نوزادان با ناهنجاریهای آمفالوسل و گاستروشیزی در چیست؟ (ذکر دو دلیل کافی است). (۱ نمره)

در مثال فوق، طراح سؤال در الگوی پاسخ مشخص کرده است که جواب‌های «ناهنجاری‌های کروموزومی»، «ناهنجاری‌های قلبی عروقی» و «ناهنجاری‌های عصبی» هر سه قابل قبول هستند. یعنی اگر دانشجو به دو مورد از این سه بیماری اشاره کند، نمره کامل را دریافت می‌کند.

به طور کلی در هنگام نمره‌دهی سؤالات کوته‌پاسخ باید به موارد زیر توجه کرد:

- **پیش‌بینی پاسخ‌های احتمالی و تصمیم‌گیری قبلی در مورد پذیرفتن یا نپذیرفتن آنها:** این کار برای یکسان کردن فرایند تصحیح برگه‌ها ضروری است. در این حالت اگر آزمون‌شوندگان پاسخ‌های مترداف صحیح دیگری را ارائه نمایند، به راحتی می‌توان آن‌ها را از پاسخ‌های نادرست تفکیک نمود و بدین ترتیب نمره‌دهی منصفانه‌تر خواهد بود.
- **اشاره به نکات لازم برای تصحیح در راهنما:** انجام این کار باعث افزایش عینیت فرایند نمره‌دهی خواهد شد و حتی در شرایطی که فردی به جز شما ارزیابی و نمره‌دهی پاسخ‌های آزمون شوندگان را انجام می‌دهد، همچنان پایایی نمرات حفظ می‌گردد.
- **در نظر گرفتن جای مشخص برای پاسخ دانشجویان:** قرار دادن یک کادر در انتهای سمت چپ هر سؤال تا دانشجو پاسخ خود را در آن وارد کند، به سهولت تصحیح برگه‌ها کمک می‌کند. البته دو دیدگاه متضاد در این خصوص وجود دارد. در برخی از منابع به این موضوع اشاره شده است که به منظور به حداقل رساندن شانس حدس زدن توسط آزمون شوندگان، حجم فضای در نظرگرفته شده برای پاسخ باید در تمام سؤالات یکسان در نظر گرفته شود اما دیدگاه متضاد که ذکر می‌کند فضای پاسخ برای هر سؤال باید متناسب با حجم مورد انتظار پاسخ در نظر گرفته شود، و بر این موضوع مورد تاکید دارد که می‌کند که آزمون‌شوندگان نباید در ارائه حجم پاسخ به سؤالات مختلف دچار سردرگمی شوند. به نظر می‌رسد در صورتی که سؤالات طراحی شده در دسته سؤالات دانشی جای می‌گیرند، پیاده سازی دیدگاه اول امکان‌پذیر باشد اما در شرایطی که سؤالات آزمون در دسته سؤالات تفسیری قرار می‌گیرند، تعیین حجم تقریبی پاسخ بهتر باشد.
- **مشخص کردن ارزش هر سؤال و نمره آن برای دانشجویان:** انجام این کار باعث می‌شود فراگیران با توجه به ارزش هر سؤال وقت آزمون را تنظیم کنند و حتی ممکن است به این طریق به آن‌ها در خصوص حجم پاسخ مورد نیاز برای هر سؤال راهنمایی ارائه شود. توجه شود که اگر جواب به گونه‌ای است که بیش از یک کلمه یا عبارت باید در پاسخ آن ذکر شود، نمره سؤال طوری در نظر گرفته شود که به راحتی قابل تقسیم باشد.
- **تصحیح سؤالات فقط توسط افراد مطلع:** گاهی از دانشجو فقط یک جواب خواسته شده است اما ممکن است چند جواب صحیح وجود داشته باشد. در این حالت در صورتی که تصحیح سؤالات به فرد دیگری سپرده شده باشد، در صورت ارائه پاسخ‌های مترادف صحیح ممکن است حق دانشجویان ضایع شود. اگر استفاده از مصححان دیگر اجتناب ناپذیر است، حتماً در برگه راهنمای تصحیح سؤالات باید مجموعه پاسخ‌های مترادف دیگری که می‌توانند صحیح باشند، ذکر شوند. به عبارت دیگر، پاسخ‌های احتمالی صحیح از قبل مشخص شوند تا در هنگام تصحیح نمره مشکل ایجاد نشود.

مرور و بازبینی سؤالات

چهارمین مرحله از فرایند طراحی سؤالات تشریحی، نقد و مرور سؤالات طرح شده است. برای انجام این کار می‌توان از چک‌لیست ارزیابی سؤالات تشریحی که در جدول ۲-۱۲ ارائه شده است، استفاده کرد. بدین ترتیب بسیاری از مواردی که ممکن است در هنگام طراحی سؤال از قلم افتاده باشد، به شکل دقیق مشخص و اصلاح خواهد شد.

جدول ۲-۱۲: چک‌لیست ارزیابی سؤالات کوتاه‌پاسخ

ردیف	سؤال	
۱	هر سؤال با توجه به یک هدف مهم آموزشی طراحی شده است؟	بله
۲	برای طراحی سؤال‌ها از بلوپرینت استفاده شده است؟	بله
۳	آیا استفاده از سؤال کوته‌پاسخ برای ارزیابی این هدف آموزشی مناسب است؟	بله
۴	آیا سطح دانشی سؤال (تاکسونومی بلوم) منطبق با هدف آموزشی است؟	بله
۵	سطح دشواری سؤال با توجه به سطح آموزشی ارائه شده متناسب است؟	بله
۶	سؤال کاملاً روشن، صریح و بدون ابهام طراحی شده است؟	بله
۷	سؤال بدون نقل مستقیم جملات کتاب نوشته شده است؟	بله
۸	در سؤال سرنخ‌های دستوری و ساختاری وجود ندارد؟	بله
۹	مقیاس اندازه‌گیری و درجه دقت مورد انتظار برای پاسخ‌دهی به سؤال کاملاً مشخص شده است؟	بله
۱۰	در سؤال‌های نوع تکمیلی: جای خالی در قسمت آخر سؤال قرار داده شده است؟ استفاده از جاهای خالی متعدد اجتناب شده است؟ طول جای خالی در سؤالات مختلف یکسان در نظر گرفته شده است؟عبارت و موضوع اصلی و نه موارد جزئی و کم اهمیت مطرح شده است؟	بله
۱۱	هر سؤال طوری طراحی شده است که تنها یک پاسخ درست برای آن وجود دارد؟	بله
۱۲	راهنمای نمره‌دهی پاسخ‌ها (شامل پاسخ‌های قابل قبول و ارزش نمره هریک) از قبل تهیه شده است؟	بله

تصحیح سؤالات کوته‌پاسخ

همان‌طور که پیشتر نیز اشاره شد، یکی از ویژگی‌های سؤالات کوته‌پاسخ، کوتاه و دقیق بودن پاسخ ارائه شده است. این ویژگی امکان ارزیابی و نمره‌دهی عینی‌تر سؤالات کوته‌پاسخ را فراهم می‌سازد. با وجود این، به منظور یکسان‌سازی فرایند ارزیابی سؤالات کوته‌پاسخ و متعاقب آن افزایش میزان پایایی بین ارزیابان، باید حتماً پیش از اجرای آزمون الگوی پاسخ تهیه شود. انجام این کار نه تنها به استادان در فرایند نمره‌دهی کمک خواهد نمود، بلکه همچنین در صورت ارائه الگوی پاسخ به دانشجویان بعد از اجرای آزمون، می‌توان به تقویت اثرآموزشی آزمون کمک شایانی کرد. به منظور افزایش پایایی نمره‌دهی بین ارزیابان علاوه بر تهیه کلید پاسخ باید روبریک[1] تصحیح برگه‌های امتحانی نیز تهیه گردد. در این حالت با توجه به نوع پاسخ دریافتی در خصوص شکل طراحی روبریک باید تصمیم گرفته شود. بر این اساس در مواردی که پاسخ سؤال به شکل لیستی از کلمات است، طراحی روبریک در قالب جدول ۳-۱۲ توصیه می‌گردد.

1. Rubric

جدول ۳-۱۲: نمونه روبریک نمره‌دهی سؤال کوته‌پاسخ ۵ نمره‌ای

میزان پاسخ دانشجو	نمره
به تمامی موارد اشاره شده است.	۵ نمره
به ۷۵ درصد موارد اشاره شده است.	۴ نمره
به ۵۰ درصد موارد اشاره شده است.	۳ نمره
به ۲۵ درصد موارد اشاره شده است.	۲ نمره
به ۱۰ درصد موارد اشاره شده است.	۱ نمره

به عنوان مثال در مواردی که فراگیر باید به شش مورد در پاسخ خود اشاره کند، ارزیاب بر اساس روبریک بالا در خصوص نحوه تخصیص نمره به پاسخ‌های ارائه شده از سوی فراگیران تصمیم‌گیری می‌کند. البته در مواردی ممکن است با توجه به کل پاسخ‌های ارائه شده از سوی فراگیران، مصحح تصمیم بگیرد که تا حدودی سهل‌گیرانه‌تر عمل کند و به دانشجویانی که به چهار تا شش مورد از پاسخ مدنظر اشاره کرده‌اند، کل نمره تعلق گیرد. تصمیم‌گیری در این مورد باید واضح و دقیق باشد تا در صورتی که بیش از یک مصحح کار ارزیابی برگه‌های امتحانی را بر عهده گرفتند، این هماهنگی بین آن‌ها وجود داشته باشد.

سودمندی سؤالات کوته‌پاسخ

روایی سؤالات کوته‌پاسخ

از جمله نقاط قوت سؤالات کوته‌پاسخ آن است که زمان پاسخگویی به آنها در مقایسه با سؤالات تشریحی کوتاه‌تر است، بنابراین آزمون می‌تواند دامنه وسیع‌تری از محتوای دوره را از طریق پرسش سؤالات مهم و مجزا پوشش دهد که این موضوع منجر به کاهش بروز مشکل ویژگی محتوا و در نتیجه افزایش روایی محتوایی آزمون خواهد شد (امین ۲۰۰۶). البته سؤالات کوته‌پاسخ در مقایسه با سؤالات چندگزینه‌ای، از روایی محتوایی پایین‌تری برخوردار هستند. در صورتی که استادان از مهارت لازم در طرح سؤالات مناسب برخوردار باشند، می‌توان سؤالات کوته‌پاسخ را به شکل سناریوی بالینی طرح کرد که به این طریق روایی صوری آزمون نیز بهبود می‌یابد. همچنین روایی ملاکی سؤالات کوته‌پاسخ با سایر انواع ابزارهای ارزیابی در مطالعات متعدد بررسی شده است که در زیر نمونه‌ای از مطالعات موجود ارائه شده است.

گارسیا و همکاران [1] ۲۰۱۲

در این مطالعه تاثیر سؤالات کوته‌پاسخ و چندگزینه‌ای بر نمرات دانشجویان در سه دوره آموزشی بیولوژی بین سال‌های ۲۰۰۸ و ۲۰۱۰ با شرکت ۵۹۳ دانشجو مورد بررسی قرار گرفت. آزمون شامل سؤالات کوته‌پاسخ و چندگزینه‌ای بود که هر دو نوع سؤالات به نحوی طراحی شده بودند که بتوانند محتوای مشابهی را پوشش دهند. همبستگی بین روش‌های مختلف با استفاده از ضریب پیرسون محاسبه شد.
نتایج حاصل از مطالعه نشان داد که دانشجویان نمرات پایین‌تری را در سؤالات چندگزینه‌ای نسبت به نوع کوته‌پاسخ کسب کردند. تفاوت بین نمرات سؤالات چندگزینه‌ای و کوته‌پاسخ در تمامی دوره‌های آموزشی معنادار بود. با این حال، همبستگی بین نمرات چندگزینه‌ای و کوته‌پاسخ در هر سه دوره آموزشی بالا بود. یافته‌های حاصل از این مطالعه نشان داد که اولا روش ارزیابی یک عامل بسیار مهم و تعیین‌کننده در نمراتی است که فراگیران از آزمون کسب می‌کنند. ثانیا، علی رغم وجود همبستگی بین نمرات فراگیران در سؤالات کوته‌پاسخ و سؤالات چندگزینه‌ای، با این حال تفاوت معناداری در میانگین نمرات هر دو روش وجود دارد.

1. García et al

رسائیان و همکاران ۱۳۸۱

این بررسی با استفاده از سؤالات امتحانی دوره‌های علوم پایه، فیزیوپاتولوژی و بالینی دانشکده پزشکی دانشگاه‌های علوم پزشکی و خدمات بهداشتی-درمانی شهید بهشتی و ایران در ا نیمسال دوم سال تحصیلی ۷۶-۱۳۷۵ با حضور ۱۳۷۲ نفر دانشجو انجام شد. برای طراحی سؤالات با هماهنگی قبلی با مدرسان و پس از توضیح در مورد اهداف طرح و چگونگی طرح سؤال، ۱۰ نکته علمی به شکل مجموعاً ۳۰ سؤال از نوع تاکسونومی II نوشته شد. برای کاستن میزان احتمال استفاده از شانس در سؤالات چندگزینه‌ای و درست-نادرست از نمره منفی استفاده گردید. تصحیح برگه‌های امتحانی با استفاده از کلید سؤالات و توسط دو نفر از همکاران انجام شد و نحوه نمره‌دهی مجدد توسط یک نفر مورد بررسی قرار گرفت. نتایج حاصل از مطالعه نشان داد که کمترین زمان، صرفِ تصحیح برگه‌های امتحانی مربوط به روش درست-نادرست و بیشترین مدت را روش کوته‌پاسخ به خود اختصاص داد. تحلیل نمرات دانشجویان، پس از حذف سؤالات با درجه تمایز کمتر از ۰/۳ نشان داد که معتبرترین روش مورد بررسی، سؤالات کوته‌پاسخ بودند (۰/۶۸۵ = r). تعداد پاسخ‌های درست در روش کوته‌پاسخ کمترین (۵۳ درصد) و در روش صحیح-غلط بیشترین (۶۵ درصد) بود. بر اساس نتایج حاصل از مطالعه، محققان پیشنهاد کردند که از سؤالات کوته‌پاسخ و درست-نادرست به همراه سؤالات چندگزینه‌ای در آزمون‌ها استفاده شود.

پایایی سؤالات کوته‌پاسخ

در اکثر مواقع ارزیاب در هنگام تصحیح این سؤالات به منظور تصمیم‌گیری در خصوص آنکه آیا پاسخ ارائه شده رضایت‌بخش است، نیازمند انجام یک قضاوت ذهنی است که البته میزان آن به طول پاسخ ارائه شده و پیچیدگی سؤال بستگی دارد. هر چه طول پاسخ و پیچیدگی آن بیشتر باشد، تصحیح سؤال نیازمند قضاوت ذهنی بیشتر ارزیاب است که این موضوع بالتبع بر روی پایایی آزمون تاثیرگذار خواهد بود. به این دلیل که پاسخ‌های ارائه شده در سؤالات کوته‌پاسخ دقیق و کوتاه هستند، نمره‌دهی آنها در مقایسه با سؤالات تشریحی آسان‌تر است. به طور کلی هر چه پاسخ‌های مربوط به سؤالات کوته‌پاسخ کوتاه‌تر باشد، عینیت فرایند نمره‌دهی افزایش خواهد یافت و از طریق امکان طرح سؤالات بیشتر، پایایی آزمون نیز بهبود خواهد یافت (ذولفقاری ۱۳۷۹).

با این حال، سعی می‌شود پایایی آزمون از طریق استفاده از پاسخ‌های استاندارد از پیش تعیین‌شده و طراحی راهنمای نمره‌دهی برای هر سؤال افزایش یابد (امین ۲۰۰۶، دانشگاه رابرت گوردن[1] ۱۹۹۷). به نحوی که استفاده از یک الگوی ساختارمند و از قبل تعیین شده جهت تصحیح سؤالات منجر به افزایش عینیت آزمون و به دنبال آن افزایش پایایی خواهد شد (امین ۲۰۰۶). پژوهشگران همچنین به این موضوع اشاره کرده‌اند که پایایی مساوی یا بالاتر آزمون می‌تواند با تعداد کمتر سؤالات کوته‌پاسخ در مقایسه با سؤالات درست–نادرست کسب گردد (تن‌کیت و همکاران[2] ۱۹۹۶، تن‌کیت ۱۹۹۷، رادمیکرز ۲۰۰۵).

تن کیت ۱۹۹۷

برنامه درسی دوره پزشکی عمومی در دانشگاه آمستردام شامل ۹ ساعت آزمون کتبی حل مسأله بود که در آن با طرح بیش از ۲۰۰ مورد بالینی، محتوای آموزشی دوره به خوبی پوشش داده می‌شد. از سؤالات درست-نادرست (۹۰ سؤال) و سؤال کوته‌پاسخ (حدود ۱۳۰ سؤال) در آزمون کتبی استفاده شد. سؤالات درست-نادرست به وسیله ماشین و سؤالات کوته‌پاسخ به صورت دستی توسط ۳۰ تا ۴۰ متخصص و بر اساس الگوی پاسخ تصحیح شدند.

نتایج حاصل از مطالعه نشان‌دهنده پایایی بالا در تمام آزمون‌ها بود (آلفای کرونباخ برابر با ۰/۹۱). زمانی که سؤالات درست-نادرست و کوته‌پاسخ از یکدیگر تفکیک شدند، تفاوت در پایایی مشاهده شد. سؤالات کوته‌پاسخ پایایی مشابه با آزمون کلی داشتند در حالی که سؤالات درست-نادرست پایایی پایین‌تری داشتند که از ۰/۳۴ تا ۰/۵۳ متغیر بود. پنج تا شش برابر سؤالات درست-نادرست موجود برای رسیدن به پایایی برابر با کل آزمون نیاز بود. روایی همزمان سؤالات درست-نادرست و سؤالات کوته‌پاسخ بین ۰/۷۳ تا ۰/۹۵ بود. محققان بر اساس یافته‌های مطالعه نتیجه‌گیری کردند که سؤالات کوته‌پاسخ از نظر کیفیت و کارآمدی نسبت به سؤالات درست-نادرست بهتر هستند.

مایرز و همکاران[1] ۲۰۰۱

این پژوهشگران در مطالعه خود به بررسی نمرات حاصل از خودارزیابی دانشجویان پزشکی در استفاده از سؤالات کوته‌پاسخ پرداختند. بر اساس نتایج حاصل از مطالعه مشخص گردید که همبستگی بین نمرات مصححان و دانشجویان در هر یک از سؤالات کوته‌پاسخ و نمره کلی آزمون بالا بود (ضریب همبستگی بین ۰/۹۱ تا ۰/۷۷) که این ارتباط از نظر آماری نیز معنادار بود. همچنین هیچ تفاوت معنادارِی بین میانگین نمرات دانشجویان و مصححان در هر یک از سؤالات کوته‌پاسخ و نمره کلیِ آزمون مشاهده نشد. با این حال دانشجویان معمولاً در درک ارزش خودارزیابی به عنوان وسیله‌ای برای دریافت بازخورد از عملکردشان و متعاقباً یک فرصت برای یادگیری موفق نبودند و آن را یک فرایند استرس‌زا درک کردند.

1. Mires et al.

1. Robert Gordon University
2. Ten Cate

مقبولیت سؤالات کوته‌پاسخ

سؤالات کوته‌پاسخ مزایای متعددی دارند که باعث جذابیت زیاد این سؤالات در بین ارزیابان و فراگیران شده است. از آنجا که امکان ارزیابی توانایی درک و کاربرد مطالب به وسیله این سؤالات بدون ارائه راهنما و سرنخ به فراگیران میسر است، استفاده از سؤالات کوته‌پاسخ در بین ارزیابان از مقبولیت بالایی برخوردار است. به‌علاوه یکی دیگر از دلایل محبوبیت بالای این سؤالات در بین ارزیابان، قدرت تمیز بیشتر آنها در افتراق مابین فراگیران ضعیف و قوی است. فندرسون و همکاران[1] بر اساس مطالعات خود نتیجه‌گیری کردند که این سؤالات در مقایسه با سؤالات صحیح–غلط و چندگزینه‌ای، قدرت بیشتری در افتراق دانشجویان خوب از دانشجویان مرزی دارند (فندرسون و همکاران ۱۹۹۷).

هزینه و قابلیت اجرای سؤالات کوته‌پاسخ

در حالی‌که زمان و هزینه طراحی و نمره‌دهی سؤالات کوته‌پاسخ در مقایسه با سؤالات تشریحی کمتر است، اما در مقایسه با آزمون‌های بسته‌پاسخ از قبیل سؤالات چندگزینه‌ای و سؤالات درست– نادرست، به زمان بیشتری برای تصحیح و نمره‌دهی نیاز دارند. بنابراین هزینه اثربخشی این سؤالات بستگی به موازنه بین نقاط قوت و ضعف آن دارد.

تاثیر آموزشی سؤالات کوته‌پاسخ

در مورد تاثیر آموزشی سؤالات کوته‌پاسخ بر یادگیری فراگیران همچنین می‌توان گفت اگر سؤالات به ارزیابی سطوح پایین حیطه شناختی بپردازند، دانشجویان بیشتر به سمت حفظ کردن مطالب کتاب روی می‌آورند و مهارت‌های حل مسأله و تفکر در آنها تقویت نخواهد شد. با این حال معمولاً سؤالات کوته‌پاسخ نسبت به آزمون‌های بسته‌پاسخ چندگزینه‌ای، درست–نادرست و جورکردنی اطلاعات تشخیصی بیشتری در اختیار استادان می‌گذارند و به‌همین دلیل در اکثر مواقع به شکل آزمون‌های تکوینی در طول دوره استفاده می شوند (سیف۱۳۹۴). توصیه می‌شود، به منظور دستیابی به اثرات آموزشی بالاتر در یادگیری فراگیران سؤالات کوتاه‌پاسخ با دقت فراوان طراحی گردند. برخی از مطالعات تجربی به بررسی تاثیر این نوع از سؤالات بر یادگیری فراگیران پرداختند.

رامراج[1] ۲۰۱۱

این مطالعه به بررسی تاثیر استفاده از آزمون‌های کلاسی معلم‌ساخته بر میزان ماندگاری اطلاعات فراگیران پرداخت. ۸۴ دانشجوی شرکت کننده در این مطالعه به صورت تصادفی به سه گروه دانشجویانی که از طریق سؤال کوتاه‌پاسخ ارزیابی می‌شدند، دانشجویانی که از طریق سؤالات چندگزینه‌ای ارزیابی می‌شدند و دانشجویانی که هیچ آزمون کلاسی دریافت نکردند، تقسیم شدند. در این تحقیق، آزمون اولیه بلافاصله بعد از آموزش و آزمون مشابه بعدی برای تعیین میزان ماندگاری اطلاعات سه هفته بعد صورت گرفت. آزمون دوم شامل سؤالات چندگزینه‌ای و کوته‌پاسخ بود که برای هر سه گروه دانشجویان اجرا گردید.
میانگین نمره آزمون دوم که به ارزیابی میزان به خاطر سپاری فراگیران می‌پرداخت در سؤالات چندگزینه‌ای برابر با ۱۰/۹۷ و برای سؤالات کوته‌پاسخ برابر با ۸/۴۲ و گروه کنترل برابر با ۶/۷۱ بود. تحلیل واریانس بین میانگین نمرات سه گروه از نظر آماری تفاوت معناداری را نشان داد. یافته‌های پژوهش نشان داد که آزمون‌های کلاسی می‌تواند تاثیر مثبتی بر ماندگاری اطلاعات هم در سؤالات کوته‌پاسخ و چندگزینه‌ای داشته باشد، هر چند سؤالات چندگزینه‌ای از این لحاظ بهتر هستند.

1. Ramraje

سؤالات رایج در مورد سؤالات کوته‌پاسخ

استفاده از کدام یک از انواع سؤالات کوته‌پاسخ بهتر است؟

از جمله سؤالاتی که همواره در خصوص سؤالات کوته‌پاسخ مطرح می‌شود آن است که طراحی کدام یک از انواع

1. Fenderson et al

سؤالات کوته‌پاسخ بهتر است. به طور کلی در منابع متعدد به این موضوع اشاره شده است که سؤالات نوع پرسشی و هدایتی به دلیل آنکه کمتر باعث ایجاد سردرگمی و ابهام دانشجویان می‌شوند، بهتر از سؤالات نوع تکمیلی هستند.

یکی دیگر از سؤالات فراروی ارزیابان در استفاده از سؤالات کوته‌پاسخ آن است که این سؤالات برای ارزیابی کدام سطح از دانش مناسب‌تر هستند. از آنجا که نمره‌دهی این سؤالات ممکن است به آسانی از نوع عینی خارج و به نوع ذهنی متمایل گردد، بهتر است در طراحی این نوع سؤالات از سؤال دانشی به جای تفسیری استفاده شود.

در چه موقعیت هایی از سؤالات کوته‌پاسخ باید استفاده شود؟

همانند سؤالات بسته‌پاسخ، از سؤالات کوته‌پاسخ می‌توان برای ارزیابی سطوح پایین تا متوسط حیطه شناختی از قبیل سنجش دانش، درک، کاربرد و تحلیل مطالب استفاده نمود. به‌علاوه همانند سؤالات بسته‌پاسخ استفاده از این سؤالات در ارزیابی مهارت‌های غیرشناختی از قبیل مهارت‌های ارتباطی، بین فردی و مهارت‌های عملی چندان کمک‌کننده نیست. بنابراین اولین گام در استفاده از سؤالات کوته‌پاسخ تصمیم‌گیری در خصوص این موضوع است که آیا پیامد مورد انتظار یادگیری که قصد ارزیابی آن وجود دارد در حقیقت مناسب با این نوع سؤالات هستند یا خیر.

منابع

1. Crean SJ, Shaikh Z, Addy L. Clinical short-answer questions for postgraduate dentistry. British dental journal 2007; 203(1):56-57
2. Doolittle P. Supply Items.Virginia Polytechnic Institute & State University (2000): 43-58.
3. Estanford P. Efficiency Analysis of Two Written Short-Answer Student Evaluation Formats. J of the American Physical Therapy Association 1988; 68:1546-1549.
4. Fenderson, Bruce A., et al. The virtues of extended matching and uncued tests as alternatives to multiple choice questions. Human pathology 1997;28(5): 526-532.
5. García M, Sempere JM, Marco F, De la Sen L, Segovia Y. The influence of assessment method on student's scores in three course of biology: short answer exam and multiple choice questions. 6th International Technology, Education and Development Conference2012 : 3144-3150
6. Intelligent Assessment Technologies Ltd. E-Assessment of short-answer questions 2009. Available at: www.intelligentassessment.com
7. Leacock C, Chodorow M. C-rater: Automated scoring of short answer questions. Computers and the Humanities 2003; 37: 389–405
8. Mires G, Ben-David MF, Preece P, Smith B. Educational benefits of student self-marking of short-answer questions. Medical Teacher 2001; 23(5): 462-466
9. Rasaian N, Nakhaei S, Sadeghi Gandehari N. Comparision of three exam techniques in medical students: multiple-choice, true-false and short answer questions. J of Hakim 2001; 5(4): 271-80
10. Rademakers TH, Ten Cate J, BAR PR. Progress testing with short answer questions. Medical Teacher 2005; 27(7): 578–582
11. Ramraje SN, Sable PL. Comparison of the effect of post-instruction multiple-choice and short-answer tests on delayed retention learning. Australas Med J 2011; 4(6): 332–339.
12. Royal College of Physicians and Surgeons of Canada. Short-Answer Questions: Guidelines for their Development. Canada 2010
13. Schuwirth LWT, van der Vleuten CP. ABC of learning and teaching in medicine: Written assessment. BMJ 2003; 326:643–5
14. Schuwirth LWT, van der Vleuten CP. Written Assessments. In: Dent J, Harden R, Eds. New York: Elsevier Churchill Livingstone 2005
15. Ten Cate TJ. Comparing Reliabilities of True/False and Short-Answer Questions in Written Problem Solving Tests. Advances in Medical Education 1997:193-96
16. Vleuten CP, van Luijk SJ, Beckers HJ. A written test as an alternative to performance testing. Medical Education 1989; 23: 97–107
17. Zucker S.Fundamentals of standardized testing. Pearson Education 2003.

۱۸. ذوالفقاری ب، ادیبی ن، درخشان فر ث، تن ساز م، کرباسی ا، نیرومند پ. آزمون‌های پیشرفت تحصیلی در علوم پزشکی، انتشارات چاپ نشاط اصفهان، ۱۳۷۹، صفحات ۲۹۰
۱۹. رساییان ن. آموزش دستگاه تنفس و تحلیل ارزشیابی آن در دانشجویان اولین دوره اصلاحات آموزش پزشکی. مجله ایرانی آموزش در علوم پزشکی ۱۳۸۴، دوره ۵، پیوست ۱۴
۲۰. ذوالفقاری ب، ادیبی ن، درخشان‌فر ث، تن‌ساز م، کرباسی ا، نیرومند پ. آزمون‌های پیشرفت تحصیلی در علوم پزشکی ۱۳۷۹. انتشارات چاپ نشاط اصفهان، صفحه ۲۹۰
۲۱. سیف، ع ا. اندازه گیری، سنجش و ارزشیابی آموزشی. نشر دوران، ۱۳۹۴.

۴

آزمون‌های استدلال بالینی

فصل ۱۳

خانواده آزمون‌های استدلال بالینی

تاریخچه آزمون‌های استدلال بالینی

ارزیابی استدلال بالینی، از آنجا که به نوعی ارزیابی مهارت طبابت است، جایگاه ویژه‌ای در آموزش پزشکی دارد. متخصصان ارزیابی همواره در پی طراحی روش‌های ارزیابی استاندارد برای سنجش مهارت‌های استدلال بالینی و چگونگی فعال‌سازی فرایندهای ذهنی دخیل در آن بوده‌اند. در ابتدا این تصور وجود داشت که تنها راه ارزیابی استدلال بالینی و فرایندهای ذهنی مربوط به آن، مشاهده مستقیم عملکرد پزشکان در بالین یا در محیط‌های شبیه‌سازی شده است. اما بعدها با گذشت زمان مشخص شد که ارزیابی استدلال بالینی از طریق مشاهده مستقیم عملکرد در بالین با فرض قابلیت انجام آن، کار بسیار دشواری است. این مشکل، تمایل ارزیابان به استفاده از روش‌های شبیه‌سازی برای سنجش این مهارت‌ها را برانگیخت. همچنین نتایج مطالعات نشان داد که هرچند OSCE ممکن است در ارزیابی استدلال بالینی از طریق مشاهده مستقیم مهارت‌ها و رفتارها در موقعیت‌های شبیه‌سازی‌شده مفید باشد، با این حال برای رسیدن به ویژگی‌های سایکومتریک مطلوب، تعداد زیادی ایستگاه مورد نیاز است که انجام آن را با چالش‌های جدی مواجه می‌سازد (نورمن[1] ۱۹۹۳) و گذشته از آن، محدودیت‌هایی از قبیل پایایی بین ارزیابان وجود دارد (نورمن ۱۹۹۳، چارلین و همکاران[2] ۲۰۰۲). از این رو عمده تلاش‌های صورت‌گرفته برای رفع چنین محدودیت‌هایی، در نهایت منجر به ارائه شکل جدیدی از آزمون‌های کتبی به نام دسته آزمون‌های استدلال بالینی شد.

به طور کلی، ارزیابی استدلال بالینی به عنوان یکی از پیامدهای مورد انتظار در حرفه پزشکی، در دهه ۱۹۶۰ و ۱۹۷۰ آغاز شد. در آن زمان استدلال بالینی به عنوان یک خصوصیت فردی در نظر گرفته می‌شد که یک پزشک با کمک آن قادر به حل مشکلات و مسائل بالینی است و از این رو، بیشتر ارزیابی‌ها در زمینه استدلال بالینی به طرح تعداد معدودی از مشکلات بالینی طویل (گاهی اوقات فقط یک مورد) محدود می‌شد که طبق آن استدلال بالینی را به عنوان یک مهارت عمومی در نظر می‌گرفتند (ون‌درولوتن و نیوبل[3] ۱۹۹۵). به این ترتیب، متداول‌ترین نوع آزمون در آن موقع، آزمون «تدبیر مشکل بیمار»[4] (PMP) بود. زیربنای اصلی معرفی این آزمون، استفاده از مسائل واقعی بیمار جهت ارزیابی فرایند استدلال بود (فلتی و همکاران[5] ۱۹۸۳). با گذشت زمان مشخص شد که اگرچه این شکل از آزمون‌ها از روایی صوری بالایی برخوردار هستند، با این حال دارای اشکالات جدی نیز هستند (ون در ولوتن و نیوبل ۱۹۹۵، چارلین و همکاران ۲۰۰۷).

مهم‌ترین انتقاد وارده به این آزمون‌ها، پایین بودن پایایی آنها بود (نورمن و همکاران ۱۹۸۵). به علاوه، بعدها با مطرح

1. Norman
2. Charlin et al.
3. Van der Vleuten & Newble
4. Patient Management Problem (PMP)
5. Feletti et al.

شدن بحث اختصاصی بودن ارزیابی مهارت استدلال بالینی یا همان ویژگی محتوا این مشکل حادتر شد. به گونه‌ای که نتایج حاصل از مطالعات صورت گرفته نشان داد که همبستگی بین نتایج حاصل از حل دو PMP در دامنه ۰/۱ تا ۰/۳ قرار دارد و معمولاً کمتر از ۱۰ تا ۱۵ درصد واریانس مربوط به عملکرد آزمون‌شوندگان در یک مورد بالینی با سایر موارد بالینی مشترک است (الستین و همکاران[1] ۱۹۷۸). بنابراین از آنجا که هر مورد بالینی بر ارزیابی ویژگی‌های منحصر به فردی تاکید دارد، نمی‌توان انتظار داشت که عملکرد آزمون‌شوندگان در یک موقعیت خاص بالینی قابل تعمیم به سایر موقعیت‌ها باشد. در نتیجه، سنجش عملکرد بالینی فراگیران، نیازمند نمونه‌گیری وسیع‌تری از مهارت‌های مورد ارزیابی است. در نهایت اینکه، مطالعات صورت گرفته نشان‌دهنده همبستگی بسیار بالای بین نمرات این آزمون‌ها با نمرات حاصل از سایر آزمون‌های کتبی بود که این موضوع را مطرح ساخت که PMP اطلاعات بیشتری در خصوص عملکرد داوطلبان فراهم نمی‌سازد (ون‌درولوتن و نیوبل ۱۹۹۵، پیج و بوردیج[2] ۱۹۹۵).

در اوایل دهه ۱۹۸۰ نگرانی‌های پیش رو در استفاده از آزمون PMP باعث گردید تا بسیاری از دانشکده‌های پزشکی و مؤسسات تأیید صلاحیت، این آزمون را از فهرست ابزارهای ارزیابی خود کنار بگذارند (پیج و همکاران ۱۹۹۵). چندی بعد لغت «ویژگی‌های کلیدی» توسط بوردیج و پیج، متعاقب مرور و تحلیلی که روی پژوهش نورمن و همکاران (۱۹۸۵) در مورد ماهیت و ارزیابی مهارت‌های تصمیم‌گیری بالینی داشتند، مطرح گردید. پیج و بوردیج بر اساس مطالعات خود به این نتیجه رسیدند که برای حل موفقیت‌آمیز یک مسأله بالینی تعدادی ویژگی کلیدی و واحد[3] وجود دارد که همواره باید آن‌ها را در نظر گرفت (بوردیج و پیج ۱۹۸۷).

این دیدگاه در انتها منجر به ارائه ابزار جدیدی در زمینه ارزیابی مهارت‌های تصمیم‌گیری بالینی به نام آزمون ویژگی‌های کلیدی شد (پیج و بوردیج ۱۹۹۵). از دیدگاه آنان، ویژگی‌های کلیدی به عنوان یک مرحله اساسی در حل مسأله بالینی در نظر گرفته می‌شوند و از طریق آزمون «ویژگی‌های کلیدی»[4] می‌توان بین متخصصان[5] و افراد تازه‌کار[6] تمایز قائل شد. در نتیجه، استفاده از «ویژگی‌های کلیدی» در فرایند ارزیابی امکان سنجش بخشی از توانمندی‌های مورد انتظار بالینی را فراهم می‌سازد.

به طور کلی، مفهوم ویژگی کلیدی مطرح‌کننده دو تغییر مهم در ارزیابی سنتی توانمندی‌های بالینی است. اول این که، مفهوم ویژگی کلیدی، تمرکز فرایند سنجش را از روش ارزیابی به موضوع مورد ارزیابی سوق می‌دهد. دوم این که مفهوم ویژگی کلیدی، ارزیابی عناصر ضروری مرتبط با هر مشکل را بر ارزیابی تمام اجزای حل یک مسأله مقدم می‌شمارد. این تغییرات مشخص می‌سازند که عناصر اساسی در حل یک مسأله برای هر مورد بالینی منحصر به فرد و ویژه خواهد بود. برای تعدادی از مسائل بالینی، ویژگی‌های کلیدی ممکن است مربوط به فرایند جمع‌آوری یا تفسیر اطلاعات باشد و برای برخی دیگر منوط به تشخیص یا انتخاب یک تدبیر بالینی مناسب باشد.

درباره تاریخچه استفاده از سؤالات «ویژگی‌های کلیدی» در مؤسسات آموزشی، می‌توان به سال ۱۹۹۲ اشاره کرد که در آزمون انجمن پزشکی کانادا[7] این شکل از سؤالات به عنوان جایگزینی برای PMP و سؤالات چندگزینه‌ای استفاده شدند (بوردیج و پیج ۱۹۸۷، پیج و همکاران ۱۹۹۵). طی سال‌های بعد، این شکل از سؤالات به سرعت در سایر سطوح و دوره‌های آموزشی در دانشکده‌های پزشکی مختلف مورد استفاده قرار گرفتند و در حال حاضر به عنوان آزمون‌های ورودی به مقاطع بالاتر فارغ‌التحصیلان رشته پزشکی نیز مورد استفاده قرار می‌گیرند.

1. Elstein et al.
2. Bordage & Page
3. Unit
4. Key features
5. Expert
6. Novice
7. Medical Council of Canada (MCC)

در ادامه، تجربه اجرای این دسته از آزمون‌ها مشخص کرد که هرچند این شکل از آزمون‌ها امکان سنجش مهارت‌های استدلال بالینی را فراهم می‌سازد، با این حال زمان‌بر بودن فرایند طراحی آن‌ها و نیاز به طرح تعداد زیادی سؤال برای رسیدن به سطح قابل قبول پایایی، از محدودیت‌های این ابزارها به شمار می‌رود (چارلین و همکاران ۲۰۰۳).

مجموع این محدودیت‌ها منجر به ابداع شکل جدیدی از دسته آزمون‌های کتبی به نام آزمون «همخوانی با شرحنامه»[1] شد. این آزمون در اواخر دهه ۱۹۹۰ توسط چارلین و همکاران به عنوان یک ابزار جهت سنجش مهارت‌های تصمیم‌گیری داوطلبان طراحی و معرفی گردید. از دیدگاه چارلین، پزشکان باید قادر باشند در موقعیت‌های پیچیده و مبهم با جمع‌آوری اطلاعات از طریق انجام مجموعه‌ای از قضاوت‌های کیفی، به تأیید یا رد فرضیه‌های تشخیصی خود بپردازند (فلوویچ و باروز[2] ۱۹۸۴، چارلین و همکاران ۲۰۰۰). لذا آزمون «همخوانی با شرحنامه» توانایی ارزیابی مهارت‌های تصمیم‌گیری بالینی داوطلبان را از طریق سنجش فرضیه‌های تشخیصی پیش رو طبق الگوی شرحنامه فراهم می‌سازد. نتایج حاصل از ارزیابی استدلال بالینی در مطالعات گوناگون نیز موید این مطلب است که در شرایط مشابه، متخصصان اطلاعات مشابهی را جمع‌آوری نمی‌کنند و الگوهای ذهنی یکسانی را دنبال نمی‌کنند (گرنت[3] ۱۹۸۸). نتایج حاصل همچنین نشان‌دهنده تفاوت اساسی در عملکرد متخصصان در شرایط واقعی و شبیه‌سازی شده است (الستین ۱۹۷۸، باروز ۱۹۷۸).

در ادامه با درک بیشتر فرایند استدلال، اَشکال جدیدی از آزمون‌های استدلال بالینی از قبیل پازل ادغام‌یافته[4]، آزمون سناریونویسی و ... توسط پژوهشگران مختلف معرفی شد که به شرح آن‌ها خواهیم پرداخت. اما قبل از معرفی تک‌تک ابزارها از آنجا که اصطلاحات و عبارات خاصی برای توضیح فرایند استدلال بالینی به کار می‌روند، ابتدا به بیان این مفاهیم می‌پردازیم.

مفاهیم پایه در استدلال بالینی

استدلال رو به جلو و رو به عقب

در استدلال رو به جلو[5] در ابتدا به دنبال جستجوی علایم بیماری می‌گردیم و سپس بر اساس مجموعه علایم و نشانه‌های جمع‌آوری شده، تشخیص‌های احتمالی بیماری را مطرح می‌کنیم. در مقابل در استدلال رو به عقب[6]، نخست تشخیص یک بیماری را به عنوان فرضیه در نظر می‌گیریم و سپس از طریق اخذ شرح حال، انجام معاینه فیزیکی و درخواست تست‌های آزمایشگاهی به دنبال علایم و نشانه‌های آن در بیمار می‌گردیم.

از استدلال رو به جلو در زمان ساخت فرضیه و از استدلال رو به عقب هنگام ارزیابی فرضیه استفاده می‌شود. در متخصصان، ساخت فرضیه‌های تشخیصی زمانی صورت می‌گیرد که اطلاعات جمع‌آوری شده به اندازه کافی برای ساخت فرضیه در دسترس باشد. برعکس آن، یکی از ویژگی‌های دانشجویان و نوآموزان، شتابزدگی در رسیدن به تشخیص‌های بیماری است که در اکثر مواقع نادرست هستند. این بدان معناست که اکثر دانشجویان و پزشکان تازه‌کار بدون جمع‌آوری اطلاعات کافی وارد مرحله ساخت فرضیه‌های تشخیصی می‌شوند.

1. Script Concordance Test
2. Feltovich & Barrows
3. Grant
4. Integrated puzzle
5. Forward reasoning
6. Backward reasoning

استدلال شرحنامه

در زندگی روزمره، ما کارهایی انجام می‌دهیم که پس از مدتی به صورت عادت درمی‌آیند و به طور خودکار و بدون اینکه نیاز به تأمل و تفکر داشته باشیم آن‌ها را انجام می‌دهیم. برای این رفتارها، ساختارهایی شناختی در ذهن ما شکل می‌گیرند که شرحنامه[1] خوانده می‌شوند. این شرحنامه‌ها در پزشکی در قالبِ بیماری‌های مختلف سازمان‌دهی می‌شوند. طبق نظریه شرحنامه، در اثر مواجهه مکرر پزشکان با موارد بالینی تکراری، مجموعه‌ای سازمان‌یافته از اطلاعات پزشکی در ذهن آن‌ها نقش می‌بندد که باعث می‌شود هنگام برخورد با بیماران بعدی از آنها در فرایند تصمیم‌گیری و حل مشکل بالینی استفاده کنند. بدین ترتیب پزشکان با تجربه، هنگام مواجه با موارد جدید بیماری که مشابه با بیماری‌هایی است که در گذشته با آن‌ها برخورد داشته‌اند، شبکه‌ای ساختارمند از اطلاعات مرتبط را به خاطر می‌آورند و از این طریق با صرف زمان کمتر به راه‌حل‌های تشخیصی و درمانی می‌رسند (چارلین و همکاران ۲۰۰۰). به عنوان مثال، یک متخصص گوش، حلق و بینی در مواجهه با بیماران سرپایی که از سرگیجه شکایت دارند، بلافاصله و به صورت خودکار به عوامل مسبب این علامت فکر می‌کند. در همین حین، اگر بیمار دیگری با شکایت توده گلو به مطب وی مراجعه کند، شرحنامه سرگیجه بلافاصله از ذهن او پاک می‌شود و شبکه‌های اطلاعاتی مرتبط با توده گلو در ذهن او فرا خوانده می‌شوند و در واقع این سؤالات در ذهن او مطرح می‌گردد که در حال حاضر چه گزینه‌های تشخیصی یا درمانی برای این بیمار کارساز است. به صورت کلی، شرحنامه بیماری‌ها به پزشکان در تعیین تشخیص‌های افتراقی، استراتژی‌های مداخله‌ای و گزینه‌های درمانی کمک‌کننده خواهد بود.

همان‌طور که گفته شد شرحنامه بیماری‌ها در یک پزشک از سال اول پزشکی شروع به شکل‌گیری می‌کند و در طی سالیان متمادی فعالیت و تجربه، تکمیل و اصلاح می‌شود (اشمیت و همکاران[2] ۱۹۹۰). به همین دلیل پزشکان باتجربه با داشتن تعداد زیادی از این شرحنامه‌ها در ذهن خود قادر هستند بیماری‌ها را در کمترین زمان ممکن و با بالاترین دقت تشخیص و درمان نمایند. در مقابل، شرحنامه شکل‌گرفته در ذهن دانشجویان و نوآموزان در مراحل ابتدایی قرار دارد و غنی از مطالب مربوط به پاتوفیزیولوژی بیماری‌ها است که باید به تدریج در نتیجه کسب تجربه، دریافت اطلاعات زمینه‌ای بیماری‌ها و تشخیص و درمان تقویت و تکمیل گردد.

استدلال فرضیه‌ای- قیاسی

در این رویکرد پزشک تلاش می‌کند تا از طریق کشف رابطه علی- معلولی میان علایم و نشانه‌های بیماری به تشخیص احتمالی برسد. در واقع استدلال فرضیه‌ای-قیاسی[3] روشی از استدلال است که بر اساس روش سعی و خطا است. این رویکرد بیشتر توسط دانشجویان و پزشکان تازه‌کار برای حل مسائل بالینی با استفاده از پاتوفیزیولوژی بیماری‌ها استفاده می‌شود. علت این امر آن است که هنوز معلومات آن‌ها از ساختار مناسبی برخوردار نیست و به علاوه تجربه بالینی کافی از مواجه با بیماران مختلف نیز ندارند. با این حال ممکن است پزشکان مجرب و متخصص نیز از این روش در مواجه با مسائل بالینی مبهم و پیچیده که به راحتی قابل حل نیستند یا خارج از حوزه تخصصی آن‌ها قرار دارد، استفاده کنند.

استدلال تحلیلی و غیرتحلیلی

به صورت کلی می‌توان گفت که فرایند استدلالی که در موارد بالا توضیح داده شد، همواره آمیزه‌ای است از دو گونه استدلال تحلیلی[4] و غیرتحلیلی[5] که در قالب نظریه پردازش دوگانه[6] ارائه می‌شود. طبق این نظریه، استدلال تحلیلی

1. Pattern recognition
2. Schmidt et al.
3. Hypothetico-deductive reasoning
4. Analytical reasoning
5. Non- analytical reasoning
6. Dual-processing

استدلالی خودآگاه، کند و زمان‌بر و مرحله‌مرحله است و انجام آن انرژی زیادی می‌برد. در مقابل استدلال غیرتحلیلی، ناخودآگاه و سریع است، در لحظه اتفاق می‌افتد و نیاز به صرف انرژی زیادی ندارد. در واژگان استدلال بالینی، استدلال غیرتحلیلی معادل بازشناسی الگو[1] و استدلال تحلیلی مترادف استدلال فرضیه‌ای قیاسی[2] است. تذکر این نکته حائز اهمیت است که بر اساس این نظریه، استدلال بالینی یک طیف است که یک سوی آن استدلال تحلیلی و سوی دیگر آن استدلال غیرتحلیلی است. به طور کلی هر استدلالی به یکی از این دو تمایل بیشتری دارد، اما درواقع آمیزه‌ای از هر دو استدلال است (کاسترز[3] ۲۰۱۳). آنچه مسلم است برای یک استدلال و حل موفق مشکلات بیمار، نیاز به هر دوگونه استدلال داریم و ناتوانی یا بی‌کفایتی در هرکدام از آنها نهایتاً به حل ناموفق مسأله منتهی می‌شود.

برای روشن‌تر شدن بحث به مثال زیر توجه کنید. بیمار مرد ۴۶ ساله دیابتی با سابقه مصرف سیگار است که با شکایت تنگی‌نفس و درد مبهم قفسه سینه که از دو ساعت پیش شروع شده است، طی کمتر از بیست دقیقه به اورژانس آورده شده است. در ادامه نحوه برخورد یک پزشک متخصص و یک دانشجوی تازه‌کار با این بیمار مورد بحث قرار می‌گیرد:

نحوه استدلال پزشک متخصص به صورت استدلال غیرتحلیلی است. وی در مواجهه با چنین بیماری بلافاصله به آنژین قلبی فکر می‌کند. به عبارت دیگر شرحنامه آنژین قلبی سریعاً به ذهنش خطور می‌کند و بر اساس شرحنامه‌ای که در ذهنش فعال شده است، شروع به جمع‌آوری اطلاعات بیشتر در مورد بیمار می‌کند. به عنوان مثال از بیمار می‌پرسد که آیا درد با تنفس تغییر می‌کند یا خیر. اگر درد با تنفس تغییر کند، به ضرر شرحنامهٔ آنژین قلبی است. اگر با گرفتن شرح‌حال بیشتر مشخص شود بیمار سابقه احساس سنگینی قفسه سینه از یک ماه پیش دارد که با پیاده‌روی تشدید و با قرص زیرزبانی آرام می‌شده است، به نفع شرحنامهٔ آنژین قلبی است. به علاوه، به نظر پزشک متخصص، تسکین درد بیمار با نوشیدن مایعات گرم ردکنندهٔ تشخیص آنژین قلبی است و حملهٔ درد پس از مصرف شام سنگین توام با نفخ شکم به نفع آنژین قلبی است. همین طور که از فرآیند تفکر این پزشک متخصص مشهود است، در برخورد اولیه با بیمار، به سرعت یک شرحنامه در ذهن وی فعال می‌شود و در ادامهٔ مسیر استدلال هر یافتهٔ جدید را در پرتو شرحنامهٔ فعال‌شده ارزیابی می‌کند. در واقع پزشک با جمع‌آوری اطلاعات بیشتر مشخص می‌کند که هر یافته، فرضیه تشخیصی وی را تائید یا رد می‌کند. اینکه چه یافته‌هایی را در بیمار باید جستجو کرد، چیزی است که شرحنامه برای پزشک مشخص می‌کند. به این ترتیب پزشک فقط ویژگی‌های کلیدی مورد نظرش را در بیمار جستجو می‌کند.

اما در مقابل، فرآیند استدلال در دانشجوی پزشکی به صورت استدلال تحلیلی است. او در مواجهه با بیمار فوق به تشخیص‌های افتراقی درد قفسه سینه فکر می‌کند: درد قلبی، ریوی، گوارشی. او با استناد به اینکه بیمار هم اکنون درد ندارد و نوار قلب بیمار هم طبیعی است، مشکل قلبی بیمار را نادیده می‌گیرد و با توجه به اینکه تنگی‌نفس وی با فعالیت بیشتر می‌شود و سیگار می‌کشد، نتیجه می‌گیرد که بیمار مشکل ریوی دارد اما با شنیدن اینکه درد بیمار با مصرف مایعات گرم بهتر شده است و در حمله اخیر هم درد به دنبال مصرف شام سنگین و توام با نفخ شکم بوده است، به بیماری گوارشی شک می‌کند. در استدلال یک دانشجوی نوآموز به خوبی می‌توان دید که او از بازشناسی الگوی آنژین قلبی از ابتدا ناتوان است و برای تشخیص دادن مشکل بیمار تنها بر یک علامت بیمار متمرکز می‌شود. دوم آنکه وی مشکل قلبی بیمار را با توجه به یافته‌هایی رد می‌کند که ردکنندهٔ مشکل قلبی نیستند و این به خوبی فقر و بی‌کفایتی شرحنامهٔ مشکل قلبی در ذهن وی را نشان می‌دهد. از سوی دیگر هنگامی که دانشجو به غلط تشخیص مشکل ریوی را مطرح می‌کند، بدون تلاش برای پیگیری همین تشخیص در ادامه به سرعت آن را رها می‌کند و به تشخیص مشکل گوارشی متوسل می‌شود. به علاوه، معمولاً دانشجویان هنگام برخورد با یک بیمار به دلیل ناقص بودن الگوهای شناختی خود، شروع به پرسش سؤالات بسیار می‌کنند که اغلب بدون هدف خاص صورت می‌گیرد. از این رو دانشجویان اغلب تمایل به گرفتن شرح حال

1. Pattern recognition
2. Hypothetico-deductive
3. Custers

کامل از بیماران دارند و حتی از سوی استادان خود برای انجام این کار تشویق می‌شوند. در حالی که متخصصان به دلیل مشخص بودن هدف از پرسش هر سؤال، شرح حال متمرکز از بیمار می‌گیرند.

با وجود مواردی که گفته شد، استدلال تحلیلی فقط مختص نوآموزان و دانشجویان نیست و متخصصان در دو حالت به سمت این استدلال تمایل پیدا می‌کنند. اول در مواجهه با آن دسته از تظاهرات و تابلوهای بالینی که در مواجههٔ اولیه با آنها یک الگوی مشخص در ذهنشان فعال نمی‌شود. این حالت معمولاً زمانی اتفاق می‌افتد که پزشکان با تظاهرات نادر یک بیماری شایع یا با یک بیماری ناشایع و نادر مواجه می‌شوند. دوم آنکه متخصصان با بیماری برخورد کنند که خارج از حیطه تخصصی‌شان است مثل آنکه یک متخصص زنان و زایمان با یک بیمار قلبی مواجه شود یا یک متخصص قلب با یک بیمار روانپزشکی. در این شرایط استدلال به سمت استدلال تحلیلی سوق پیدا می‌کند. به این ترتیب با آنکه در مسیر خبره شدن و کسب تجربه در طبابت انتظار این است که استدلال پزشکان به سمت استدلال غیرتحلیلی تمایل پیدا کند، بسیار اتفاق می‌افتد که پزشکان با شرایط دشواری مواجه می‌شوند که مجبور هستند از استدلال تحلیلی بهره گیرند و این انعطاف‌پذیری در تغییر وضعیت دادن از استدلال غیرتحلیلی به تحلیلی یکی از ویژگی‌های خبرگان است.

به این ترتیب اگر استدلال بالینی آمیزه‌ای از استدلال تحلیلی و غیرتحلیلی است، ارزیابی آن هم باید هر دو گونه استدلال را در بربگیرد. مطالعات مختلف نشان داده‌اند که هر کدام از این دو گونه استدلال را می‌توان با دستورالعمل‌ها یا مداخلاتی فعال کرد. مثلاً پرسیدن اینکه پس از خواندن سناریوی این بیمار اولین تشخیصی را که به ذهن‌تان خطور می‌کند بنویسید، آشکارا استدلال غیرتحلیلی را می‌طلبد. در حالی که، درخواست از داوطلب برای نوشتن تعدادی تشخیص افتراقی برای سناریوی بیمار، استدلال تحلیلی را فعال می‌کند. پرسیدن دلیل انتخاب یک تشخیص، توضیح خواستن در مورد شواهد و دلایل تائیدکننده تشخیص یا تبیین پاتوفیزیولوژی بیماری همه از مصادیق استدلال تحلیلی هستند. در مقابل، نوشتن یک سناریوی بالینی برای یک تشخیص مشخص از نمونه‌های استدلال غیرتحلیلی است.

ساختار کلی آزمون‌های استدلال بالینی

برخی آزمون‌های استدلال بالینی را در دسته‌ای از آزمون‌ها قرار می‌دهند که ارزیابی جایگزین[1] نامیده می‌شود. ارزیابی‌های جایگزین با نزدیک کردن شرایط ارزیابی به شرایط واقعی، دانش و مهارت استدلال داوطلبان را ارزیابی می‌کنند (منجمی ۱۳۸۹). به علاوه، از آنجا که مسائل بالینی در این آزمون‌ها بر اساس شرایط واقعی طرح می‌شوند، بنابراین فقط یک پاسخ درست برایشان وجود نخواهند داشت و این دقیقاً برعکس آزمون‌های عینی است که برای هر سؤال، فقط و فقط یک پاسخ صحیح در نظر گرفته می‌شود. در این صورت تصحیح این گونه آزمون‌ها هم با آزمون‌های عینی تفاوت‌هایی خواهد داشت، چرا که باید به دنبال تمام پاسخ‌های محتمل باشیم. از این رو آزمون‌های جایگزین، هم از بُعد طراحی و هم از نظر تهیه کلید و تصحیح، کاملاً با سایر آزمون‌های کتبی مرسوم تفاوت دارند. برای استخراج گزینه‌های صحیح در این آزمون‌ها، کار به متخصصان آن رشته واگذار می‌شود که گروه خبرگان[2] نامیده می‌شود. در این روش گروه خبرگان بر اساس روشی استاندارد، کلید آزمون را تهیه می‌کنند که در فصل‌های بعدی به آن اشاره خواهد شد. یکی دیگر از ویژگی‌های این آزمون‌ها، تاکید بر انجام یک اقدام یا اتخاذ یک تصمیم است. به عنوان مثال، اگر سناریوی بیماری با درد ناحیه اپیگاستر مطرح شود و درباره‌ی چگونگی تولید اسید معده و ارتباط آن با بروز درد سؤالاتی پرسیده شود، معلومات دانشی آزمون‌شونده ارزیابی می‌شود. اما اگر در سناریوی بالا از آزمون شونده خواسته شود که بگوید برای تشخیص بیماری، جمع‌آوری چه اطلاعات دیگری کمک‌کننده است، مهارت تصمیم‌گیری و استدلال او مورد ارزیابی قرار گرفته است.

1. Alternative assessment
2. Expert panel

به طور کلی ساختار آزمون‌های استدلال بالینی شامل سه جزء اصلی است: سناریو، سؤالات و دستورالعمل (منجمی ۱۳۸۹). در تمام آزمون‌های استدلال بالینی این سه جزء وجود دارد اما ترتیب و شکل آن‌ها در آزمون‌های مختلف متفاوت است. در ادامه به شرح مختصری از این اجزاء خواهیم پرداخت و در فصول بعد نحوه طراحی هر یک از آزمون‌های استدلال بالینی را به شکل اختصاصی شرح خواهیم داد.

سناریوی بیمار

مهم‌ترین جزء یک آزمون استدلال بالینی، سناریو است. همان‌طور که پیشتر مطرح شد، آزمون‌های استدلال بالینی درصدد بازنمایی شرایط واقعی هستند بنابراین سناریوی طراحی‌شده باید تا حد امکان به شرایط تجربه شده در بالین نزدیک باشند. در طراحی سناریوی بیمار باید به موارد زیر توجه کرد:

- **اندازه سناریو:** اندازه سناریو بسته به هدف ارزیابی و نوع آزمون متفاوت است. به عنوان مثال در آزمون «ویژگی‌های کلیدی» که هدف از آن سنجش توانایی جمع‌آوری اطلاعات فراگیران است، سناریو باید کوتاه و مختصر باشد تا شرایط مناسب برای جمع‌آوری اطلاعات فراهم شود. اگر اطلاعات به طور کامل ارائه شوند، پرسیدن و کسب اطلاعات بیشتر موضوعیتی نخواهد داشت. بر عکس، در آزمون «همخوانی با شرحنامه» که هدف آن ارزیابی توانایی دانشجو در ساختن فرضیه‌های تشخیصی است، سناریو باید نسبت به آزمون «ویژگی‌های کلیدی» کامل‌تر باشد تا بتوان بر اساس آن تشخیص‌های احتمالی را مطرح نمود.
- **ابهام در سناریو:** این نکته بسیار حائز اهمیت است چرا که اگر مشکل و مسأله بالینی بسیار واضح و مشخص باشند، دیگر مجالی برای استدلال و حل مسأله باقی نمی‌ماند. به عنوان نمونه به مثال زیر توجه کنید که با ارائه یک سناریو از دانشجو خواسته شده است تشخیص احتمالی را ذکر کند.

سناریوی ضعیف

آقای ۶۵ ساله با درد قفسه سینه و تنگی‌نفس مراجعه کرده است. سابقه چند بار بستری در CCU را دارد و از چربی و فشارخون نیز رنج می‌برد. در نوار قلبی، بالا رفتن قطعه ST در لیدهای تحتانی مشاهده می‌شود و افزایش قابل ملاحظه آنزیم‌های قلبی هم گزارش شده است.

مسلماً بعد از خواندن این شرح‌حال، تشخیص انفارکتوس میوکارد مطرح می‌شود و تشخیص دیگری را به سختی می‌توان مطرح کرد. حالا فرض کنید که اطلاعات مربوط به نوار قلب و آنزیم‌های قلبی را حذف کنید. بدین ترتیب، تشخیص‌های دیگری از قبیل آنژین ناپایدار، آمبولی ریه و دایسکشن آئورت را هم می‌توان به این بیمار نسبت داد. البته باید توجه داشت که میزان ابهام مورد نیاز در آزمون‌های مختلف قدری متفاوت است. نکته‌ای که باید در ذهن داشت، این است که ابهام با پیچیدگی و دشواری سؤال اشتباه گرفته نشود. ابهام به معنای دشواری نیست بلکه این مفهوم را می‌رساند که در شرایط واقعی همه اطلاعات بیمار از ابتدا در اختیار پزشک قرار ندارد و این موضوع تشخیص در وهله اول را غیرممکن می‌سازد.

- **انواع سناریو:** سناریوها انواع و اقسام متفاوتی دارند. آن‌ها می‌توانند در مورد شکایات مبهم و تعریف‌نشده بیمار، تابلوی بالینی مشخص، درگیری همزمان چند سیستم یا پیشگیری از بیماری باشند. هر کدام از این سناریوها، مناسب یکی از انواع آزمون‌های استدلال بالینی هستند. به عنوان مثال، در آزمون «ویژگی‌های کلیدی» بهتر است در مورد شکایات اولیه بیمار سناریو طراحی شود چرا که هدف از این آزمون، سنجش توانایی جمع‌آوری اطلاعات از بیماران است. در حالی که در آزمون پازل ادغام‌یافته، سناریو باید در مورد یک تابلوی بالینی مشخص باشد.
- **ترکیب سناریو:** از موضوعات مهم این است که چگونه باید از حیطه‌ها و حوزه‌های مختلف سؤال انتخاب کرد.

همان‌طور که قبلاً گفته شد، آزمون‌های استدلال بالینی باید تا حد امکان به شرایط واقعی نزدیک باشند. بنابراین گستره سناریوها باید به گونه‌ای باشد که شرایط واقعی را شبیه‌سازی کند. قاعدتاً یکی از رویکردهای مؤثر برای انجام این مهم، استفاده از آمار مراجعان است که مشخص می‌کند کدام یک از بیماری‌ها شایع‌تر است. بدین ترتیب سناریوها بر اساس توزیع بیماری‌ها طراحی می‌شوند. به علاوه اگر هدف، ارزیابی خطاهای پزشکی و مسائل تهدیدکننده زندگی است باید مشخص شود چه بیماری‌هایی وجود دارند که عدم تشخیص و تدبیر بالینی به موقع در خصوص آن‌ها جان بیمار را به مخاطره می‌اندازد. به هر صورت، بسته به هدفی که در ذهن داریم و حوزه‌ای از طب که انتخاب کرده‌ایم، رویکردهای متفاوتی را می‌توان اتخاذ کرد اما نکته مهم آن است که هنگام طراحی، یک نقشه کلی از نحوه بودجه‌بندی سناریوها در ذهن داشته باشیم.

سؤالات/ گزینه‌ها

در آزمون‌های استدلال بالینی، بسته به نوع آزمون ممکن است سؤالات کوته‌پاسخ یا چندگزینه‌ای طرح شود. اصول حاکم در طراحی سؤالات تفاوت عمده‌ای با سایر سؤالات کتبی ندارد (جهت کسب اطلاعات بیشتر در خصوص می‌توانید به بخش دو و سه کتاب مراجعه نمایید)

دستورالعمل

برخلاف قسمت ابتدایی یعنی سناریو که عمدتاً باید مبهم نوشته شود، دستورالعمل سؤال باید کاملاً واضح و گویا باشد و بر تصمیم‌گیری تاکید کند. در غیر این صورت، روند استدلال بالینی در دانشجو فعال نخواهد شد و تبعاً مورد ارزیابی هم قرار نخواهد گرفت. تاکید بر تصمیم‌گیری به چند شکل میسر می‌شود که مهمترین آن‌ها محدود کردن تعداد پاسخ است. محدودیت تعداد تصمیم درست در شرایط واقعی نیز وجود دارد. اگر تعداد تصمیم‌های مجاز نامحدود باشد، عملاً نمی‌توانیم بین متخصصان و دانشجویان تمایزی قائل شویم. زیرا معمولاً دانشجویان تعداد زیادی پاسخ ارائه می‌کنند که تعدادی از آن‌ها درست نیست. بنابراین باید دانشجویان را متوجه کنیم که در ارائه تصمیم‌های پیشنهادی یا انتخابی محدودیت وجود دارد. برای این امر، بهتر است تعداد حداکثر پاسخ مجاز را مشخص نماییم. به عنوان مثال، بگوییم که حداکثر پنج پاسخ را می‌توانند ارائه یا انتخاب کنند. اما حداقل را مشخص نکنیم تا پنج مورد حالت اجباری نداشته باشد. به علاوه، از آنجا که این باور اشتباه در بین دانشجویان وجود دارد که انتخاب یا ارائه پاسخ‌های بیشتر به کسب نمره بالاتر می‌انجامد، در متن دستورالعمل باید به این نکته تاکید شود که ارائه یا انتخاب پاسخ‌های بیشتر به معنای به دست آوردن نمره بیشتر نیست.

طبقه‌بندی آزمون‌های استدلال بالینی

طبق گفته کریتر و برگز[1]، استدلال بالینی شبیه «جعبه سیاه»[2] است و ارزیابی محتوای آن مستلزم سنجش خروجی (رویکردهای سنتی متداول از قبیل ارزیابی پیامدهای حاصل از فرایند حل مسأله) یا ورودی‌های آن (رویکردهای کمتر متداول از قبیل ارزیابی نحوه سازماندهی دانش و توانایی ادغام اطلاعات جدید) است. آنچه کریتر و برگز به عنوان استدلال بالینی توصیف می‌کنند شامل یادگیری دانش شناختی، سازمان‌دهی دانش، استدلال بالینی و در نهایت حل مسائل بالینی است.

به طور کلی چهار مؤلفه اصلی استدلال بالینی شامل جمع‌آوری اطلاعات، ساختن فرضیه‌ها، ارزیابی فرضیه‌ها و حل مسأله است. روش‌های بسیار متنوع ارزیابی استدلال بالینی هر کدام یکی از چهار مؤلفه اصلی استدلال بالینی یا آمیزه‌ای

1. Kreiter & Bergus
2. Black box

از چند مؤلفه را می‌سنجند. به عنوان مثال، در آزمون «ویژگی‌های کلیدی» ابتدا شرح‌حال بیمار مطرح می‌شود و سپس از فراگیران پرسیده می‌شود برای رسیدن به تشخیص این بیمار نیاز به چه داده‌های دیگری دارند. تمرکز این آزمون بر جنبه‌های مختلف جمع‌آوری اطلاعات است. برعکس در آزمون استدلال بالینی که بیشتر روی ساختن فرضیه‌های تشخیصی متمرکز است، از داوطلبان خواسته می‌شود که پس از خواندن یک شرح‌حال مختصر بیمار، محتمل‌ترین تشخیص را انتخاب کنند.

در ادامه این بحث، به شرح خصوصیات هر طبقه پرداخته می‌شود.

سنجش جمع‌آوری اطلاعات

هدف این مرحله، ارزیابی توانایی گرفتن شرح‌حال و به دست آوردن داده‌های پایا[1] و روا[2] از خلال نتایج حاصل از معاینات فیزیکی و اطلاعات پاراکلینیک است. نکته مهم در جمع‌آوری اطلاعات، سرعت، دقت و صحت داده‌های به دست آمده است. مسأله مهم دیگر، توانایی استفاده از منابع مختلف شرح‌حال و سنجیدن اعتبار اطلاعات به دست آمده بر اساس مقایسه منابع مختلف است. همچنین در این مرحله داوطلب باید بتواند بین داده‌های مرتبط و بی‌ربط[3] تمایز قائل شود.

ساختن فرضیه‌های تشخیصی

هدف این مرحله، بررسی توانایی پزشک در ساخت فرضیه‌ها بر پایه اطلاعات موجود است. همانطور که پیشتر نیز اشاره شد در ساخت فرضیه‌ها، استدلال همگرا یا رو به جلو نقشی حیاتی دارد بدین ترتیب که از جمع‌بندی یافته‌های موجود فرضیه‌های تشخیصی به دست می‌آیند. نکته مهم در این فرایند، ارائه فرضیه بر پایه یافته‌های کافی و مرتبط است. به عبارت دیگر یافته‌های حاصل از نظر تعداد و کیفیت باید به حدی برسند که منجر به شکل‌گیری فرضیه‌های تشخیصی شوند. البته توجه به این نکته حائز اهمیت است که رویکرد پزشک مجرب و یک دانشجوی نوآموز در رسیدن به فرضیه‌های تشخیصی قدری متفاوت است. پزشکان باتجربه عمدتاً با مشاهده وضعیت بیمار و پرسیدن سؤالات کم به تشخیص می‌رسند، در حالی‌که یک دانشجوی نوآموز ممکن است حتی با صرف زمان طولانی نتواند به تشخیصی برسد.

ارزیابی فرضیه‌های تشخیصی

در ارزیابی فرضیه‌های تشخیصی، سه روش کلی وجود دارد. متداول‌ترین روش اثبات[4] است. در این روش سعی می‌شود، داده‌هایی به نفع فرضیه‌های تشخیصی در بیمار پیدا شوند. این روند آنقدر ادامه می‌یابد تا فرضیه تشخیصی نهایتاً تأیید شود. این روش معمولاً وقتی کاربرد دارد که شکایت بیمار ساده و غیرپیچیده باشد یا دانش ما در خصوص آن بسیار ساختارمند باشد. روش دیگر حذف[5] است که در آن سعی می‌کنیم بر اساس داده‌های بیمار فرضیه‌های تشخیصی را حذف کنیم. این روش زمانی کارساز است که فهرست بلندبالایی از تشخیص‌های افتراقی محتمل وجود داشته باشد. این حالت معمولاً زمانی رخ می‌دهد که یا تابلوی بالینی پیچیده است یا دانش ما در خصوص آن بسیار محدود است. در نهایت در روش تمیز[6]، تشخیص‌های افتراقی بر اساس معیارهایی از هم تمیز داده می‌شوند و بر این اساس، روند ارزیابی فرضیه‌ها به سمتی خاص سوق داده می‌شود. این روش در مورد شکایاتی کاربرد دارد که پاتوفیزیولوژی شناخته شده‌ای دارند، منتها طیف وسیعی از بیماری‌ها را در بر می‌گیرند. در این روش، بر اساس تابلوی بیماری فرضیه یا فرضیه‌های تشخیصی مطرح می‌شوند. در ادامه، داده‌های جدیدی از بیمار مطرح می‌گردند و از داوطلبان درخواست می‌شود که مشخص کنند که این داده‌ها فرضیه‌ها را رد می‌کنند یا تأیید.

---

1. Reliable
2. Valid
3. Relevant and inrelevant
4. Confirmation method
5. Deletion method
6. Discrimination method

حل مسأله و تصمیم‌گیری

در این مرحله از داوطلبان درخواست می‌شود که از اطلاعات حاصل از مراحل قبلی یک جمع‌بندی ارائه نمایند و بر اساس آن برنامه تشخیصی و درمانی خود را ذکر کنند.

ذکر این نکته کلی ضروری است که برخی معتقد هستند تصمیم‌گیری و استدلال بالینی در خلال یک فرایند رخ می‌دهد که سنجش دقیق آن مستلزم اندازه‌گیری تمام اتفاقاتی است که در جریان آن رخ می‌دهد. هر چند در برخی از مستندات، ارزیابی «فرایند» حل مسأله و تصمیم‌گیری به دلیل مشکلات موجود مورد نقد قرار گرفته است و در مقابل بر ارزیابی «پیامدها» تاکید شده است (شوورث[1] و ون‌درولوتن ۲۰۱۱). در جدول ۱-۱۳ نمونه آزمون‌هایی که در هر دسته قرار می‌گیرند، آورده شده است.

جدول ۱-۱۳: طبقه‌بندی آزمون‌های استدلال بالینی

ارزیابی جمع‌آوری اطلاعات	ساختن فرضیه‌های تشخیصی	ارزیابی فرضیه‌های تشخیصی	حل مسأله و تصمیم‌گیری
آزمون ویژگی‌های کلیدی آزمون جمع‌آوری اطلاعات	آزمون استدلال بالینی آزمون سناریونویسی آزمون استدلال تحلیلی آزمون پازل بیماری‌ها	آزمون همخوانی با شرح نامه آزمون استدلال تحلیلی	آزمون PMP
آزمون جامع استدلال بالینی			

انواع آزمون‌های استدلال بالینی

همان‌طور که پیش‌تر گفته شد، امروزه اَشکال متعددی از آزمون‌های استدلال بالینی در دسترس است که هر کدام بخشی از فرایند استدلال و تصمیم‌گیری بالینی را می‌سنجند. در ادامه به برخی از انواع آزمون‌های استدلال بالینی اشاره مختصری خواهیم داشت و سپس در فصول بعدی به توضیح مفصل تعدادی از آزمون‌های متداول‌تر خواهیم پرداخت.

آزمون ویژگی‌های کلیدی

آزمون «ویژگی‌های کلیدی» بر پایهٔ این پیش فرض بنا شده است که در حل یک مسأله، همهٔ اطلاعات بیمار ارزش یکسانی ندارند بلکه نکات کلیدی وجود دارد که اهمیت آن‌ها در حل مسأله بیشتر از سایر نشانه‌ها و علائم است و اشتباه در شناسایی آن‌ها باعث شکست در حل درست مسأله می‌شود. در این آزمون ابتدا یک سناریو مطرح می‌شود. سناریو معمولاً کوتاه است چرا که هدف اصلی از آن جمع‌آوری اطلاعات بیشتر است. در ادامه ممکن است تعدادی سؤال کوته‌پاسخ طرح شود و از فراگیر خواسته شود تا خود پاسخ‌های مدنظر را ارائه نماید یا تعدادی گزینه طرح گردد و فراگیر از بین گزینه‌های ارائه شده پاسخ‌های احتمالی را مشخص نماید. جهت مطالعه جزئیات بیشتر و مشاهده نمونه مثال‌های مربوط به این آزمون به فصل دوم این بخش مراجعه کنید.

آزمون همخوانی با شرحنامه

در آزمون «همخوانی با شرحنامه»، فرضیه‌های تشخیصی در متن سؤال ارائه می‌شوند و وظیفه دانشجو ارزیابی فرضیه‌ها است. به این ترتیب در این آزمون وزن و اهمیت داده‌ها در شرحنامه بیماری بر اساس سناریو سنجیده می‌شود. آزمون با یک سناریو آغاز می‌شود که معمولاً کمی مبهم است و راه‌های متعددی را برای فرضیه‌های تشخیصی و درمانی باز می‌گذارد. پس از آن تعدادی سؤال پرسیده می‌شود که هر کدام سه بخش دارند. در بخش اول عبارتی ذکر می‌شود که شامل یک گزینه تشخیصی یا درمانی

1. Schuwirth

مرتبط است. در بخش دوم مطالبی ذکر می‌شوند که به یک یافته جدید در شرح‌حال، معاینه فیزیکی یا آزمایش‌های پاراکلینیک اشاره دارد. در نهایت بخش سوم یک مقیاس لیکرت پنج‌تایی است که فراگیران باید تأثیر یافته جدید را بر احتمال وقوع تشخیص یا درمان مورد نظر (به صورت مثبت یا منفی) مشخص نمایند و شدت این تأثیر را (به صورت صفر، یک یا دو) برآورد کنند. جزئیات مربوط به این آزمون و همچنین نمونه مثال‌های آن، به شکل مفصل در فصل سوم این بخش ارائه شده است.

آزمون استدلال بالینی

آزمون «استدلال بالینی»[1] برخلاف آزمون «همخوانی با شرحنامه» جهت ارزیابی مهارت ساختن فرضیه‌های تشخیصی طراحی شده است. در این آزمون ابتدا سناریویی مطرح می‌شود که اطلاعات آن، برای ارائه تشخیص نهایی کافی نیست و چندین تشخیص محتمل برای آن مطرح است. سپس از داوطلب خواسته می‌شود که تشخیص‌های احتمالی را به صورت پاسخ کوتاه بنویسد. البته در نوع تغییریافته آن از فراگیر خواسته می‌شود که از میان گزینه‌های موجود یک تشخیص را انتخاب کند. یعنی باید از میان یافته‌های بیمار که به صورت گزینه‌هایی مرتب شده‌اند، حداکثر پنج مورد را که با تشخیص انتخاب‌شده همخوانی دارند یا مرتبط هستند، انتخاب کند. در ادامه دانشجو باید به هریک از پاسخ‌ها بر اساس تشخیص موردنظر علامت مثبت یا منفی بدهد. علامت مثبت به معنای این است که این یافته، تشخیص مورد نظر را تأیید می‌کند و علامت منفی به این معناست که یافته، تضعیف‌کننده تشخیص انتخاب شده است. بدیهی است که تمام یافته‌ها نمی‌توانند با علامت منفی مشخص شوند اما آمیزه‌ای از علایم مثبت و منفی یا تماماً مثبت قابل قبول است. در ادامه این آزمون سؤال دیگری ارائه می‌شود که در آن از فراگیر خواسته می‌شود با فرض نادرست بودن تشخیص اولیه، چه تشخیص دیگری برای بیمار مطرح است و مشابه با فرمت سؤال ابتدایی، یافته‌های مرتبط با آن را انتخاب نمایند. جهت مشاهده نمونه مثال‌های مربوط به این آزمون به فصل چهارم همین بخش مراجعه نمایید.

آزمون پازل ادغام‌یافته

آزمون «پازل ادغام‌یافته بیماری‌ها» با این هدف طراحی شده است که مهارت داوطلب در شناسایی شرحنامه یا الگوی بیماری‌ها را ارزیابی کند. بنابراین برعکس بسیاری از آزمون‌های استدلال بالینی از جمله آزمون همخوانی با شرحنامه، در این آزمون برای هر سناریو تنها یک تشخیص مطرح خواهد بود که این امر مستلزم بازشناسی[2] الگوی بیماری‌هاست. این روش مانند آن است که پرونده تعدادی از بیماران در هم ریخته شده است و ما تلاش می‌کنیم تا با خواندن مجدد اطلاعات مربوط به پرونده‌ها، آن‌ها را دوباره مرتب کنیم. در این آزمون، پرونده بیماران به چند قسمت شامل شکایت اصلی و بیماری فعلی[3]، سابقه پزشکی قبلی[4]، معاینه و نتایج پاراکلینیک تفکیک شده است. دانشجویان باید قطعات در هم ریخته را جور کنند و آزمون‌دهنده باید برای هر تابلوی بالینی فقط یک تشخیص انتخاب کند تا موفق شود قطعات مختلف پرونده را با هم جور کند. جهت کسب اطلاعات بیشتر در خصوص این آزمون به فصل چهارم همین بخش مراجعه کنید.

آزمون ساختن فرضیه یا سناریونویسی

هدف از آزمون ساختن فرضیه، ارزیابی توانایی آزمون‌شوندگان در ساخت فرضیه است. این فرضیه‌ها می‌توانند در محور تشخیص یا ارائه مراقبت‌های درمانی باشند. در این آزمون، تعدادی علامت و نشانه به داوطلب داده می‌شود و از او خواسته می‌شود بر اساس این علایم و نشانه‌ها، سناریوی یک بیمار را بنویسد به گونه‌ای که حداکثر تعداد علایم و نشانه‌ها در این سناریو گنجانده شده باشد. در ضمن فراگیر در نهایت باید تشخیص نهایی سناریو را نیز بنویسید. جهت کسب اطلاعات بیشتر و مشاهده نمونه مثال‌های موجود به فصل چهارم همین بخش مراجعه کنید.

1. Clinical Reasoning Problem (CRP)
2. Recognition
3. Present illness
4. Past medical history

آزمون «تدبیر مشکل بیمار» یا PMP

آزمون PMP معمولاً با یک عبارت بالینی در ارتباط با مشکل موجود بیمار، همراه با خلاصه‌ای از شرح‌حال شروع می‌شود. بعد از آن شرح‌حال در چند مرحله به صورت پی در پی مطرح می‌گردد و در هر مرحله از داوطلب خواسته می‌شود تا در مورد اداره و ارزیابی بیمار تصمیم بگیرد. بنابراین در این نوع از ارزشیابی آزمون‌شونده باید مشکل بیمار را درک کند، اطلاعات لازم را برای حل مسأله جمع‌آوری کند، اطلاعات جمع‌آوری شده را تجزیه و تحلیل نماید و از اطلاعات برای حل مسأله استفاده کند.

بدین ترتیب در آزمون PMP داوطلب با یک بیمار که اطلاعات محدودی از او در دسترس می‌باشد، مواجهه می‌شود و بعد از مطالعه اطلاعات باید تصمیم‌گیری نماید که چه اقدامی برای بیمار لازم است. این اقدامات ممکن است شامل درخواست یک‌سری آزمایش‌های پاراکلینیکی یا سایر روش‌های تشخیصی باشد. نهایتاً داوطلب باید در رابطه با درمان و مدیریت روند بیماری تصمیماتی را اتخاذ نماید. امروزه از این آزمون به دلیل روایی و پایایی پایین، زمان‌بر بودن و مشکلات اجرایی کمتر استفاده می‌شود (نورسینی و همکاران ۲۰۰۳). جهت کسب اطلاعات بیشتر به فصل چهارم همین بخش مراجعه کنید.

آزمون جمع‌آوری اطلاعات

هدف اصلی از آزمون «جمع‌آوری اطلاعات»[1]، ارزیابی مهارت دانشجویان در پیدا کردن نکات کلیدی و سنجش میزان اعتماد آنها به داده‌های جمع‌آوری شده است. در این آزمون ابتدا یک تابلوی بیماری ارائه می‌شود و از فراگیر خواسته می‌شود تا به ذکر مواردی بپردازد که از نظر او چندان واضح نیست و در مورد آنها نیاز به کسب اطلاعات بیشتر و اطمینان‌بخشی از صحت اطلاعات وجود دارد. در نهایت از وی خواسته می‌شود که پیشنهادهای خود را برای کسب اطمینان از موارد ذکرشده بیان نماید. به عنوان مثال، فراگیر ممکن است در مورد بیمار مبتلا به فشارخون بالا، اندازه‌گیری مجدد فشارخون یا اخذ شرح‌حال دقیق‌تر از بیمار را توصیه کند. یا در مورد مصرف گلی‌بنکلامید، داوطلب ممکن است تزریق ویال گلوکز ۵۰ درصد، بررسی دقیق آزمایش‌های قبلی و درخواست اندازه‌گیری قند خون را مطرح نماید.

جهت طراحی این سؤال ابتدا تابلوی اولیه بیمار تنظیم شده و سپس بر اساس آن مجموعه‌ای از تشخیص‌های افتراقی مرتبط تعیین می‌گردد و منابع مختلف جمع‌آوری داده از بیمار تعیین می‌شود (به عنوان مثال همراهان بیمار، پرسنل درمانی، مدارک پزشکی، داروها و ...).

آزمون جمع‌آوری اطلاعات

داوطلب گرامی پس از مطالعه سناریوی بیمار، لطفاً موارد زیر را مشخص نمایید.
به چه مواردی در شرح‌حال بیمار برخورد کردید که نیاز به کسب اطمینان از صحت اطلاعات موجود داشت؟ شماره آن را بنویسید.
برای کسب اطمینان از صحت داده‌ها چه پیشنهادهایی دارید؟ راه‌کار ارائه کنید.
بیمار آقای (۱) ۶۵ ساله‌ای (۲) است که با حالت کما به اورژانس آورده شده است. به گفته همراهان، وی نظامی بازنشسته است و تنها زندگی می‌کند. ساعت ۱۰ صبح که فرزندان برای ملاقات به منزل وی رفته‌اند، او را بیهوش (۳) روی زمین پیدا کرده‌اند و آثار استفراغ (۴) روی زمین مشهود بوده است.همراهان وی سابقه‌ای از فشارخون بالا (۵) را ذکر می‌کنند که قرص‌های (۶) آن را مرتب مصرف می‌کرده است. کلاً آدمی است که مقید به مصرف دارو و مراجعه به پزشک است. یک قطره چشمی (۷) هم مصرف می‌کرده است که اسم آن را نمی‌دانند. گزارش پرسنل اورژانس حاکی از آن است که در زمان پذیرش فشارخون بیمار طبیعی (۸) بوده است، اما ضربان قلب زیاد (۹) بوده است، از مریض رگ گرفته شده است و سرمی (۱۰) نیز به وی تزریق شده است. داروهای همراه بیمار شامل قرص کاپتوپریل (۱۱)، گلی‌بن‌کلامید (۱۲)، آسپیرین بچه (۱۳)، دیگوکسین (۱۴)، و دیفنوکسیلات (۱۵) است. یک نسخه مربوط به دو روز قبل (۱۶) است که هیوسین (۱۷)، متوکلوپرامید (۱۸)، و سرم نمکی (۱۹) تجویز شده است. در معاینه مرد لاغری (۲۰) در حالت کما (۲۱) با سردی اندام تحتانی (۲۲) است.
T= 36.7 (۲۵) RR = 22 (۲۴) PR=110/min (۲۳) BP=13/p (۲۶)
مردمک‌ها میوتیک (۲۷) است. خونمردگی (۲۸) در محل پیشانی مشاهده می‌شود. معاینه قلب و ریه طبیعی (۲۹) است. شکم نرمال است. ترگور پوست کاهش یافته است و مخاط‌ها خشک است.

1. Information gathering test

آزمون «استدلال تحلیلی»

از طریق آزمون «استدلال تحلیلی»[1] می‌توان مسیر ساخته شدن فرضیه‌های تشخیصی در ذهن دانشجو را بررسی نمود. در این آزمون سناریو به گونه‌ای طراحی می‌شود که داده‌ها به صورت مرحله به مرحله در اختیار داوطلب قرار گیرند و در پایان هر مرحله از دانشجو خواسته می‌شود که فرضیه‌های تشخیصی خود را بنویسد. برخی از داده‌های ارائه شده در هر مرحله با داده‌های قبلی تناقض دارند. به این شکل توانایی داوطلب در اضافه یا حذف کردن فرضیه‌ها بر اساس داده‌های جدید نیز سنجیده می‌شود. در این آزمون هر چه به سمت قسمت‌های انتهایی سناریو نزدیک‌تر می‌شویم، تعداد فرضیه‌های تشخیصی کاهش می‌یابد.

از وجوه متمایزکننده دانشجویان با متخصصان در این آزمون آن است که دانشجویان پزشکی معمولاً زمانی که اطلاعات مختصری از یک تابلوی بیماری در اختیار دارند، تشخیص‌های بسیار اختصاصی در حد نام مشخص بیماری را مطرح می‌کنند و هر چه اطلاعات بیشتری در طول سناریو به آن‌ها ارائه می‌شود، تشخیص‌های آنان به سمت عمومی‌تر شدن پیش می‌رود. به عنوان مثال، در ابتدا تشخیص بیماری انفارکتوس میوکارد را مطرح می‌کنند اما با ارائه داده‌های بیشتر در سناریو دچار سردرگمی شده و نهایتاً به تشخیص بیماری قلبی بسنده می‌کنند. حال آن که، پزشکان باتجربه کاملاً عکس آن عمل می‌کنند. آنان هنگام مواجه با اطلاعات کم، تشخیص‌های عام‌تری را مطرح کرده و در ادامه با کسب اطلاعات بیشتر و دقیق‌تر از وضعیت بیمار، ضمنِ کاهش چشمگیر تعداد فرضیه‌های تشخیصی، در نهایت به یک بیماری مشخص اشاره می‌کنند.

آزمون جمع‌آوری اطلاعات

داوطلب گرامی لطفاً پس از پایان هر کدام از قسمت‌ها، فرضیه‌های تشخیصی خود را بنویسید و آنها را به ترتیب اولویت شماره‌گذاری کنید. پس از تکمیل هر قسمت، برگهٔ پاسخ از شما پس گرفته می‌شود و اطلاعات جدید در اختیار شما قرار می‌گیرد. اما در مرحله بعدی به اطلاعات مرحله قبلی دسترسی خواهید داشت.

۱. بیمار مردی است ۴۸ ساله و بیکار که با تنگی‌نفس، سرفه و درد قفسه سینه پیشرونده از صبح دیروز به اورژانس آورده شده است.

تشخیص‌های افتراقی:

۲. بیمار بدون سابقه قبلی و چهار هفته پس از سرماخوردگی، از دو سه روز پیش، دچار حملات درد قفسه سینه و تنگی‌نفس گذرا شده است که از صبح دیروز مداوم و پیشرونده شده است. از امروز عصر حتی در حالت استراحت هم درد قفسه سینه، تنگی‌نفس و سرفه داشته است و به اورژانس آورده شده است. دردهای گذرای روز گذشته در همی‌توراکس چپ بوده است که به هردو شانه تیر می‌کشد، با نفس کشیدن و خوابیدن بدتر می‌شود و با تعریق همراه است. با فعالیت، ارتباط چندانی ندارد. بیمار اظهار می‌کند پس از سرماخوردگی چهار هفته پیش، هنوز تب، ضعف و بی‌حالی وی به طور کامل برطرف نشده است. در حال حاضر از سرفه‌های شدید خلط دار شکایت دارد.

تشخیص‌های افتراقی:

۳. بیمار بیست پاکت/سال سیگار مصرف می‌کند. به مدت ۲۰ سال کارگر ذوب آهن بوده است و پنج سال پیش بازنشسته شده است. سابقه هفت ساله پرفشاری خون را می‌دهد که تحت درمان با اتنولول، تریامترن اچ بوده است. در سابقه خانوادگی نکته مثبتی ندارد. در مرور سیستم‌ها از بی‌اشتهایی و بی‌حالی عمومی شکایت دارد.

تشخیص‌های افتراقی:

۴. در معاینه فیزیکی:

Toral=38.2 PR=110/m RR=28/m BP=155/95 mmHg

بسیار چاق، آژیته، بدحال، اورینته و دیافورتیک است. همکاری نسبتاً مطلوب است. دیسترس تنفسی متوسط تا شدید دارد. سیانوز و ایکتر ندارد.

JVP S3 - S4+

در سمع قلب، سوفل سیستولیک ۳/۶ هارش در کنار چپ استرنوم و زایفوئید دارد که در حالت خوابیده بهتر شنیده می‌شود. سمع ریه رال خشن و نرم دمی و بازدمی در سرتاسر ریه دارد. در معاینه شکم، آسیت و هپاتومگالی ندارد.

طحال ده بند انگشت زیر لبه دندهای لمس می‌شود. در معاینه اندام‌ها محل سوراخ شدگی در کوبیتال فوسای هر دو دست مشاهده می‌شود. ادم دارد. نبض‌ها نرمال است.

۵. پاسخ تست‌های درخواستی: EKG: تاکی کاردی سینوسی و CXR: احتقان هر دو ریه، افزایش مارکینگ عروقی، اندازه قلب نرمال، زوایای جنبی باز، دیافراگم‌ها نرمال.

Hb=15.5 with normal indices WBC=22000 with 90% PMN

BUN=35 cr=1.4 Electrolytes= normal

CkMB=8.5

U/A: protein 1+ Granular cast

BC [1]: CoNS [2]: positive

ABG: PH=7.38 Fio2=0.6 HCO3-=32 O2sat=100 PCO2=60

1. Blood Culture
2. Coagulase Negative Staphylococci (CoNS)

1. Analytical Reasoning Test

همان‌طور که در نمونه مثال مشاهده می‌کنید در این آزمون ابتدا یک تابلوی بالینی کوتاه نوشته می‌شود و دانشجو باید تشخیص‌های افتراقی آن را مشخص کند. در مرحلهٔ بعدی سعی می‌شود اطلاعاتی گنجانده شوند که قسمتی از تشخیص‌های افتراقی قسمت اول را حذف می‌کنند. در ضمن اطلاعاتی نیز اضافه خواهند شد که مطرح‌کننده تشخیص یا تشخیص‌های دیگری هستند. این روند به همین منوال ادامه می‌یابد.

پرسشنامه تفکر تشخیصی

پرسشنامه تفکر تشخیصی[1] شامل ۴۱ سؤال است. این پرسشنامه در واقع یک آزمون نیست بلکه در مورد روند استدلال بالینی فراگیران اطلاعاتی را در اختیار ارزیابان قرار می‌دهد. بیست و یک سؤال اول این پرسشنامه جهت سنجش قابلیت انعطاف در استدلال بالینی هستند که در حقیقت چگونگی رخداد تفکر در برخورد با مسائل بالینی را می‌سنجند و معمولاً افراد خبره در آن نمره بالاتری کسب می‌کنند. بیست سؤال بعدی برای سنجش دانش ساختاریافته طرح‌ریزی شده است و نشان‌دهنده این پیش‌فرض است که افراد خبره با ساختار متفاوتی نسبت به افراد تازه کار دانش را در ذهنشان طبقه‌بندی می‌کنند.

در این پرسشنامه از دانشجویان خواسته می‌شود جمله ابتدایی را بخوانند و بعد به دنبال آن، عبارات موجود در دو سویه مقیاس شش‌تایی را مطالعه کنند و میزان تمایل خود را به یکی از این دو وضعیت، به صورت یک علامت ضرب در وسط خطوط و نه روی خطوط نشان بدهند. سؤالات این پرسشنامه، پاسخ درست یا غلط ندارند بلکه مجموع جواب‌های فراگیران در ارتباط با سؤالات معنادار است. همچنین فراگیران به سؤالات باید به گونه‌ای پاسخ دهند که واقعاً عمل می‌کنند نه آنگونه که می‌اندیشند باید باشند. فراگیران ممکن است در بعضی سؤالات هر دو گزینه را درست بیابند ولی در نهایت باید گزینه‌ای را انتخاب کنند که بیشتر به آن صورت عمل می‌کنند.

نمره‌دهی در آزمون‌های استدلال بالینی

همان‌طور که در ابتدای فصل هم گفته شد، دو تفاوت عمده میان آزمون‌های استدلال بالینی و آزمون‌های متداول وجود دارد: اول اینکه در سؤالات استدلال بالینی، جواب‌های محتمل و درست وجود دارد و نه فقط یک جواب قطعی و کاملاً درست. دوم آن که در این سؤالات، نحوهٔ نمره‌دهی متفاوت است. به عبارت دیگر، جواب‌های درست ممکن توسط گروهی از خبرگان مشخص می‌شوند و کلید آزمون بر اساس ارجاع به صفحهٔ مشخصی از یک کتاب مرجع مشخص نمی‌شود.

بنابراین با توجه به تعداد زیاد سؤال مربوط به آزمون‌های استدلال بالینی (به دلیل حل مشکل ویژگی مسأله[2]) و طبعاً تعداد زیاد گزینه‌ها، لازم است تصویر روشنی از شیوهٔ نمره‌دهی وجود داشته باشد. نمره‌دهی در این گونه آزمون‌ها به دو بخش تقسیم می‌شود: نمره‌دهی به هر سؤال و نمره‌دهی به کل آزمون. این دو بخش می‌توانند به روش‌های متفاوتی انجام شوند. حتی در آزمون‌های مختلف استدلال بالینی ممکن است بنا به هدف موردنظر و اجرایی بودن آن شیوه‌های متفاوتی به کار گرفته شود اما باید توجه داشت که در تهیهٔ کلید تمام این آزمون‌ها از گروه خبرگان استفاده خواهد شد که مبتنی بر دو روش زیر است:

- **روش تجمیعی[3]:** اصل زیربنایی این روش نمره‌دهی آن است که پاسخ‌های هر کدام از اعضای گروه خبرگان منعکس‌کننده نظرات ارزشمندی است. به این ترتیب کلیه پاسخ‌های ارائه شده در فرایند نمره‌دهی لحاظ می‌شوند. به عبارت دیگر، نمره هر سؤال بر اساس پاسخ‌های انفرادی پانل متخصصان تعیین می‌شود.

1. Diagnostic thinking inventory
2. Problem specifity
3. Aggregate method

- **روش اجماع‌نظر**[1]: در روش نمره‌دهی اجماع‌نظر، پانل متخصصان به صورت گروهی و مشترک در خصوص پاسخ‌های صحیح مجاز به توافق نظر می‌رسند.

انواع نمره‌دهی به هر سؤال

در نمره‌دهی به هر سؤال این نکته را باید در نظر گرفت که هم نمره هر گزینه و هم نمره مجموعه گزینه‌ها باید مد نظر قرار گیرند. در این مورد سه روش وجود دارد:

- **روش دوتایی**[2]: اگر روش دوتایی را در مورد هر گزینه به کار می‌گیریم، برای جواب‌های درست، نمره یک و برای جواب‌های نادرست، نمره صفر در نظر گرفته می‌شود. این روش نمره‌دهی در مورد آزمون «استدلال بالینی» استفاده می‌شود. تقریباً تمام افراد با این روش آشنایی دارند. تنها نکته قابل توجه و متفاوت در این روش آن است که مقصود از پاسخ صحیح، اشاره به تمام گزینه‌های مورد انتظار است. یعنی اگر داوطلب مجموع گزینه‌هایی مشخصی را که مورد نظر است، انتخاب کرده باشد، امتیاز یک به او تعلق می‌گیرد اما حتی اگر فقط یک گزینه درست را از تمام مجموعه انتخاب نکرده باشد، صفر می‌گیرد.
- **روش امتیازدهی نسبی**[3]: در این روش، به هر سؤال نمره‌ای بین صفر و یک تعلق می‌گیرد؛ بسته به این که جواب‌های داده شده چه میزان با پاسخ‌های درست در کلید تطابق دارند. وزن هر گزینه در این آزمون توسط گروه خبرگان تعیین می‌شود. در آزمون‌های «ویژگی‌های کلیدی»، «جمع‌آوری اطلاعات» و «همخوانی با شرحنامه» از این روش نمره‌دهی استفاده می‌شود. این روش معمولاً در مورد مجموع گزینه‌ها استفاده نمی‌شود.
- **روش مبتنی بر کارایی**[4]: در این روش جواب درست و نادرست انتخاب شده، هر دو لحاظ می‌شوند. در واقع، نوعی روش امتیازدهی نسبی است، با این تفاوت که برای انتخاب گزینه‌های نادرست، امتیاز منفی در نظر گرفته می‌شود. این روش معمولاً موقعی کاربرد دارد که انتخاب‌های نادرست یا برای جان بیمار خطرناک هستند (مثل تجویز یک داروی خطرناک) یا انتخاب آنها هزینه بر دوش سیستم درمانی می‌گذارد (مانند درخواست یک تست پاراکلینیک پرهزینه و نامناسب). در این روش تک‌تک گزینه‌ها و مجموع آنها همزمان با هم در نظر گرفته می‌شود.

انواع نمره‌دهی به کل آزمون

- **روش تجمعی**[5]: در این شیوه سؤالات متفاوت، بر حسب تعداد گزینه‌های درست و وزن‌های آن، نمرات متفاوتی خواهند داشت و مجموع نمرات سؤال‌ها به عنوان نمرهٔ کلی آزمون اعلام می‌شود. بنابراین، نمرهٔ کل یک آزمون از قبل قابل پیش‌بینی نیست بلکه به کلیدی که توسط گروه خبرگان تهیه می‌شود، بستگی دارد. پس از محاسبهٔ نمرهٔ کل آزمون، با یک تناسب ساده می‌توان آن را به پایهٔ ۲۰ یا ۱۰۰ برد.
- **روش میانگینی**[6]: در این روش، میانگین نمرات سؤال‌ها به عنوان نمرهٔ کلی اعلام می‌شود. در این صورت وزن سؤالات متفاوت، یکسان خواهد بود و سؤالات متفاوت، تفاوتی با هم نخواهند داشت.

همانطور که گفته شد در یک آزمون، ترکیب متفاوتی از روش‌های نمره‌دهی به هر سؤال و به کل آزمون می‌تواند وجود داشته باشد. به نظر می‌رسد بسته به شرایط آزمون و نوع سؤالات و گزینه‌ها می‌توان ترکیب‌های متفاوتی از روش‌ها را به کار گرفت. برخی از مطالعات نشان داده‌اند که هر چند ترکیب روش امتیازدهی نسبی و تجمعی به بهبود روایی آزمون

1. Consensus method
2. Dichotomous
3. Partial credit
4. Efficiency
5. Summative method
6. Averaging method

می‌انجامد اما در آزمون‌هایی که از روش میانگینی استفاده کرده‌اند، این میزان بیشتر است (پیج و همکاران ۱۹۹۵، پیج و بوردیج ۱۹۹۵). به طور اختصاصی‌تر، در آزمون «ویژگی‌های کلیدی» تلفیق روش امتیازدهی نسبی و نمره‌دهی میانگینی بهترین نتیجه را به همراه داشته است (پیج و همکاران ۱۹۹۵، پیج و بوردیج ۱۹۹۵). در آزمون «همخوانی با شرحنامه»، روش امتیازدهی نسبی در مورد گزینه‌ها به کار گرفته می‌شود اما از آنجا که بر خلاف آزمون «ویژگی‌های کلیدی» پاسخ‌ها در یک طیف قرار دارند، فاصله از گزینه‌ای که بیشترین امتیاز را در گروه خبرگان آورده است، معنادار است و باید در امتیازدهی لحاظ شود. به عنوان مثال، اگر گزینهٔ ۲- از میان ۱۰ نفر اعضاء گروه خبرگان ۶/۱۰ امتیاز را آورده است و دو گزینهٔ ۱- و ۰ به ترتیب ۲/۱۰ و ۲/۱۰، مسلماً چون ۱- به ۲- نزدیک‌تر است باید از ۰ امتیاز بیشتری بگیرد. منجمی (۱۳۸۹) پیشنهاد می‌کند که بهتر است وزن داده شده به هر یک از این گزینه‌ها در گروه خبرگان را تقسیم بر «۱+ قدرمطلق فاصله تا جواب حداکثر» نمود. به این ترتیب امتیاز ۱-، عدد ۰/۱ و فاصلهٔ صفر ۰/۰۶ خواهد بود (منجمی ۱۳۸۹).

یکی از موارد مناقشه برانگیز در آزمون‌های استدلال بالینی استفاده از گروه خبرگان در تهیهٔ کلید و تصحیح آزمون‌های استدلال بالینی است. عده‌ای بر این گمان هستند که استفاده از چنین روشی تصحیح را سلیقه‌ای می‌کند و به اعتبار و سلامت آزمون خدشهٔ جدی وارد می‌کند. این افراد به دنبال عینی کردن آزمون‌های استدلال بالینی هستند (یعنی تهیه یک کلید که جواب‌های منعطف در آن راهی ندارد) اما توجه ندارند که این ایراد دقیقاً به همان محدودیتی برمی‌گردد که اساساً این گونه آزمون‌ها برای برطرف کردن آن طراحی شده‌اند یعنی سنجش توانایی استدلال و تصمیم‌گیری در شرایط عدم قطعیت. در واقع اگر قرار باشد مسأله فقط یک راه‌حل صحیح نداشته باشد، چگونه می‌شود راه‌حل‌های درست و ممکن را مشخص کرد. سپردن حل مسأله به دست عده‌ای از متخصصان که بر اساس معیارهای مشخص انتخاب شده‌اند، می‌تواند راهکار خوبی باشد. مبنای استفاده از گروه خبرگان، دقیقاً همین است اما باید توجه داشت که این گروه نباید همان طراحان سؤال باشند. از سوی دیگر هر کدام باید به تنهایی به سؤالات پاسخ دهند و نباید در پاسخگویی به سؤالات با هم مشورت کنند چرا که در آن صورت، پاسخ‌ها بر مبنای توافق شکل می‌گیرد و آرای اقلیت حذف می‌شود و این همان چیزی است که در این گونه آزمون‌ها باید از آن احتراز کرد.

این که گروه متخصصان با چه معیار و ملاکی باید انتخاب شوند و چه تعداد باید باشند، در ادامه همین فصل مورد بحث قرار خواهد گرفت اما نکته حائز اهمیت این است که بر عکس تصور عموم مبنی بر آنکه گروه‌های مختلف متخصصان، کلیدهای کاملاً متفاوتی را استخراج می‌کنند، نتایج مطالعات نشان داده است که گزینه‌هایی که بیشترین امتیازها را به خود اختصاص می‌دهند، در میان هیأت‌های مجرب متفاوت، یکسان هستند.

انتخاب گروه خبرگان بر اساس هدف ما از ارزیابی و سطح آزمون‌دهندگان صورت می‌گیرد. به عنوان مثال، اگر قرار است گروهی از دانشجویان پزشکی را ارزیابی کنیم در مقایسه با وقتی که می‌خواهیم دستیاران قلب را امتحان کنیم به گروه متفاوتی نیاز داریم. اگر از فوق تخصص‌های مختلف دعوت کرده‌ایم تا به عنوان گروه خبرگان برای آزمودن استدلال بالینی دانشجویان دوره پزشکی عمومی شرکت کنند، باید به آنها گوشزد کنیم که با دید کل‌نگر و همسطح با انتظار ما از عملکرد یک دانشجوی پزشکی عمومی به آزمون نگاه کنند. اگر احساس کردیم که این گروه از متخصصان نتوانسته‌اند انتظار ما را برآورده کنند، باید از متخصصان داخلی یا پزشکان عمومی آموزش‌دیده و باتجربه استفاده کنیم. هر عضو گروه خبرگان یک رأی دارد و نباید بر اساس نوع مسأله طرح‌شده، وزن بیشتری به یکی از تخصص‌ها داد. به عنوان مثال اگر یکی از سؤالات در مورد بیماری کلیه است، نباید وزن بیشتری به نظر فوق تخصص کلیه داده شود.

مزایا و محدودیت‌های آزمون‌های استدلال بالینی

مزایای آزمون‌های استدلال بالینی

- **ارزیابی در موقعیت‌های مشابه واقعی:** در این سؤالات امکان طراحی سؤالات مشابه با موقعیت‌های واقعی‌تر فراهم است که موجب بهبود روایی صوری می‌شود. مجموعه آزمون‌هایی که در این دسته قرار می‌گیرند، فراگیر را به ارائه راه‌حل یا استدلال سوق می‌دهند. به عبارت دیگر در این آزمون‌ها داوطلب باید در زمان تصمیم‌گیری، واحدهای مختلفی اطلاعاتی را در برابر سایر اطلاعات موجود ارزش‌گذاری کند و سپس بهترین تصمیم و راه‌حل ممکن را ارائه نماید. هرچند به نظر می‌رسد سنجش مهارت‌های استدلال بالینی در ارتباط با موقعیت‌های واقعی و در بالین بیمار از اعتبار بیشتری برخوردار است، همواره سعی می‌شود تا طراحی و ابداع ابزارهای ارزیابی که از قابلیت اجرای بالاتری برخوردار هستند و مشکلات مربوط به ارزیابی در بالین از قبیل احتمال بروز آسیب به بیمار در آن‌ها وجود ندارد، صورت گیرد.
- **امکان سنجش سطوح بالای یادگیری:** در این سؤالات استفاده از سناریوی بالینی کمک می‌کند تا سطوح شناختی بالا مورد ارزیابی قرار گیرند زیرا از دانشجو خواسته می‌شود تا معلومات و محفوظات خود را در بستر یک موقعیت نزدیک به واقعیت به کار برد. در این شرایط فراگیران به منظور پاسخ‌دهی صحیح به این سؤالات نیازمند درک مطالب آموخته شده، برقراری ارتباط مناسب بین آن‌ها و در نهایت نحوه کاربرد دانسته‌ها است.
- **کمرنگ بودن مشکل ویژگی محتوا:** این شکل از آزمون‌ها مخصوصاً در مقابل آزمون‌های تشریحی و شفاهی می‌توانند از طریق طرح چندین تابلوی بالینی در یک آزمون واحد، مشکل ویژگی محتوا را برطرف کنند.
- **حذف امکان حدس زدن:** در این سؤالات امکان حدس زدن و تقلب در مقایسه با تعدادی از سؤالات کتبی از جمله سؤالات چندگزینه‌ای بسیار کمتر است. همان‌طور که در فصل دوم همین بخش نیز شرح داده می‌شود، آزمون‌های استدلال بالینی می‌توانند به شکل کوته‌پاسخ یا ارائه مجموعه‌ای از گزینه‌ها طرح شوند. بنابراین امکان حدس زدن در آزمون‌های استدلال بالینی که از نوع کوته‌پاسخ هستند، وجود ندارد. در آزمون‌هایی نیز که در قالب ارائه مجموعه‌ای از گزینه‌ها طرح می‌شوند، از آنجا که اغلب تعداد گزینه‌های ارائه شده به طور متوسط بین ۱۵ تا ۲۰ گزینه است، احتمال حدس زدن در مقایسه با سؤالات چندگزینه‌ای بسیار پایین می‌آید.
- **انعطاف‌پذیری در ارائه پاسخ‌ها:** در آزمون‌های استدلال بالینی وجود چند پاسخ درست به رسمیت شناخته می‌شود. این ویژگی منجر گردیده است تا فرایند تصحیح این گونه سؤالات با آزمون‌های عینی قدری متفاوت باشد چرا که باید به دنبال تمام پاسخ‌های محتمل بود. از این رو تهیه کلید پاسخ این سؤالات همواره توسط تیمی از متخصصان صورت می‌گیرد تا بدین ترتیب تمام پاسخ‌های محتمل شناسایی شود.

محدودیت‌های آزمون‌های استدلال بالینی

- **زمان‌بر بودن فرایند طراحی سؤالات:** مسلماً طراحی آزمون‌های استدلال بالینی طبق راهنماها و دستورالعمل‌های موجود، فرایندی زمان‌بر و دشوار است. البته قسمتی از این موضوع از عدم آشنایی استادان با این شکل از آزمون‌ها نشأت می‌گیرد و به مرور زمان که توانمندی طراحان افزیش می‌یابد، بهتر می‌شود.
- **دشواری تهیه کلید و تصحیح سؤالات:** از آنجا که در آزمون‌های استدلال بالینی صورت سؤال به شکل سناریوی بالینی طراحی می‌گردد، تنها یک جواب صحیح منحصر به فرد وجود ندارد و باید به دنبال جواب‌های درست متعدد گشت. اگر به این شیوه عمل شود، تصحیح این گونه آزمون‌ها نیز با آزمون‌های عینی قدری متفاوت خواهد بود، چرا که باید به دنبال تمام پاسخ‌های محتمل بود. از این رو همان‌طور که پیشتر نیز اشاره شد، فرایند تهیه کلید و تصحیح

اوراق امتحانی نیز به زمان زیادی نیاز دارد و مستلزم تشکیل جلسات متعدد با حضور متخصصان می‌باشد.

- **زمان‌بر بودن پاسخ‌دهی:** همان‌طور که در مطالب پیشتر نیز اشاره شد، فراگیران برای پاسخ‌دهی به این سؤالات باید به تحلیل و سازماندهی مطالب مختلف بپردازند. در اغلب اوقات این فرایند به زمان بیشتری نیاز دارد.

در جدول ۲-۱۳ مزایا و محدودیت آزمون‌های استدلال بالینی به شکل خلاصه ارائه شده است.

جدول ۲-۱۳: مزایا و محدودیت‌های آزمون‌های استدلال بالینی

نکات مثبت	نکات منفی
• روایی صوری بالا • امکان ارائه اطلاعات منطبق با شرایط واقعی • امکان سنجش سطوح بالای حیطه شناختی • امکان پرسش از موقعیت‌های بالینی متنوع • احتمال تشویق فراگیران به یادگیری عمیق مطالب • نبود احتمال حدس زدن	• طراحی دشوار • زمان بر بودن پاسخ‌دهی • دشواری تهیه کلید و تصحیح سؤالات

سودمندی آزمون‌های استدلال بالینی

مسلم است هیچ روش ارزیابی وجود ندارد که برای تمام موقعیت‌های ارزیابی قابل به کارگیری و مناسب باشد. هر کدام از ابزارهای ارزیابی دارای مزایا و محدودیت‌هایی هستند که باعث می‌شود بسته به شرایط از آنها استفاده کرد. از این رو، یکی از دغدغه‌های همیشگی طراحان سؤال این است که در موقعیت حاضر استفاده از کدام یک از ابزارها کارآمدتر خواهد بود. همان‌طور که پیشتر نیز اشاره شد، یکی از رویکردهای خوب برای تصمیم‌گیری در مورد سودمندی یک ابزار، استفاده از فرمول سودمندی ون‌درولوتن است که پنج معیار روایی، پایایی، قابلیت پذیرش، تأثیر آموزشی و هزینه را در بر می‌گیرد. در ادامه به ذکر ویژگی‌های سودمندی آزمون‌های استدلال بالینی به صورت کلی خواهیم پرداخت.

روایی

آزمون‌های استدلال بالینی به این دلیل که عملکرد فراگیر را در مواجهه با سناریوهای بالینی می‌سنجند، دارایی روایی قابل ملاحظه‌ای هستند. در آزمون‌های استدلال بالینی، روایی محتوایی یکی از مسائل مهم به شمار می‌رود. بدین معنا که سؤالات به چه میزان می‌توانند حوزه‌ی محتوایی مورد نظر را پوشش دهند. به این منظور باید ملاک‌هایی تعیین شود تا بر اساس آن‌ها، سؤالات انتخاب شوند. به عنوان مثال اگر قرار است در مورد طب اورژانس تعدادی سؤال طرح شوند، انتخاب می‌تواند بر اساس شیوع بیماری‌ها، شدت و میزان مخاطره‌آمیز بودن آن برای حیات بیماران و سن و جنس بیماران صورت گیرد. بدیهی است که در این شرایط استفاده از جدول مشخصات آزمون یا بلوپرینت بسیار کمک‌کننده است. بنابراین به صورت کلی می‌توان گفت که مشخص کردن ارتباط محتوای آزمون با اهداف دوره و همچنین استفاده از گروه خبرگان برای طراحی آزمون‌های استدلال بالینی به بهبود روایی محتوایی این آزمون‌ها می‌انجامد. یکی از مزایای آزمون «ویژگی‌های کلیدی» این است که می‌توانند بین عملکرد متخصصان و افراد غیرمتخصص تمایز قائل شوند. این موضوع مؤید بالا بودن روایی سازه این آزمون‌هاست. همچنین طراحی سؤالات مشابه با موقعیت‌های واقعی روایی صوری را در این شکل از آزمون‌ها بهبود می‌بخشد.

پایایی

یکی از جنبه‌های مفهوم پایایی در آزمون‌های استدلال بالینی این است که اگر فراگیران یکسان توسط مصححان مختلف و با فاصله زمان‌های متفاوت توسط این آزمون‌ها ارزیابی شوند، تا چه حد نتایج مشابهی به دست خواهد آمد. بنابراین در آزمون‌های استدلال بالینی، اطمینان از ثبات نتایج حاصل به فاکتورهای متعددی از قبیل تهیه کلید پاسخ ساختارمند، تعداد و شکل سؤالات آزمون، زمان آزمون و تعداد مصححان وابسته است.

مقبولیت

در نظر گرفتن این نکته ضروری است که هر چقدر یک آزمون خوب باشد، تا زمانی که از سوی ذی‌نفعان آن پذیرفته نشود، قابلیت اجرا نخواهد داشت. این موضوع به حدی مهم است که همواره به دنبال طراحی و ارائه یک ابزار ارزیابی جدید، این پرسش مطرح می‌شود که قابلیت پذیرش آن چقدر است. بنابراین در این بین دریافت بازخوردها و نظرات تمام ذی‌نفعان از قبیل دانشجویان، آزمونگران و طراحان سؤال اهمیت دارد. اکثر مطالعات انجام شده در زمینه آزمون‌های استدلال بالینی، مقبولیت این ابزارها را با بررسی نگرش فراگیران و اعضای هیأت علمی انجام داده‌اند.

در صورت اجرای این آزمون‌ها برای اولین بار، حتماً ایجاد آمادگی در فراگیران بسیار ضروری است. این کار می‌تواند از طریق ارائه نمونه سؤالات مشابه صورت گیرد. همچنین توضیح دادن روند اجرای آزمون و ارائه دستورالعمل‌های آزمون در پذیرش دانشجویان تاثیرگذار است. به علاوه، از طریق برگزاری کارگاه‌های آموزشی و ارائه راهنماهای طراحی سؤالات می‌توان به افزایش قابلیت پذیرش این آزمون در بین استادان و ارزیابان پرداخت.

تأثیر آموزشی

آزمون‌های استدلال بالینی مانند سایر ابزارهای ارزیابی علاوه بر سنجش مهارت‌های فراگیران، بر میزان یادگیری آنان تأثیر دارد. هر چند آزمون‌های استدلال بالینی به سبب طراحی به شکل سناریو منجر به هدایت فراگیران به سمت یادگیری کاربردی و عملیاتی مطالب درسی می‌شوند، با این حال باید توجه داشت که این آزمون‌ها ممکن است همواره تأثیر مثبتی را بر رفتار فراگیران به دنبال نداشته باشند. به عنوان مثال، در صورت طراحی نامناسب این شکل از آزمون‌ها فراگیر ممکن است تنها به یادگیری حقایق هدایت شود.

هزینه

هزینه آزمون‌های استدلال بالینی را می‌توان بسته به مرحله طراحی و اجرای آن‌ها تحلیل نمود.بخشی از هزینه، صرف آماده‌سازی مقدمات لازم برای طراحی از قبیل توانمندسازی و آموزش آزمونگران، مصححان و فراگیران می‌شود. بخشی دیگر از هزینه‌ها به فرایند طراحی از قبیل طراحی سناریوها، سؤالات و کلید پاسخ مربوط می‌شود. همچنین در فرایند اجرای این آزمون‌ها، قسمتی از هزینه صرف تأمین حفاظت آزمون و تصحیح برگه‌های امتحانی می‌شود.

مسائل چالشی آزمون‌های استدلال بالینی

گروه خبرگان در فرایند نمره‌دهی آزمون‌های استدلال بالینی باید شامل چند متخصص باشد؟

تعداد متخصصان در گروه خبرگان باید به اندازه‌ای باشد که بتواند بیان کننده تنوع بین پاسخ‌های متخصصان مختلف باشد. گفته می‌شود افزایش تعداد افراد موجب افزایش پایایی می‌گردد. همچنین توصیه می‌شود که بهتر است طراحی

سؤالات توسط یک گروه کوچک صورت گیرد و سپس یک گروه بزرگتر سؤالات را ارزیابی کند (منجمی ۱۳۸۹). گروه خبرگان نباید با گروه طراحان سؤالات یکسان باشد، چرا که طراحان سؤال ممکن است نسبت به سؤالاتی که طراحی کردند سوگیری داشته باشند و بر نظر بقیه گروه خبرگان نیز تأثیر بگذارند. ضمناً ممکن است تعداد طراحان بسیار کمتر از تعداد اعضاء گروه خبرگان باشد.

در مورد تعداد اعضای گروه خبرگان نیز در مطالعات متعدد نتایج متفاوتی ارائه شده است، بنا به برخی مطالعات، تعداد اعضای گروه خبرگان از ۵ تا ۲۶ نفر پیشنهاد شده است (چارلین و همکاران ۲۰۰۰، سیبرت و همکاران[1] ۲۰۰۵). هر چند گانون و همکاران[2] (۲۰۰۵) در پژوهشی، حضور ۲۰ متخصص را در امتحانات مهم و سرنوشت‌ساز توصیه کردند، هامبر و همکاران[3] (۲۰۱۱) در پژوهش خود توصیه کردند که در امتحانات مهم و سرنوشت‌ساز با حضور حداقل ۱۵ متخصص در گروه خبرگان می‌توان به سطح قابل قبولی از پایایی دست یافت. نکته مهم اینکه، فورنیر و همکاران[4] (۲۰۰۸) دریافتند که هرچند پایایی بیشتر با تعداد متخصصان بیشتر حاصل می‌گردد، این افزایش پایایی تنها تا مرز ۲۰ متخصص در گروه خبرگان است و پس از آن بهبودی در پایایی رخ نمی‌دهد. همچنین نتایج یک مرور نظام مند نشان داده است که حضور ۱۰ تا ۲۰ متخصص در گروه خبرگان برای رسیدن به سطح قابل قبولی در پایایی نمرات ضروری است (دوری و همکاران[5] ۲۰۱۲). تعداد بیشتر از این معمولاً بر گستره پاسخ‌ها نمی‌افزاید و تعداد کمتر از این هم همهٔ پاسخ‌های درست ممکن را پوشش نمی‌دهد.

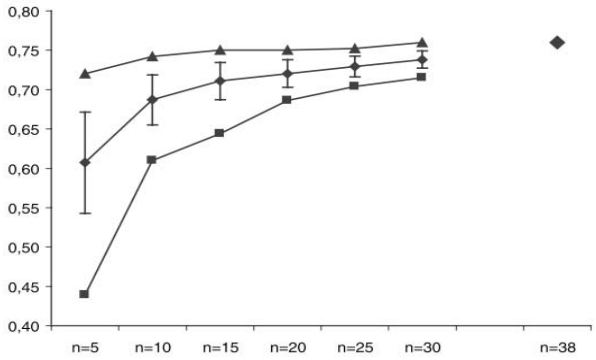
البته این تعداد عضو در گروه خبرگان در مواقعی مفید است که آزمون سطح بالا برگزار می‌کنیم. در آزمون‌های با درجه اهمیت پایین، تشکیل گروه خبرگان با تعداد کمتر (بین ۵ تا ۱۰ نفر) نیز امکان‌پذیر خواهد بود اما باید توجه داشت که تعداد کمتر از ۱۰ نفر ممکن است به پایایی آزمون لطمه بزند.

در رابطه با افزایش تعداد گروه خبرگان نتیجه غیرقابل انتظاری در مطالعه گانون و همکاران (۲۰۰۵) مشاهده شد. این نویسندگان از مجموع ۳۸ متخصصی که در گروه خبرگان حضور داشتند، با استفاده از روش‌های آماری به صورت تصادفی نمونه‌گیری کردند (۵، ۱۰، ۱۵، ۲۰، ۲۵ و ۳۰) و برای هر یک از هیأت‌های مجرب نمونه‌گیری شده، موارد زیر را انجام دادند:

- پاسخ برای هر یک از هیأت‌های مجرب نمونه‌گیری شده محاسبه شد.
- بر اساس پاسخ هر گروه خبرگان، نمرات فردی داوطلبان محاسبه شد.
- میانگین نمرات و ضریب آلفای کرونباخ برای هر یک از هیأت‌های مجرب تعیین شد.

گروه خبرگان شامل ۳۸ پزشک خانواده بود. میانگین سنی آن‌ها ۵۱/۴ سال و میانگین سابقه کار در بالینی ۲۲/۶ بود. پایایی آزمون در گروه خبرگان ۳۸ نفره، ۰/۷۶ گزارش شد. زمانی که اندازه گروه خبرگان از پنج نفر به ۱۰ نفر افزایش یافت، تفاوت واضحی در میزان دقت اندازه‌گیری‌ها مشخص شد (به ترتیب پایایی ۰/۶۲ و ۰/۷۰). زمانی که اندازه گروه خبرگان به ۲۰ نفر افزایش یافت، این افزایش پایایی کمتر بود (پایایی معادل ۰/۷۴). مقایسه افزایش تعداد گروه خبرگان و تأثیر آن در پایایی نمره‌دهی در شکل شماره ۱-۱۳ نشان داده شده است.

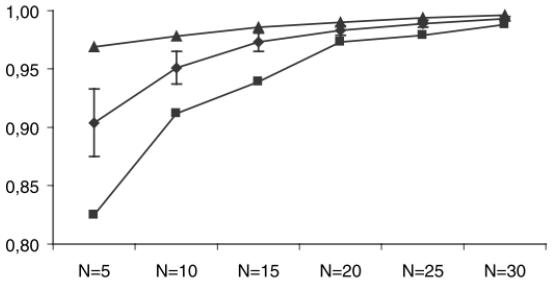
1. Sibert et al.
2. Gagnon et al.
3. Humber et al.
4. Fournier et al.
5. Dory et al.



شکل ۱-۱۳. میانگین ضریب پایایی برای اندازه‌های متفاوت گروه خبرگان

محققان بر اساس نتایج حاصل نتیجه گرفتند که حضور ۲۰ متخصص در گروه خبرگان، تخمین قابل قبولی از پایایی را فراهم می‌سازد. به طور کلی، به دلیل وجود تنوع زیاد در پاسخ‌های ارائه شده توسط اعضای گروه خبرگان حتی در گروه‌های همگن، تعداد متخصصان در گروه خبرگان باید به اندازه‌ای بزرگ باشد تا بتوان نتایج پایایی کسب کرد.

همچنین همبستگی بین نمرات به دست آمده از گروه‌های خبرگان با تعداد اعضای متفاوت با نمرات حاصل از گروه خبرگان ۳۸ نفره، در پانل ۵ نفره برابر با ۰/۹۰، پانل ۱۰ نفره ۰/۹۵، ۲۰ نفر ۰/۹۸ و ۳۰ نفره ۰/۹۹ بود. محققان ذکر کردند که حضور ۱۵ تا ۲۰ متخصص در گروه خبرگان منجر شد که نمرات همبستگی بالایی با نمرات حاصل از گروه خبرگان ۳۸ نفره داشته باشد (شکل ۲-۱۳).



شکل ۲-۱۳. ضریب همبستگی پیرسون برای اندازه‌های متفاوت گروه خبرگان

اما مساله غیرقابل انتظار این بود که با بالا رفتن تعداد اعضای گروه خبرگان، علاوه بر موارد فوق، نمرات داوطلبان نیز افزایش می‌یافت. به عبارت دیگر، میانگین نمرات داوطلبان در گروه خبرگان شامل پنج متخصص ۶۲/۹ بود، در حالی که میانگین نمرات داوطلبان در گروه خبرگان شامل ۳۰ متخصص ۶۹/۶ بود. یعنی امتیاز تعدادی از گزینه‌ها تحت تأثیر تعداد متخصصان گروه خبرگان قرار داشت. هر چند این نتیجه علی‌رغم معنادار بودن در هیچ مطالعه دیگری تأیید نشده است.

گروه خبرگان در فرایند نمره‌دهی آزمون‌های استدلال بالینی باید دارای چه ویژگی‌هایی باشند؟

یکی از سؤالاتی که در مورد تعیین اعضای گروه خبرگان مطرح می‌شود، مسأله انتخاب اعضا بر اساس ویژگی‌های موردانتظار است. اعضای گروه خبرگان می‌توانند از میان اعضای هیأت علمی، متخصصان با تجربه و حتی دانشجویان سال آخر انتخاب شوند. نتایج مطالعات مؤید این امر است که در آزمون‌های مهم و سرنوشت ساز باید دقت کافی در انتخاب گروه خبرگان به عمل آید تا بتوان طیف وسیعی از محیطهای بالینی مورد انتظار را پوشش داد.

- دو مطالعه به بررسی نتایج حاصل از آزمون از منظر شباهت بین محیطهای کاری بین داوطلبان و اعضای گروه خبرگان پرداختند. بر این اساس مشخص شد، زمانی که اعضای گروه خبرگان در محیطهای مشابه با محیطهای بالینی داوطلبان اشتغال دارند، میانگین نمرات داوطلبان قدری بالاتر خواهد بود (سیبرت و همکاران ۲۰۰۲، چارلین و همکاران ۲۰۰۷). در مطالعه چارلین و همکاران (۲۰۰۷) مشخص شد که هیچ تفاوت معناداری در الگوی پاسخ بین اعضای هیأت علمی و غیرهیأت علمی وجود ندارد.
- در مطالعه‌ای دیگر، عملکرد اعضای دارای تخصص عمومی در زمینه مامایی با افراد دارای تخصص‌های ویژه در حیطه‌های مختلف مورد ارزیابی در گروه خبرگان مقایسه شد (گانتلت[1] ۲۰۰۸). بر اساس نتایج حاصل، در نمرات داوطلبان و ویژگی‌های روانسنجی آزمون تفاوت قابل توجهی مشاهده نشد.
- همچنین بر اساس نتایج حاصل از یک مطالعه دیگر محققان پیشنهاد دادند که گروه خبرگان باید شامل پزشکان با عملکرد بالینی بسیار خوب در حوزه‌های مورد سنجش باشد (فورنیر و همکاران ۲۰۰۸).
- پژوهش سیبرت و همکاران (۲۰۰۹) نشان داد که نمرات تعیین‌شده توسط گروه خبرگانی که عملکرد آموزشی داشتند، به طور معنادار بالاتر از گروه خبرگانی بود که عملکرد غیرآموزشی داشتند. بر اساس نتایج حاصل از این مطالعه محققان نتیجه گرفتند که نمرات دستیاران همبستگی بیشتری با استادان خود دارد تا پزشکان شاغل.
- در پژوهش بروشتین و همکاران[2] (۲۰۱۰) روی دستیاران درماتولوژی مشخص شد که نمرات آزمون اساساً تحت تأثیر نوع شرایط کاری دستیاران (بیمارستان در مقابل مطب خصوصی) قرار دارد تا سطح تخصصی آن‌ها (درماتولوژیست در مقابل پزشک عمومی).

در هر حال، مقایسه نتایج حاصل از ترکیب‌های متفاوت از گروه خبرگان، توجه ما را به این واقعیت جلب می‌کند که هر کدام از ویژگی‌های گروه خبرگان بر بخشی از خطای اندازه‌گیری تاثیرگذار خواهند بود. در حال حاضر، هیچ گونه شواهد نظام‌مندی برای تصمیم‌گیری در خصوص شیوه انتخاب اعضاء گروه خبرگان وجود ندارد. به علاوه، در هیچ مطالعه‌ای تأثیر آموزش مصححان بر نتایج آزمون بررسی نشده است.

افزایش تعداد موارد بالینی یا افزایش سؤالات، کدام یک بر پایایی آزمون‌های استدلال بالینی بیشتر تاثیرگذار است؟

یکی از سؤالاتی که ذهن طراحان سؤال و دست‌اندرکاران آموزش را به خود مشغول کرده است، این است که در آزمون استدلال بالینی بهتر است تعداد سؤالات بیشتری طراحی نمود یا تعداد موارد بالینی بیشتری را طرح کرد.

در آزمون‌های استدلال بالینی با در نظر گرفتن مسأله ویژگی مورد، انتظار می‌رود که سهم واریانس خطای مربوط به موارد بالینی بالا باشد و سهم واریانس خطای مربوط به سؤالات طرح شده در هر سناریو پایین باشد. جالب آن که یافته‌های مطالعه نورمن و همکاران در ارزیابی آزمون «ویژگی‌های کلیدی» نشان داد که واریانس نسبتاً کمی به تفاوت بین موارد

1. Gantelet
2. Bursztejn et al.

بالینی مربوط می‌شود و در حدود ۸۰ درصد واریانس خطا به تفاوت عملکردی داوطلبان در پاسخ به سؤالات هر مورد بالینی مربوط می‌شود. محققان بر اساس یافته‌های حاصل مشاهده کردند که طبق مفروضات قابل‌قبول در زمینه توزیع زمانی آزمون و محدودیت‌های اجرایی (از قبیل محدودیت اجرا آزمون در مدت زمان ۳ ساعت)، طرح موارد بالینی کمتر (مشتمل بر ۲ تا ۳ سؤال به ازای هر سناریو) در عمل به بهبود بیشتر پایایی آزمون می‌انجامد. به علاوه از نظر محققان، این رویکرد مزایایی هم از نظر عملیاتی و هم قابلیت پذیرش به همراه خواهد داشت از جمله آنکه فرایند طراحی تعدادی سناریو مشتمل بر چندین سؤال آسانتر از طرح چندین سناریو حاوی یک سؤال است. از دیدگاه دانشجویان نیز پاسخگویی به چندین سؤال مرتبط با یک مورد بالینی (از قبیل جمع آوری اطلاعات، تشخیص و مدیریت) راحت‌تر از زمانی است که پاسخگویی به هر سؤال نیازمند مطالعه سناریو مجزایی است. از این رو براساس یافته‌های پژوهش محققان پیشنهاد کردند که افزودن سؤالات بیشتر مرتبط به هر موارد بالینی از منظر پایایی آزمون به جای افزودن تعداد بیشتر موارد بالینی، کارسازتر است (نورمن و همکاران ۲۰۰۶).

بر این اساس، گانون و همکاران (۲۰۰۹) با استفاده از نظریه تعمیم‌پذیری به بررسی نتایج حاصل از سه آزمون «همخوانی با شرحنامه» در رشته‌های انکولوژی، پرستاری و کودکان پرداختند تا بدین ترتیب منابع تاثیرگذار بر واریانس آزمون را در ارتباط با تعداد موارد بالینی و سؤالات هر مورد بالینی شناسایی نمایند (G study). ضریب تعمیم‌پذیری[1] برای آزمون‌های اجرا شده در پرستاری و انکولوژی در حد قابل قبول (۰/۷۸ و ۰/۸۸) و برای کودکان به طور قابل ملاحظه‌ای پایین بود (۰/۶۳). نتایج نشان داد که منبع اصلی واریانس نمرات به تعامل بین داوطلبان و تعداد سؤالات مربوط به هر مورد بالینی مربوط می‌شد (۹۲ درصد و ۸۵ درصد). برخلاف آن، واریانس مربوط به تعامل بین داوطلبان و تعداد موارد بالینی پایین بود (صفر درصد و و ۱/۸ درصد). به علاوه نتایج نشان داد که واریانس مربوط به سؤالات بیشتر از واریانس مربوط به موارد بالینی بود. نویسندگان بر اساس نتایج حاصل پیشنهاد کردند که احتمالاً تأثیر افزایش تعداد سؤالات بالینی بر پایایی آزمون به مراتب بیشتر از افزایش تعداد موارد خواهد بود. به عنوان مثال، در پرستاری افزایش تعداد سؤال از یک به سه منجر به بالا رفتن ضریب تعمیم‌پذیری در حدود ۰/۲۵ شد، در حالی که طرح پنج مورد بالینی به افزایش ۰/۱۰ انجامید. به علاوه محققان با استفاده از D study به تعیین تعداد مطلوب موارد بالینی و سؤالات هر مورد بالینی که به حداکثر میزان پایایی آزمون منجر می‌شود، پرداختند و بر اساس نتایج پیشنهاد کردند که طرح دو سؤال (۳۸–۳۷ مورد بالینی)، سه سؤال (۲۵ مورد بالینی) یا چهار سؤال (۱۸–۱۹ مورد بالینی) برای هر مورد بالینی کفایت خواهد کرد.

1. G coefficient

منابع

1. Barrows HS, Feightner JW, Neufeld VR, Norman GR. Analysis of the Clinical Methods of Medical Students and Physicians. Hamilton, Ontario: McMaster University1978 .
2. Bursztejn AC, Cuny JF, Adam JL, Sido L, Schmutz JL, de Korwin JD, Latarche C, Braun M, Barbaud A. Test de concordance de script en dermatologie: e´valuation du choix du panel d'experts. Pe´dagogie Me´dicale 2010; 11 (Suppl 1):69.
3. Bordage, G, Page G. An Alternative to PMPs: The "Key Features" Concept. Further Developments in Assessing Clinical Competence, 2nd Ottawa Conference, 1987, 59-75.
4. Charlin B, Desaulniers M, Gagnon R, Blouin D, Van der Vleuten C. Comparison of an aggregate scoring method with a consensus scoring method in a measure of clinical reasoning capacity. Teach Learn Med 2002, 14:150-156
5. Charlin B, Boshuizen HPA, Custers EJ, & Feltovich PJ. Scripts and clinical reasoning. Medical Education 2007; 41: 1178-1184.
6. Charlin B, Gagnon R, Sauve E, Coletti M. Composition of the panel of reference for concordance tests: do teaching functions have an impact on examinees' ranks and absolute scores? Medical Teacher 2007;29 (1):49–53.
7. Charlin B, Roy L, Brailovsky C, et al. The Script Concordance Test: a tool to assess the reflective clinician. Teaching and Learning Medicine 2000;12:189-195
8. Custers, Eugène JFM. Medical education and cognitive continuum theory: an alternative perspective on medical problem solving and clinical reasoning. Academic Medicine 2013; 88(8): 1074-1080.
9. Dory V, Gagnon R, Vanpee D, Charlin B. How to construct and implement script concordance tests: insights from a systematic review. Medical Education 2012; 46: 552–563
10. Elstein AS, Shulman LS, Sprafka SA. Medical problem solving an analysis of clinical reasoning. Cambridge, Massachusetts: Harvard University Press; 1978.
11. Elstein AS, Schwarz A. Clinical problem solving and diagnostic decision making: selective review of the cognitive literature. British Medical Journal 2002;324.7339: 729.
12. Feltovich PJ, Barrows HS. Issues of generality in medical problems solving. In: Schmidt HG, de Volder ML, eds. Tutorials in Problem-Based Learning: A New Direction in Teaching the Health Professions. Assen, The Netherlands: Van Garcum42–1984:128 .
13. Feletti GI, Saunders NA, Smith AJ. Comprehensive assessment of final year medical student performance based on undergraduate program objectives. The Lancet 1983; 322(40): 34-7.
14. Fournier P, Demeester A, Charlin B. Script concordance tests: guidleines for construction. BMC medical informatics and decision making 2008;8(18): 1.

15. Gantelet M. Impact du panel de re´fe´rence sur les re´sultats d'un test de concordance de script (TCS) de´veloppe´ en formation initiale des sages-femmes. Dissertation for the completion of a diploma of the Ecole des cadres sages-femmes. Dijon: Ecole de Cadres Sages-Femmes 2008.
16. Gagnon, R., Charlin, B., Coletti, M., Sauvé, E., & van der Vleuten, C. Assessment in the context of uncertainty: How many members are needed on the panel of reference of a script concordance test? Medical Education, 2005; 39, 284-291
17. Grant J, Marsden P. Primary knowledge, medical education and consultant expertise. Med Educ 1988; 22:9–173 .
18. Kreiter CD, Bergus G. The validity of performance- based measures of clinical reasoning and alternative approaches. Med Educ 2009;43 (4):320–5
19. Monajemi A. Paradigm shift in clinical reasoning assessment. 3rd Asia Pacific Medical Education Conference, 2006, Singapore.
20. Newble D, Norman G, van der Vleuten C. Assessing clinical reasoning, in editors Higgs J & Jones M, Clinical reasoning in the health professions. Second edition, 1996.
21. Norman G, Bordage G, Curry L et al. Review of recentinnovations in assessment. In: Wakeford R, ed. Directions in Clinical Assessment. Report of the Cambridge Conference on the Assessment of Clinical Competence.Cambridge: Office of the Regius Professor of Physic, Cambridge University School of Clinical Medicine, Addenbrooks Hospital 1985;8–27
22. Norman GR. Theoretical and psychometric considerations. Report on the evaluation system for specialist certification. Task force of the evaluation committee.Ottawa, Ontario, Canada: Royal College of Phy-sicians and Surgeons of Canada. 1993.
23. Norman GR, Bordage G, Page G, Keane D. How specific is case specificity? Medical education 2006;40(7):618-623
24. Page G, Bordage G, Allen T. Developing key-feature problems and examinations to assess clinical deision-making skills. Acad Med 1995;70:194-201.
25. Page G, Bordage G. The medical council of Canada's key features project: A more valid written examination of clinical decision-making skills. Acad Med 1995;70:104-110.
26. Schmidt HG, Norman GR, Boshuizen HP. A cognitive perspective on medical expertise: Theory implications. Acad Med 1990;65:611-621
27. Sibert L, Charlin B, Corcos J, Gagnon R, Grise P, van der Vleuten C. Stability of clinical reasoning assessment results with the script concordance test across two different linguistic, cultural and learning environments. Med Teach 2002;24 (5):522–7
28. Sibert L, Darmoni SJ, Dahamna B, et al. Online clinical reasoning assessment with the script concordance test: a feasibility study. BMC Med Inform Decis Mak 2005;5:18.
29. Sibert L, Giorgi R, Dahamna B, Doucet J, Charlin B, Darmoni SJ. Is a web-based concordance test feasible to assess therapeutic decision-making skills in a French context? Med Teach 2009;31(4):162–8

30. Schuwirth, Lambert WT, and Cees PM van der Vleuten. «General overview of the theories used in assessment: AMEE Guide No. 57.» Medical teacher 33.10 (2011): 783-797.
31. Van der Vleuten C, Newble D, How can we test clinical reasoning? Lancet 1995 345:1032-1034.
۳۲. منجمی ع. استدلال بالینی: مفاهیم، آموزش و ارزیابی. انتشارات دانشگاه علوم پزشکی اصفهان، ۱۳۸۹، صفحات ۱۰۸.

فصل ۱۴

آزمون «ویژگی‌های کلیدی»

ساختار آزمون «ویژگی‌های کلیدی»

این آزمون بر پایهٔ این پیش‌فرض بنا شده است که در حل یک مسأله، همهٔ اطلاعات بیمار ارزش یکسانی ندارند، بلکه نکات کلیدی وجود دارد که اهمیت آن‌ها در حل مسأله بیش از سایر نشانه‌ها و علائم است و اشتباه در شناسایی آن‌ها باعث شکست در حل درست مسأله می‌شود. این نکات کلیدی همان‌طور که گفته شد، به صورت بسته‌های اطلاعاتی[1] در ذهن پزشکان ذخیره می‌شود که می‌توان آن را یکی از ویژگی‌های بارز پزشکان با تجربه در مقایسه با دانشجویان بر شمرد (منجمی ۱۳۸۹). لذا در این آزمون، توانایی آزمون‌شونده در انتخاب مرتبط‌ترین نشانه‌ها و علائم مربوط به یک سناریو اندازه گیری می‌شود.

پیج و بوردیج ۱۹۹۵

انجمن پزشکی کانادا[1] در سال ۱۹۸۶ در قالب یک پروژه تحقیقاتی و توسعه‌ای، اقدام به تدوین و طراحی یک ابزار جدید و معتبر جهت ارزیابی مهارت‌های استدلال بالینی در امتحانات تأیید صلاحیت در پزشکی[2] کرد. در آن زمان، امتحانات تأیید صلاحیت شامل سه دفترچه از سؤالات چندگزینه‌ای و یک دفترچه سؤالات PMP بود که به شکل دو روزه برگزار می‌شد. تمام دانشجویان قبل از اینکه به طبابت مشغول شوند، باید در این امتحان شرکت می‌کردند و قبول می‌شدند این پروژه به این دلیل انجام شد که اولاً بسیاری از مطالعات انجام شده استفاده از آزمون PMP را در ارزیابی مهارت‌های تصمیم‌گیری بالینی توصیه نمی‌کردند و ثانیاً تحقیقات انجام شده پیرامون خصوصات مهارت‌های استدلال بالینی در حکم راهنمایی برای طراحی ابزار جدید ارزیابی بودند. این پروژه در سه فاز موازی در شش سال از ۱۹۸۶ تا ۱۹۹۲ انجام گرفت. آزمون‌های پایلوت از طریق ارائه مجموعه‌ای از سؤالات «ویژگی‌های کلیدی» به دانشجویان پزشکی در زمان فارغ‌التحصیلی اجرا شد. بر اساس نتایج حاصل از پروژه، انجمن پزشکی کانادا در سال ۱۹۹۲، دفترچه آزمون PMP را با «آزمون ویژگی‌های کلیدی» جایگزین نمود.

1. Medical Council of Canada (MCC)
2. Qualifying Examination in Medicine

همان‌طور که در فصل اول این بخش بیان شد، یکی از ویژگی‌های متمایزکننده آزمون «ویژگی‌های کلیدی» از سایر آزمون‌های کتبی این است که در این آزمون‌ها می‌توان بیش از یک مورد را به عنوان پاسخ صحیح در سؤال در نظر گرفت. انعطاف‌پذیری در انتخاب بیش از یک پاسخ صحیح باعث می‌شود تا این نوع از آزمون‌ها در مقایسه با سایر آزمون‌های کتبی تک‌پاسخی از قبیل سؤالات چندگزینه‌ای و سؤالات «جورکردنی گسترده»، بیشتر منعکس کننده عملکرد واقعی فراگیران در محیط کار باشد.

به علاوه، آزمون «ویژگی‌های کلیدی» همانند آزمون PMP قادر است در مرحله بررسی نشانه‌ها و علایم بیماری، یا درخواست تست‌های تشخیصی و معاینات فیزیکی، توانایی حل مسأله فراگیران را ضمن ارائه سناریوی بالینی ارزیابی نماید. این ویژگی آزمون «ویژگی‌های کلیدی» مشابه با آزمون تشریحی تغییریافته است که به صورت متوالی[2] به ارزیابی

1. Page
2. Sequential format

مهارت‌های تصمیم‌گیری فراگیران می‌پردازد. آزمون ویژگی‌های کلیدی می‌تواند به شکل کاغذی یا الکترونیکی ارائه شود.

در این آزمون ابتدا یک سناریو مطرح می‌شود که معمولاً کوتاه است چرا که هدف اصلی جمع‌آوری اطلاعات بیشتر است. در ادامه بر اساس نوع آزمون ممکن است با توجه به سناریو تعدادی سؤال کوته‌پاسخ طرح شود که از فراگیر خواسته می‌شود خود پاسخ را ارائه نماید یا تعدادی گزینه طراحی می‌گردد و فراگیر از بین گزینه‌های ارائه شده پاسخ‌های احتمالی را مشخص می‌نماید. در صورت طراحی گزینه در این آزمون، به طور معمول ۲۰ گزینه ارائه می‌شود. در این شکل از سؤالات، آزمون‌شوندگان مجاز هستند که از میان ۲۰ گزینه حداکثر پنج مورد را انتخاب نمایند و اولویت انتخابی خود را نیز با دادن اعداد ۱ تا ۵ به گزینه‌ها مشخص نمایند. با اینکه دانشجویان تنها پنج انتخاب مجاز دارند، تعداد جواب‌های درست احتمالی، می‌تواند بیش از پنج مورد باشد. به عنوان یک قاعده کلی، تعداد کل گزینه‌ها به طور معمول چهار برابر حداکثر تعداد انتخاب‌های مجاز است.

فارمر و هینچی[1] ۲۰۰۵

نویسندگان در این پژوهش به بررسی آزمون «ویژگی‌های کلیدی» امتحانات کالج سلطنتی پزشکان عمومی استرالیا[2] برای دوره فلوشیپ پرداختند. این آزمون در سال ۱۹۹۷ به عنوان جایگزین آزمون PMP اجرا شد. اکثر سؤالات توسط تیمی از ارزیابان که پزشکان عمومی و یا فلوشیپ بورد امتحانات دانشگاه رویال استرالیا بودند طراحی شد. از طراحان سؤالات خواسته شده بود که مشکلات بالینی را طرح کنند که منعکس کننده چالش‌هایی است که همه روزه فراگیران با آن مواجه می‌شوند. تعدادی از مشکلات بالینی با تصاویری از قبیل الکتروکاردیوگرافی، رادیوگرافی و ... ارائه می‌شد. فرایند تهیه کلید پاسخ نیز همزمان با مرحله طراحی سؤالات صورت گرفت. بعد از طراحی و مرور، نمونه سؤالات مناسب در بانک سؤال[3] ذخیره شد. دو معیار برای انتخاب سؤالات از بانک سؤال در نظر گرفته شد. اولا، موارد بالینی نباید بیشتر از یک بار در هر دو سال استفاده می‌شدند و ثانیاً انتخاب بر اساس میزان بروز آن مشکل در بالین صورت می‌گرفت. سؤالات نوشتاری توسط یک ارزیاب و سؤالات نوع گزینه‌ای به وسیله کامپیوتر تصحیح می‌شد. ارزیابان همچنین پاسخ‌های نادرست غیرقابل‌انتظار را یادداشت می‌نمودند و از آنان در فرایند طراحی گزینه‌های انحرافی و تهیه کلید پاسخ استفاده می‌کردند.

1. Hinchy
2. Royal Australian College of General Practitioners Examination
3. Item bank

مزایا و محدودیت‌های آزمون «ویژگی‌های کلیدی»

مزایای آزمون «ویژگی‌های کلیدی»

- **ارزیابی توانایی تصمیم‌گیری بالینی:** آزمون «ویژگی‌های کلیدی» روشی معتبر برای ارزیابی شایستگی‌های تصمیم‌گیری بالینی است. سناریو در آزمون «ویژگی‌های کلیدی» باید به نحوی طراحی گردد که بتوان از طریق آن توانایی حل مسأله و استدلال بالینی آزمون‌شوندگان را به جای ارزیابی صرف حقایق سنجید. هر چند داشتن اطلاعات و حقایق پایه در حکم پیش‌نیازی در فرایند حل مسأله محسوب می‌شود، در آزمون «ویژگی‌های کلیدی» کاربرد این حقایق در موقعیت بالینی ارائه‌شده مورد تاکید قرار می‌گیرد.

هرتز و همکاران[1] ۲۰۱۲

پژوهشگران مطالعه‌ای با هدف بررسی اینکه آیا آزمون «ویژگی‌های کلیدی» فرایندهای شناختی سطح بالا را می‌سنجد، انجام دادند. تحلیل داده‌های عینی[2] (طول سؤالات و سطح دشواری آن‌ها، عملکرد سؤالات، مدت زمان پاسخگویی) و داده‌های ذهنی[3] (نمره‌دهی متخصصان از سطح شناختی سؤالات) نشان داد که آزمون «ویژگی‌های کلیدی» قابلیت ارزیابی سطوح بالاتر حیطه شناختی را در مقایسه با سؤالات چندگزینه‌ای متداول دارد.سطح دشواری آزمون «ویژگی‌های کلیدی» نه تنها پیچیده‌تر بود، بلکه این پیچیدگی منجر شده بود تا زمان پاسخ‌دهی به برخی سؤالات افزایش یابد.
نتایج حاصل از مطالعه دال بر تأثیر استفاده از آزمون «ویژگی‌های کلیدی» برای ارزیابی سطوح بالای حیطه شناختی از قبیل تصمیم‌گیری بالینی است.

1. Hurtz et al.
2. Objective
3. Subjective

- **امکان پوشش وسیع‌تری از موارد بالینی در مقایسه با سایر آزمون‌های استدلال بالینی:** تمرکز این سؤالات بر سنجش ویژگی‌های کلیدی منجر شده است تا تعداد بیشتری از این سؤالات را در هر بار اجرای آزمون بتوان ارائه نمود.
- **امکان طراحی شکل‌های متنوعی از پاسخ:** یکی از ویژگی‌های منحصر به فرد آزمون «ویژگی‌های کلیدی» این است که در این شکل از سؤالات، گزینه‌های پاسخ می‌تواند در قالب‌های متفاوتی از علایم و نشانه‌های بیماری تا روند تشخیص و درمان طرح گردد. به علاوه، در این آزمون امکان طرح سؤالات به اشکال مختلف بازپاسخ یا بسته‌پاسخ و به روش کاغذی یا الکترونیکی فراهم است.

شاپر و همکاران[1] ۲۰۱۳

این پژوهشگران از سؤالات «ویژگی‌های کلیدی» در ارزیابی نهایی دانشجویان دامپزشکی استفاده کردند. سؤالات توسط تیمی از متخصصان طراحی شد و سپس توسط کمیته طراحی سؤالات مرور و بررسی شد. در فاز ابتدایی مطالعه، آزمون به شکل تکوینی و پایلوت اجرا شد که هیچ تاثیری روی قبولی یا ردی دانشجویان نداشت. در این مرحله، ۵۴ دانشجوی دامپزشکی که دوره بالینی خود را گذارنده بودند (ترم ۹ و ۱۰) و همچنین ۱۱ دانشجوی ترم ۶ تا ۸ نیز که در کلاس‌های اختیاری ویروس‌شناسی حضور داشتند، در این مطالعه شرکت کردند. در مرحله بعد یعنی پژوهش اصلی بین ۲۲۵ تا ۲۴۴ دانشجو دامپزشکی در آزمون نهایی ویژگی‌های کلیدی شرکت داشتند. از دانشجویان درخواست شد تا به سؤالات «ویژگی‌های کلیدی» که از طریق سیستم الکترونیکی ارائه می‌شد، پاسخ دهند. تنها در صورت پاسخگویی به هر سؤال، امکان پاسخگویی به سؤال بعدی وجود داشت و پاسخ صحیح سؤال قبلی نمایش داده می‌شد و امکان بازگشت جهت تغییر پاسخ قبلی وجود نداشت. همه پاسخ‌های فراگیران در سیستم ثبت و سپس آنالیز می‌شد. مزایا و معایب حاصل از آزمون «ویژگی‌های کلیدی» از طریق بحث و گفتگو در قالب گروه‌های متمرکز[2] با دانشجویان تعیین شد و دو نفر محقق به صورت مستقل از یکدیگر متن پیاده‌سازی‌شده بحث متمرکز را تحلیل و کدبندی نمودند. همچنین به منظور تحلیل نتایج آزمون از ضریب دشواری و آلفای کرونباخ استفاده شد. بر اساس نتایج حاصله هم فراگیران و هم استادان نظر مثبتی نسبت به آزمون «ویژگی‌های کلیدی» داشتند. میانگین ضریب دشواری بین ۷۰/۵۴ درصد و ۷۷/۳۶ درصد بود. آلفای کرونباخ کل آزمون بین ۰/۶۷ تا ۰/۸۰ بود. بر اساس نتایج حاصل محققان پیشنهاد کردند که آزمون «ویژگی‌های کلیدی» الکترونیکی یک ابزار ارزیابی مفید است.

1. Schaper et al
2. Focus group

- **پایایی بالا:** از آنجا که در این آزمون‌ها امکان طرح تعداد بیشتری سؤال در یک آزمون میسر است، این موضوع به پایایی بالاتر این آزمون‌ها می‌انجامد (ترودل و همکاران[1] ۲۰۰۸). به‌علاوه مشخص بودن نحوه نمره‌دهی به این سؤالات به بهبود پایایی بین نمره‌دهی ارزیابان مختلف منجر می‌شود. به عبارت دیگر، این آزمون دارای نمره‌گذاری عینی‌تری در مقایسه با سایر آزمون‌های استدلال بالینی است.
- **قابل دفاع بودن تصمیم‌گیری‌های مربوط به رد یا قبولی در این آزمون:** بر خلاف سؤالات بازپاسخ به ویژه نوع تشریحی، در این سؤالات ارائه گزینه‌های پاسخ احتمالی به تسهیل فرایند نمره‌دهی این سؤالات می‌انجامد، به نحوی که بر اساس نتایج حاصل از ارزیابی به راحتی می‌توان در مورد رد یا قبولی آزمون‌شوندگان قضاوت نمود. در واقع در این آزمون، تعریف دقیق و مشخص پاسخ‌های صحیح احتمالی باعث شده است تا تصمیم‌گیری در خصوص عملکرد داوطلبان راحت‌تر صورت گیرد و حتی این امر به بهبود پایایی بین نمرات مصححان نیز می‌انجامد.
- **امکان ارزیابی ساختارمندتر و متمرکزتری از موارد بالینی:** سناریوی بالینی در این سؤالات بر سنجش مهارت‌های تصمیم‌گیری و استدلال در یک موقعیت خاص بالینی متمرکز می‌شود و به این طریق می‌توان افتراق خوبی میان افراد تازه‌کار و با تجربه به دست آورد.

محدودیت‌های «آزمون ویژگی‌های کلیدی»

- **دشواری طراحی:** فرایند طراحی این سؤالات معمولاً زمان‌بر است. به ویژه استادان کم تجربه به مدت زمان بیشتری برای طراحی این سؤالات نیاز دارند. به علاوه، عدم آشنایی استادان با این سؤالات، نیازمند برگزاری کارگاه‌های آموزشی است که مستلزم صرف وقت و هزینه است. زمان‌بر بودن فرایند طراحی این سؤالات و از طرف دیگر نیاز به طرح تعداد زیادی سؤال

1. Trudel et al.

برای رسیدن به سطح قابل قبول پایایی، از محدودیت‌های این ابزارها به شمار می‌رود (چارلین و همکاران ۲۰۰۲).

- **عدم پذیرش فراگیران و استادان:** یکی از سؤالات پیش روی هر ابزار ارزیابی جدید همواره این است که این ابزار تا چه حد از سوی فراگیران و استادان مورد پذیرش قرار خواهد گرفت. آزمون «ویژگی‌های کلیدی» نیز از این قاعده مستثنی نیست. عدم آشنایی فراگیران با این آزمون می‌تواند به ایجاد اضطراب و سردرگمی در جلسه امتحان منجر شود که این موضوع بر نتایج حاصل از امتحان تاثیرگذار خواهد بود. به همین دلیل توصیه می‌شود که ابتدا این آزمون به شکل تکوینی و در طول‌ترم برگزار گردد و پس از حصول اطمینان از آشنایی فراگیران از آن در امتحانات نهایی استفاده شود. همچنین طراحی یک دستورالعمل واضح در آزمون به آشناسازی بیشتر کمک خواهد نمود. آشنا نمودن استادان با این شکل از آزمون‌ها نیز از طریق برگزاری کارگاه‌های آموزشی، تهیه کتابچه‌ها و راهنماهای طراحی سؤالات امکان‌پذیر است.

گام‌های طراحی آزمون «ویژگی‌های کلیدی»

در ادامه بر اساس راهنماهای موجود، کل فرایند طراحی آزمون «ویژگی‌های کلیدی» در هشت مرحله با ذکر اصول و مثال‌های مرتبط توضیح داده می‌شود. به طور خلاصه، گام‌های زیر در طراحی آزمون «ویژگی‌های کلیدی» برداشته می‌شود (جدول ۱-۱۴):

جدول ۱-۱٤: خلاصه مراحل طراحی آزمون «ویژگی‌های کلیدی

ردیف	عنوان	توضیح
۱	انتخاب یک مشکل بالینی مناسب	هنگام انتخاب موضوع، روی یک مشکل مهم بالینی باید متمرکز شد. برای این کار می‌توان فهرستی از مشکلات بالینی مهمی که فراگیران در آینده با آن‌ها مواجه خواهند شد، تهیه کرد.
۲	انتخاب محور مناسب	محور موقعیتی است که تعیین می‌کند چه اطلاعاتی (تفسیر داده‌ها، تشخیص، مراقبت از بیمار) باید در سؤال ارائه شود.
۳	تعیین ویژگی‌های کلیدی مشکل بالینی	انتخاب و تعریف ویژگی‌های کلیدی برای هر تابلوی بیماری، یک مرحله اساسی در طراحی سؤالات مناسب به شمار می‌رود. بهترین کار برای تقویت و ارتقاء این فرایند، ارائه و مرور ویژگی‌های کلیدی تعیین شده توسط سایر همکاران یا اعضای کمیته ارزیابی دانشکده است.
۴	نگارش سناریو (تابلوی بیماری)	شکل ارائه سناریو در آزمون ویژگی‌های کلیدی باید به صورتی باشد که آزمون‌شونده را به تشخیص ویژگی‌های کلیدی در یک بیماری خاص ترغیب نماید تا بتواند روش‌های تشخیص و درمانی مناسب را برای آن انتخاب کند.
۵	نگارش سؤال هدایت‌کننده	سؤال هدایت‌کننده در این آزمون مستقیماً از ویژگی‌های کلیدی گرفته می‌شود و باید منحصراً بر یک عملکرد یا اقدام بالینی (مداخلات درمانی، تشخیص‌های افتراقی، علایم و نشانه‌های بیماری، تست‌های آزمایشگاهی و ...) متمرکز باشد.
۶	تصمیم‌گیری در مورد نوع سؤال (بسته‌پاسخ یا بازپاسخ)	تصمیم‌گیری در مورد نوع سؤال در آزمون ویژگی‌های کلیدی معمولاً با توجه به محور انتخابی ارزیابی (شرح حال‌گیری و علایم و نشانه‌های بالینی..) و با در نظر گرفتن شرایط تصحیح آن به شکل دستی یا با کمک رایانه صورت می‌گیرد.
۷	تصمیم‌گیری در خصوص نحوه نمره‌دهی	در آزمون ویژگی‌های کلیدی با توجه به تعداد زیاد سؤال و طبعاً گزینه‌های متعدد صحیح، ضروری است که تعداد پاسخ‌های درست مجاز و بارم آن‌ها به وضوح تعیین شود. نمره‌دهی این آزمون‌ها به دو بخش تقسیم می‌شود: نمره‌دهی به هر سؤال و نمره‌دهی به کل آزمون.
۸	تدوین دستورالعمل آزمون	اگر این شیوه ارزیابی برای بار اول مورد استفاده قرار می‌گیرد و دانشجویان با آن آشنایی ندارند، تدوین دستورالعمل آزمون ضروری است. دستورالعمل آزمون باید واضح و روشن تنظیم گردد.
۹	مرور سؤال	

انتخاب یک مشکل بالینی مناسب

اولین نکته‌ای که باید در طراحی لحاظ شود، مشخص کردن شرحنامه‌ای است که به دنبال یافتن ویژگی‌های کلیدی آن هستیم. بنابراین در ابتدا یک تابلوی بالینی در نظر گرفته می‌شود. باید توجه شود که هنگام انتخاب موضوع، روی یک مشکل مهم بالینی تمرکز شود. به عنوان مثال، تابلوی سرفه مزمن در یک مرد میانسال و یا اسهال در یک نوزاد.

بهتر است برای این کار فهرستی از مشکلات بالینی مهمی که فراگیران در آینده با آن‌ها مواجه خواهند شد، تهیه گردد. در تهیه این فهرست می‌توان از تجاربی که خود استادان در مواجهه با موارد بالینی در طی سالیان مختلف داشته‌اند، نیز استفاده کرد. مراجعه به بلوپرینت نیز در این مرحله بسیار کمک‌کننده خواهد بود. به این ترتیب با کمک گرفتن از فهرست موارد بالینی و بلوپرینت می‌توان سؤالاتی را طراحی نمود که از لحاظ بالینی مهم و مرتبط با آینده حرفه‌ای فراگیران است.

علاوه بر انتخاب موارد بالینی مهم، نکته‌ی دیگری که باید هنگام طراحی سؤال به آن توجه داشت، تناسب سطح دشواری سؤال با ویژگی‌های آزمون‌شوندگان و سطح اطلاعاتی فراگیران است. به عنوان مثال، در صورتی که قرار است سؤالی برای دانشجویان فیزیوپاتولوژی طراحی شود، بهتر است تمرکز سؤال بیشتر بر ارزیابی علل احتمالی بروز بیماری باشد در حالی که برای ارزیابی دانشجویان سال‌های بالاتر می‌توان مواردی از قبیل تشخیص‌های افتراقی و برنامه درمانی را مطرح نمود.

هنگام در نظر گرفتن مشکل بالینی، مهم است که گروه سنی مورد بالینی، موقعیت بالینی و محل ارائه مراقبت مشخص شود:

- تعداد زیادی از مشکلات در دوره‌های خاصی از زندگی بیمار اتفاق می‌افتند که تظاهرات آنها بستگی به سن بیمار دارد. گروه‌های سنی که همواره باید هنگام طرح مشکل بالینی در ذهن داشت، شامل موارد زیر هستند: دوران بارداری، نوزادی، خردسالی، نوجوانی، جوانی، میانسالی وکهنسالی.
- مشکلات بیماری مطرح شده می‌توانند به صورت موقعیت‌های بالینی تظاهر پیدا کنند: شکایت‌های غیر اختصاصی و تمایز نیافته[1]، یک مشکل معمول، مشکلات متعدد و چند سیستمی، مشکلات تهدیدکننده زندگی[2] و پیشگیری و ارتقاء سلامت. توجه داشته باشید در هر سؤال تنها یک موقعیت بالینی یا در برخی مواقع نهایتاً دو موقعیت بالینی می‌تواند مورد ارزیابی قرار گیرد. موقعیت بالینی در توصیف ویژگی‌های کلیدی مشکل بالینی تاثیرگذار است. در سطح دوره پزشکی عمومی، توصیه می‌شود تظاهرات متداول و معمول مشکلات بالینی انتخاب شوند. زمانی که تظاهرات بالینی غیرمعمول در طراحی سؤال استفاده شوند، تقریباً تمام داوطلبان در حل این سؤالات ناموفق خواهند بود و سؤال در این شرایط قدرت تمیز و افتراق بین دانشجویان قوی و ضعیف را نخواهد داشت.
- محل ارائه خدمات مراقبتی نیز بر عملکرد و تصمیم‌گیری‌های بالینی مورد انتظار تاثیرگذار است، به این دلیل که در هر شرایط منابع مختلفی در دسترس است. به عنوان مثال، در مورد درد حاد قفسه سینه، درخواست مشاوره قلبی فوری یا انجام آزمایش‌های اختصاصی ممکن است در یک درمانگاه امکان‌پذیر نباشد اما در اورژانس منطقی به نظر برسد. محل‌های ارائه خدمات می‌توانند شامل موارد زیر باشند: درمانگاه سرپایی، اورژانس، بخش بستری و اتاق عمل.

توجه داشته باشید که تنها زمانی که مشکل بالینی (به عنوان مثال، درد شکم)، گروه هدف بیماران (به عنوان مثال، میانسالی)، موقعیت بالینی (مشکلات تهدیدکننده زندگی) و محل ارائه خدمات (درمانگاه سرپایی) مشخص و تعریف شده باشند، می‌توان به تعیین ویژگی‌های کلیدی پرداخت.

انتخاب محور مناسب برای سؤال

محور سؤال تعیین‌کننده دامنه‌ای از موضوع است که قصد ارزیابی آن وجود دارد. اهمیت تصمیم در خصوص محور مناسب سؤال، در طراحی گزینه‌های سؤال و نحوه ارائه اطلاعات در پایه سؤال مشخص می‌شود. محور در سؤالات

1. Undifferentiated complaint
2. Life-threatening event

ویژگی‌های کلیدی می‌تواند بررسی درخواست تست‌های آزمایشگاهی، معاینه فیزیکی و شرح حال‌گیری یا توجه به علایم و نشانه‌های بالینی و ... باشد. به دو مثال زیر توجه کنید:

موضوع: درد قفسه سینه	محور: بررسی یافته‌های حاصل از تست‌های آزمایشگاهی	گروه هدف: میانسالی

آقای ۵۰ ساله‌ای با سابقه درد شدید در قفسه سینه، تهوع و تعریق مراجعه کرده است. در سمع قلب تاکیکارد است. شما به کدام یک از موارد ذیل برای تشخیص این بیمار نیاز دارید؟ ۵ مورد را انتخاب کنید.

۱.	آزمایش کامل خون	۷.	پلاکت خون
۲.	اسید اوریک سرم	۸.	الکترولیت‌های خونی
۳.	الکتروکاردیوگرافی	۹.	کراتینین
۴.	اکوکاردیوگرافی	۱۰.	تست‌های انعقادی
۵.	آلبومین سرم		

موضوع: درد قفسه سینه	محور: علایم و نشانه‌های بالینی	گروه هدف: میانسالی

آقای ۵۰ ساله‌ای به دنبال فعالیت دچار درد شدید در قفسه سینه شده و به بخش اورژانس بیمارستان مراجعه کرده است. پزشک در بررسی‌های اولیه خود به بروز حمله قلبی شک نموده است. در صورت وجود چه علایم دیگری در این بیمار تشخیص قطعی‌تر خواهد شد؟ ۵ مورد را انتخاب کنید.

۱.	آدم اندام تحتانی	۸.	درد فک
۲.	استفراغ	۹.	سرگیجه
۳.	افزایش ضربان قلب	۱۰.	سردرد
۴.	افزایش درجه حرارت	۱۱.	سوء هاضمه
۵.	تعریق	۱۲.	سیانوز
۶.	تنگی نفس	۱۳.	فشار خون بالا
۷.	خواب آلودگ	۱۴.	نبض نخی شکل

همان‌طور که مشاهده می‌کنید، در مثال‌های بالا، هر دو سؤال یک موضوع واحد یعنی تشخیص درد قفسه سینه را می‌سنجند اما محور ارزیابی متفاوت است. همین تغییر در محور سؤال باعث شده است که اطلاعاتی که در دو سؤال ارائه شده است، کاملاً متفاوت باشد.

به طور کلی، اگر سناریویی طراحی شود، بدون اینکه دقیقاً هدف طراح سؤال مشخص باشد پاسخگویی به سؤال مشخص دشوار خواهد بود. برای طراح سؤال باید کاملاً روشن شود که پرسش در مورد شرحنامه چه بیماری یا چه علامتی است. به عنوان مثال، اگر در سناریوی یک علامت بالینی، اطلاعات غیرضروری به سناریو اضافه شود بدون آنکه به تشخیص کمک کند، این سؤال منجر به سردرگمی دانشجویان می‌شود.

تعیین ویژگی‌های کلیدی برای مورد بالینی

انتخاب و تعریف ویژگی‌های کلیدی برای تابلوی بیماری، یک مرحله اساسی در طراحی سؤالات مناسب به شمار می‌رود. این قسمت در متن سؤالی که نهایتاً در اختیار دانشجو قرار می‌گیرد، ذکر نمی‌شود اما برای روشن شدن ذهن طراح سؤال و شفاف شدن مسیر پیش رو کاملاً ضروری است.

ویژگی‌های کلیدی به طور معمول شامل موارد زیر هستند:

- اخذ شرح‌حال یا تاریخچه
- جستجو و تفسیر علائم و نشانه‌های بالینی
- تدوین فهرست تشخیصی یا ارائه تشخیص‌های افتراقی
- درخواست بررسی‌هایی برای تأیید یا رد تشخیص‌های افتراقی

- تعیین اهداف درمانی یا تصمیمات بالینی
- تجویز دارو
- پیگیری روند درمان

به طور کلی، هر ویژگی کلیدی از سه بخش تشکیل شده است. همان‌طور که مشاهده خواهید کرد، هر ویژگی کلیدی باید حتماً دو بخش ابتدایی را داشته باشد، در حالی که عنصر سوم لزوماً همیشه وجود ندارد.

- **مجموعه‌ای از شرایط:** اطلاعات بالینی اولیه[1] مشخص کننده مشکل، سن بیمار، موقعیت بالینی است. اطلاعات بالینی اولیه می‌توانند شامل علایم و نشانه‌ها (به عنوان مثال، بیماری با پای متورم و دردناک) و یافته‌های آزمایشگاهی باشند. انتخاب این اطلاعات بستگی به ماهیت ویژگی‌های کلیدی مورد سنجش دارد. مثال زیر نمونه‌ای از ارائه اطلاعات بالینی اولیه است.

نمونه‌هایی از ارائه مجموعه شرایط در یک تابلوی بیماری

۱. خانم باردار سه ماهه با خونریزی واژینال بدون درد شکمی
۲. مردی میانسال با سابقه تنگی نفس شدید و ادم اندام تحتانی و سطح کراتنین بالا
۳. آقای ۲۵ ساله با درد شدید ناگهانی در سمت راست قفسه سینه و ناحیه تحتانی ریه

عملکرد یا تصمیم بالینی: شامل یک وظیفه بالینی، تصمیم و گام ضروری در تشخیص و درمان تابلوی بالینی است که انتظار می‌رود فراگیر توانایی انجام آن را داشته باشد (به عنوان مثال، درخواست انجام ونوگرام در افتراق سایر تشخیص‌های بالینی در ترومبوز وریدهای عمقی).

نمونه‌هایی از عملکرد یا تصمیم بالینی مورد انتظار در سؤال «ویژگی‌های کلیدی»

۱. توجه به وجود جفت سر راهی به عنوان یک تشخیص محتمل
۲. توجه به علایم نارسایی حاد کلیه
۳. توجه به تشخیص های افتراقی اندوکاردیت قلبی

- **کلمات توصیفی:** در صورت نیاز می‌توان از کلمات توصیفی برای شرح بیشتر ویژگی‌های کلیدی استفاده کرد (به عنوان مثال، محتمل‌ترین تشخیص ممکن). این کلمات توصیفی منعکس‌کننده فوری بودن اتخاذ یک تصمیم (به عنوان مثال، اولین اقدام) و اولویت‌های تصمیم‌گیری (به عنوان مثال مهم‌ترین اقدام) است.

نمونه‌ایی از کلمات توصیفی در سؤال «ویژگی‌های کلیدی»

۱. درخواست اولین اقدام ...
۲. درخواست آزمایش‌های فوری
۳. فهرست کردن محتمل‌ترین تشخیص‌های احتمالی
۴. مهم‌ترین اقدام درمانی

تجربه نشان داده است که معمولاً دو تا سه ویژگی کلیدی در هر مورد بالینی قابل طرح است یعنی برای یک مشکل بالینی می‌توان دو یا سه سؤال طراحی کرد. هر چند در این میان، تعدادی از مشکلات بالینی ممکن است تنها یک اقدام حیاتی و برخی دیگر بیش از پنج اقدام را شامل شوند. با این حال از نظر ویژگی‌های سایکومتریک، در نظر گرفتن دو تا سه ویژگی کلیدی و در نتیجه طرح دو یا سه سؤال برای هر مورد بالینی به بهبود پایایی آزمون می‌انجامد (نورمن‌و همکاران ۲۰۰۶).

1. Initial clinical information

البته بسیار مهم است که هر سؤال تنها بر یک ویژگی کلیدی متمرکز شود. آنچه به شفاف شدن موضوع کمک می‌کند، طراحی جدولی است که به صورت خلاصه نحوه ارتباط سؤالات آزمون و ویژگی‌های کلیدی را نشان می‌دهد. به عنوان مثال در مورد بالینی زیر شامل سه سؤال است که به ارزیابی چهار ویژگی‌های کلیدی می‌پردازد.

	ویژگی کلیدی ۱	ویژگی کلیدی ۲	ویژگی کلیدی ۳	ویژگی کلیدی ۴
سؤال یک	×			
سؤال دو		×	×	
سؤال سه				×

به طور کلی در این مرحله باید به نکات زیر توجه داشت:

- در صورتی که مشکلات بالینی بسیار وسیع انتخاب شده باشند، فرایند تعریف ویژگی‌های کلیدی، می‌تواند بسیار خسته‌کننده باشد. به عنوان مثال، تعریف ویژگی‌های کلیدی در اختلال ایسکمیک قلبی–عروقی[1] و سوءمصرف الکل[2] بسیار دشوار خواهد بود. در حالی که با تمرکز بر یک تابلوی بیماری ویژه به عنوان مثال آنژین پایدار و ناپایدار[3] یا سندرم ترک الکل[4] می‌توان این مشکل را بر طرف نمود. در مقابل، باید توجه داشت که انتخاب یک بیماری نادر به عنوان پایه‌ای برای تعریف ویژگی‌های کلیدی، باعث ایجاد دید محدود در ذهن می‌شود و بدین ترتیب تعریف ویژگی‌های کلیدی دشوار خواهد شد.
- باید در نظر داشت که از تعریف اقدامات بالینی به صورت مبهم و کلی از قبیل «درمان مناسب» یا «درخواست انجام آزمایش مناسب» خودداری شود. اقدامات و عملکردهای مورد انتظار در این مرحله باید تا حد امکان به صورت ویژه تعیین شوند، به عنوان مثال، «تجویز مورفین» یا «درخواست تعیین گازهای خون شریانی» مناسب هستند. همچنین از ترکیب چندین اقدام بالینی به صورت همزمان و در قالب یک گزینه باید اجتناب کرد. به عنوان مثال «در نظر گرفتن احتمال بروز ترمبوز وریدهای عمقی در فهرست تشخیص‌های افتراقی و درخواست انجام ونوگرام» باید به صورت دو ویژگی کلیدی مجزا بیان شود.
- در مواقعی می‌توان ویژگی کلیدی را به صورت یک عبارت اجتنابی مطرح کرد که فراگیر باید بداند از انجام آن کار باید خودداری شود. به عنوان مثال ممکن است در فهرست گزینه‌ها به عنوان ویژگی کلیدی، تابلوی بالینی خونریزی واژینال، اجتناب از معاینات لگنی (به دلیل احتمال بروز خونریزی کشنده متعاقب انجام آن) یا ممانعت از ترخیص بیمار از بخش اورژانس ذکر گردد.
- در تعریف ویژگی‌های کلیدی باید همواره سطح تجربه و آموزش داوطلبان را در نظر داشت. مسائل بالینی بسیار مشابه می‌توانند ویژگی‌های کلیدی متفاوتی را شامل شوند. انتخاب این ویژگی‌ها بستگی به این دارد که قصد ارزیابی کارآموز، کارورز یا حتی دستیار را دارید.
- کار برای تقویت و ارتقاء این فرایند، ارائه و مرور ویژگی‌های کلیدی تعیین شده توسط سایر همکاران یا اعضاء کمیته ارزیابی دانشکده است. اهمیت این موضوع در مواردی که طراح سؤال از تجربه کافی برای طراحی این‌گونه سؤالات برخوردار نیست، دو چندان می‌شود. مطالعات نشان داده‌اند که اعضای کمیته ارزیابی در صورتی که با سطح و ویژگی داوطلبان آشنا باشند، می‌توانند ویژگی‌های کلیدی بسیار مناسبی را تعریف کنند (بوردیج و همکاران ۱۹۹۵).

1. Cardiovascular ischemic disorder
2. Alcohol abuse
3. Stable and unstable angina
4. Alcohol withdrawal

نگارش سناریو (تابلوی بیماری)

ارائه سناریو در آزمون «ویژگی‌های کلیدی»، موجب تقویت توانایی داوطلب در حل مسائل بالینی می‌شود. همواره باید توجه داشت که موارد بالینی انتخابی نباید به شکلی ارائه گردند که‌صرفاً توانایی آزمون‌شونده را در توصیف ویژگی‌های مربوط به یک بیماری بسنجند (ارزیابی حقایق، داده‌ها و اطلاعات)، بلکه شکل ارائه سناریو در آزمون ویژگی‌های کلیدی باید به صورتی باشد که آزمون‌شونده را به تشخیص این ویژگی‌ها در یک بیمار خاص ترغیب نماید (ارزیابی استدلال بالینی) تا بتواند روش‌های تشخیص و درمانی مناسب را برای آن انتخاب نمود (ارزیابی تصمیم‌گیری بالینی). از این رو سناریوهای بالینی باید بر اساس ویژگی‌های کلیدی مهم طرح شوند.

مثال زیر، نمونه‌ای از سناریوی بالینی است که تنها به ارزیابی صرف حقایق می‌پردازد. داوطلب در این سؤال بدون مطالعه سناریو هم می‌تواند به سؤالات طرح شده پاسخ دهد.

سناریوی ضعیف

یک خانم ۲۰ ساله در اولین حاملگی خود مبتلا به دیابت نوع ۱ شده است. ایشان بعد از مراجعه به مطب شما راجع به مشکلات دیابت در دوران بارداری صحبت می‌کند.
عوارض مادری ناشی از دیابت نوع ۱ در دوران بارداری چه مواردی است؟ (چهار مورد را بیان کنید)
عوارض جنینی ناشی از دیابت نوع ۱ در دوران بارداری چه مواردی است؟ (دو مورد را بیان کنید)

هنگام نوشتن سناریو، مهم است که به ویژگی‌های کلیدی که در مرحله قبل استخراج شده رجوع شود. در غیر این صورت سناریو به بیراهه می‌رود و احتمال اینکه به جای استدلال بالینی، صرفاً اطلاعات سطحی دانشجو سنجیده شود زیاد می‌شود. مواردی که در توضیح بیمار می‌توان استفاده کرد شامل موارد زیر است:

- سن و جنس (خانم ۴۵ ساله)
- محل مراجعه (به درمانگاه مراجعه کرده است)
- شکایت اصلی[1] (مبتلا به درد شکم)
- مدت بیماری (از سه روز قبل)
- سابقه بیمار (با سابقه خانوادگی فشارخون)
- یافته‌های معاینه (در لمس، ربع فوقانی شکم دردناک است)
- نتیجه تست‌های آزمایشگاهی (افزایش سطح گلبول‌های سفید)

موضوع: ادم **محور: بررسی یافته‌های حاصل از تست‌های آزمایشگاهی و شرح‌حال** **گروه هدف: خانم باردار**

خانم ۲۸ ساله‌ای در هفته ۳۳ حاملگی با آدم ۲+ مراجعه کرده است. در معاینات ایکتر دارد، سمع قلب تاکیکارد است. شما به کدام یک از موارد ذیل برای تشخیص این بیمار نیاز دارید؟ ۵ مورد را انتخاب کنید.

۱. آزمایش کامل خون
۲. اسید اوریک سرم
۳. آلبومین سرم
۴. آنالیز ادرار
۵. الکترولیت‌های خونی
۶. آنالیز مایع آمنیون
۷. اندازه‌گیری پلاکت خون
۸. آندوسکوپی فوقانی
۹. آنزیم‌های کبدی
۱۰. اندازه‌گیری کراتینین
۱۱. تست‌های انعقادی
۱۲. سابقه پره اکلاپسی
۱۳. سطح هوشیاری بیمار فشار خون

1. Chief complaint

در این مثال، همان‌طور که می‌بینید شرح‌حال کوتاهی از بیمار ارائه شده است. در این شرح‌حال هدف، پیدا کردن نکات کلیدی مربوط به حاملگی و ایکتر است.

طول سناریو در سؤالات مختلف می‌تواند متفاوت باشد. این موضوع بستگی به این دارد که چه مقدار اطلاعات مورد نیاز است تا بتوان به ارزیابی ویژگی‌های کلیدی در هر تابلوی بالینی پرداخت. در مسائل بالینی که ویژگی‌های کلیدی بر تشخیص مشکل بیمار متمرکز هستند، سناریو در اغلب موارد به صورت بسیار مختصر بیان می‌شود. در مسائل بالینی که ویژگی‌های کلیدی بر روند درمان یا تست‌های آزمایشگاهی متمرکز می‌شوند، سناریو معمولاً طولانی‌تر خواهد بود و شامل اطلاعات بالینی حاصل از شرح‌حال‌گیری، معاینه فیزیکی یا تست‌های آزمایشگاهی است.

برای طراحی سناریو، استفاده از زبان غیرتخصصی به استفاده از زبان تخصصی و اصطلاحات پزشکی ترجیح داده می‌شود. علت این امر این است که در شرایط واقعی زندگی، اطلاعاتی که در دسترس داوطلب قرار داده می‌شود، عبارات و کلمات غیرتخصصی مانند سرفه خون آلود[1] است، نه اصطلاحات پزشکی از قبیل هموپتزی[2]. توجه داشت باشید زمانی که سناریوی بالینی با استفاده از اصطلاحات پزشکی (به عنوان مثال، علامت هومن[3]) ارائه می‌شود، داوطلبان به ویژه آن‌هایی که ضعیف‌تر هستند، با احتمال بیشتری می‌توانند پاسخ صحیح را تشخیص بدهند (اوا و همکاران[4] ۲۰۱۰). به عبارت دیگر، استفاده از اصطلاحات پزشکی به نوعی در حکم سرنخی برای رسیدن به پاسخ است و به این طریق قدرت تمیز آزمون کاهش می‌یابد. استفاده از توصیفات غیرتخصصی (از قبیل درد ساق پا) برای داوطلبان ضعیف چالش‌برانگیزتر است و در این حالت قدرت تمیز ویژگی‌های کلیدی مورد ارزیابی افزایش می‌یابد. در زیر نمونه‌ای از استفاده از اصطلاحات غیرتخصصی در سناریوی بالینی ارائه شده است.

نمونه‌ای از سناریوی بالینی

یک خانم ۳۵ ساله در سومین مراجعه به مطب شما، با شکایت اسهال آبکی که از دیروز صبح شروع شده است، مراجعه می‌کند. او بیان می‌کند که در طول ۲۴ ساعت گذشته، ۱۵ بار اجابت مزاج داشته است. در طول این مدت، او احساس ناراحتی در معده داشته که همراه با استفراغ نبوده است. او به عنوان آشپز در یک رستوران مشغول به کار است که به دنبال شرایط ایجاد شده مجبور به گرفتن مرخصی از محل کار خود شده است. در برگه ثبت علایم حیاتی، فشارخون او از ۱۰۵/۵۰ میلی‌متر جیوه به ۹۰/۴۰ نزول داشته است و درجه حرارت دهانی ۳۶/۸ درجه سانتی‌گراد گزارش شده است. در معاینه فیزیکی، متوجه خشکی دهان و حرکات زیاد روده می‌شوید. در آنالیز ادرار، فاکتورهای میکروسکوپی طبیعی است و وزن مخصوص ادرار ۱۰۳۰ گزارش شده است.

همچنین باید به خاطر داشته باشید که مهم است فراگیر در یک موقعیت ایفای نقش واقعی قرار داده شود. به عنوان مثال، به جای اینکه ذکر شود «بیمار خانم ۲۹ ساله با قرمزی چشم که در طی دو روز گذشته بدتر شده است ...» بهتر است نوشته شود «از شما درخواست می‌شود تا یک خانم ۲۹ ساله را در مطب ویزیت کنید. او در صحبت‌های خود به شما می‌گوید که چشم راستش در طول دو روز گذشته قرمز شده است و این وضعیت رفته‌رفته شدت پیدا کرده است.» همچنین به جای اینکه پرسیده شود که «چه تشخیص‌هایی در این بیمار باید در نظر گرفته شود؟» بهتر است سؤال شود «تشخیص شما در این بیمار چیست؟». در واقع با تبدیل کلمه «باید» به «تشخیص شما»، آزمون‌شوندگان در نقش بالینی‌شان در مواجهه با یک بیمار قرار می‌گیرند.

به طور کلی هنگام نوشتن سناریو باید به موارد زیر توجه داشت:

- در هنگام نگارش سناریوی بالینی، به جای آنکه خودتان از ابتدا بخواهید آن را طرح نمایید، با استفاده از بیماران واقعی یا موقعیت حقیقی زندگی می‌توانید سناریوهای نزدیک‌تر به واقعیت و کمتر ساختگی را طراحی نمایید.

1. Coughing up blood
2. Hemoptysis
3. Homan sign
4. Eva et al.

- مطمئن شوید که اطلاعات مندرج در سناریو، واضح، روان و بدون ابهام است. برخی از استادان که از تجربه کافی برای طراحی سؤالات برخوردار نیستند، به اشتباه تصور می‌کنند که وجود ابهام در سناریو بر دشواری سؤال می‌افزاید و به این طریق بهتر می‌توان توانایی داوطلبان را ارزیابی نمود. در حالی که چنین تصوری اشتباه است. در این شرایط تنها داوطلبان از مسیر درست برای پاسخگویی به سؤالات خارج می‌شوند.
- همان‌طور که در قسمت طراحی ویژگی‌های کلیدی ذکر شد، هنگام طرح سناریوی بالینی نیز باید ویژگی‌های داوطلبان را در ذهن داشته باشید. بدین ترتیب باید همواره سعی کنید که زبان مورد استفاده در طرح سناریو متناسب با گروه داوطلبان مورد نظر باشد. توجه به این مهم در طراحی سؤالات با سطح دشواری متناسب با فراگیران نیز تاثیرگذار است.
- بدنه سؤال از نظر لغوی و دستوری نباید پیچیده باشد. از استفاده از مخففها در سناریو خودداری کنید و برعکس، سعی کنید که از کلمات و عباراتی که درشرایط واقعی مواجهه با بیمار در اختیار دانشجو قرار می‌گیرد، استفاده کنید. به این طریق از وجود ابهام در سناریو جلوگیری می‌شود.
- نکته آخر اینکه هنگام نوشتن سناریو در عین اینکه به شرح اطلاعات ضروری می‌پردازید، از ارائه مطالب غیرضروری خودداری کنید. ذکر مطالب غیر ضروری در سناریو باعث می‌شود تا فراگیران بخشی از زمان آزمون را به مطالعه مطالبی سپری کنند که هیچ اطلاعات مفیدی در اختیارشان نمی‌گذارد. با این حال، اگر اطلاعات اضافی مربوطه بسیار مرتبط با موقعیت مورد نظر هستند، مطمئن شوید که این اطلاعات نیز در نظر گرفته شده‌اند.

نگارش سؤال هدایت‌کننده

بعد از تعریف ویژگی‌های کلیدی و نوشتن سناریوی بالینی، قدم بعدی نوشتن سؤال هدایت‌کننده[1] است. سؤال هدایت‌کننده در آزمون ویژگی‌های کلیدی شامل یک عبارت مستقیم است که به صورت پرسشی نوشته می‌شود. به عنوان مثال، ممکن است بر اساس تابلوی بالینی مربوطه و ویژگی‌های کلیدی مورد انتظار از فراگیر بپرسید که «کسب چه اطلاعاتی از فرایند شرح‌حال‌گیری/ معاینه فیزیکی بیمار می‌تواند در ادامه روند کار به او کمک کند؟»

موضوع: لنفوم	محور: تشخیص	گروه هدف: میانسال	موقعیت: وضعیت بحرانی و تهدید کننده حیات بیمار

دختر ۲۱ ساله‌ای که مورد شناخته‌شدهٔ لنفوم هوچکین از سه سال قبل است با شکایت تنگی‌نفس پیشرونده از دو هفته پیش به مطب شما مراجعه کرده است. کدام یک از موارد زیر برای تشخیص مشکل فعلی وی کمک بیشتری به شما می‌کنند؟ از بین موارد زیر حداکثر چهار مورد را انتخاب کنید.

۱.	آزمون عملکرد تیروئید	۸.	تست ورزش
۲.	اسپیرومتری	۹.	رادیوگرافی قفسه سینه
۳.	اکوکاردیوگرافی	۱۰.	سونوگرافی شکم
۴.	اندازه‌گیری فشارخون	۱۱.	سی‌تی‌اسکن قفسه سینه
۵.	اندازه گیری لیپید	۱۲.	شمارش گلبول‌های قرمز خونی
۶.	بررسی ادم اندام تحتانی	۱۳.	معاینه پالس پارادوکس
۷.	بررسی فشار ورید ژگولار	۱۴.	نمونه مغز استخوان

آزمون غربالگری حیطهٔ استدلال بالینی المپیاد علمی دانشجویان علوم پزشکی تهران/ خرداد ۹۱

در مثال فوق، عبارت «کدام یک از موارد زیر برای تشخیص مشکل فعلی وی کمک بیشتری به شما می‌کنند»، سؤال هدایت‌کننده است. هنگام طراحی سؤال هدایت‌کننده همواره باید چند نکته را به خاطر داشت:

1. Leading question

- سؤال هدایت‌کننده مستقیماً از ویژگی‌های کلیدی گرفته می‌شود و باید منحصراً بر یک عملکرد یا اقدام بالینی (مداخلات درمانی، تشخیص‌های افتراقی، علایم و نشانه‌های بیماری، تست‌های آزمایشگاهی و ...) متمرکز باشد. در نوشتن دستورالعمل باید یک وظیفهٔ مشخص از داوطلب خواست. یعنی اینکه اگر بخواهیم مشکل بیمار را تشخیص دهیم به چه اطلاعاتی نیاز داریم یا اینکه اگر بخواهیم به تدبیر بالینی بیمار فکر کنیم، چه داده‌هایی را احتیاج داریم.
- چنانچه برای یک مشکل بالینی واحد قرار است چند سؤال طراحی شود، ممکن است لازم باشد برای سؤالات بعدی، یک سری اطلاعات اضافی دیگر نیز در سؤال هدایت‌کننده ارائه شود. به عنوان مثال، «در حال حاضر دو روز از بستری شدن بیمار در بیمارستان می‌گذرد و شرایط فعلی او ... است، در این شرایط چه درمانی را برای او پیشنهاد می‌کنید؟»

در ادامه نمونه‌هایی از سؤالات هدایت‌کننده در محورهای مختلف معاینه فیزیکی، شرح‌حال گیری، تشخیص و درمان ارائه شده.

نمونه‌هایی از سؤالات هدایت‌کننده در آزمون «ویژگی‌های کلیدی»

الف) معاینه فیزیکی و شرح‌حال گیری:
- چه اطلاعاتی از فرایند شرح‌حال گیری/ معاینه فیزیکی این بیمار جمع‌آوری می‌کنید (متمرکز می‌شوید، می‌پرسید، کسب می‌کنید و ...)؟
- در حال حاضر کسب چه اطلاعاتی از فرایند شرح‌حال گیری/ معاینه فیزیکی بیمار می‌تواند در ادامه روند کار به شما کمک کند؟
- در فرایند شرح‌حال گیری/ معاینه فیزیکی به جستجوی چه اطلاعات دیگری در بیمار می‌پردازید؟
- چه اطلاعات بالینی (مرتبط) دیگری در حال حاضر نیاز خواهید داشت؟
- در حین مرور برگه شرح‌حال گیری و علایم حیاتی این بیمار به چه مواردی توجه می‌کنید؟
- در صورت لزوم، چه اطلاعات دیگری را از بیمار کسب خواهید کرد؟

ب) تشخیص افتراقی:
- تشخیص اولیه شما تا به اینجا چیست؟ (لطفاً فقط یک مورد را ذکر کنید)
- به چه تشخیص‌های افتراقی باید تا به اینجا توجه دارید؟ (حداکثر سه مورد را مشخص نمایید)
- با توجه به نتایج حاصل از رادیوگرافی قفسه سینه بیمار، تشخیص فعلی شما چیست؟
- محتمل‌ترین علت بروز بیماری چیست؟

ج) مداخلات تشخیصی:
- با توجه به اطلاعات حاصل، درخواست انجام چه آزمایش‌ها یا تست‌های تشخیصی را در بیمار دارید؟
- در قدم بعدی درخواست انجام چه آزمایش‌ها و تست‌های تشخیصی را می‌دهید؟
- در صورت نیاز، چه مداخلات تشخیصی دیگری را در این بیمار درخواست خواهید کرد؟

د) درمان:
- چه برنامه درمانی در حال حاضر برای این بیمار پیشنهاد می‌کنید؟
- چه اقدامات دیگری در برنامه درمانی بیمار لحاظ می‌کنید؟
- در قدم بعدی، چه اقدام درمانی برای بیمار تجویز می‌کنید؟
- با توجه به اطلاعات حاضر، از نظر شما اقدام بعدی ضروری در روند درمان این بیمار چیست؟
- فوری‌ترین مداخله درمانی شما برای این بیمار در حال حاضر چیست؟
- بعد از انجام اقدامات اولیه درمانی، با توجه به شرایط بیمار، شما چه مداخلات دیگری را پیگیری می‌کنید؟
- چه توصیه‌های درمانی به بیمار ارائه می‌دهید؟

تصمیم‌گیری در مورد نوع سؤال (بسته‌پاسخ یا بازپاسخ)

همان‌طور که قبلاً ذکر شد، به صورت کلی، دو مدل پاسخگویی در آزمون «ویژگی‌های کلیدی» وجود دارد: سؤالات بازپاسخ و سؤالات بسته‌پاسخ[1].

در مدل بازپاسخ، در واقع سؤالات کوته‌پاسخ[2] مدنظر هستند که داوطلب باید پاسخ موردنظر را در قالب عبارات کوتاه ارائه نماید. به عنوان مثال، در پاسخ به سؤال فراگیر ممکن است عبارت «تجویز پنی‌سیلین» یا «دیابت وابسته به انسولین» را بنویسد.

1. short-menu (SM) responses
2. "write-in" (WI) responses

آزمون «ویژگی‌های کلیدی» از نوع کوته پاسخ

موضوع: حمله تشنج **محور: تشخیص** **گروه هدف: میانسالی** **موقعیت: وضعیت بحرانی و تهدید کننده حیات بیمار**

شما به عنوان پزشک اورژانس مشغول طبابت هستید که آقای ۳۶ ساله‌ای توسط آمبولانس به واحد اورژانس بیمارستان انتقال می‌یابد. در شرح حال‌گیری ابتدایی از تیم فوریت‌های پزشکی متوجه می‌شوید که ایشان در پیاده‌روی خیابان در حالی که منتظر اتوبوس بوده است، هوشیاری خود را از دست داده و به زمین افتاده است. ناظران سریعاً با اورژانس تماس گرفته و به تکنسین‌های اورژانس گزارش داده‌اند که وی قبل از افتادن به زمین هوشیار، بی‌قرار و دچاراختلال دید شده است. بیمار بعد از دست دادن هوشیاری برای مدت زمان کوتاهی دچار انقباض ناگهانی گردیده و رنگ صورت وی آبی شده و سپس حرکات تشنجی متناوب در تمام بدن برای حدود یک دقیقه بروز کرده است. وی هوشیاری خود را بعد از حمله به دست نیاورده است. در طول زمان ۱۰ دقیقه‌ای که در آمبولانس بوده است، او دو حمله مشابه را بدون اینکه هوشیاری خود را به دست آورد، تجربه کرده است و شما هنگام ورود، شاهد حمله سوم آن هستید. درجه حرارت وی ۳۷/۸ درجه سانتی‌گراد است و به نظر بیهوش می‌رسد. هیچ‌یک از اعضای خانواده یا نزدیکان همراه بیمار نیستند.

سؤال ۱: تشخیص یا تشخیص‌های شما تا به اینجا چیست؟ دو تشخیص بالینی را ذکر نمایید.
سؤال ۲: درمان فوری شما تا به اینجا چیست؟ هر چیزی که فکر می‌کنید مناسب است، فهرست نمایید.
سؤال ۳: ده دقیقه بعد از ورود، بیمار هنوز هوشیار نیست. پرستار در کیف وی یک شماره تلفن پیدا می‌کند و شما تصمیم می‌گیرید فوراً با آن شماره تماس بگیرید. به فرض اینکه فرد پشت خط بیمار را بشناسد، چه سؤالاتی را جهت پرسیدن از آن فرد در ذهن مرور می‌کنید؟ شش سؤال را ذکر کنید.
سؤال ۴: ۱۵ دقیقه از زمانی که بیمار وارد شده است، گذشته است. در این لحظه چه آزمایش‌هایی درخواست می‌دهید؟

در مدل بسته‌پاسخ، داوطلب ملزم می‌شود که از بین مجموعه پاسخ‌های احتمالی ارائه شده، پاسخ یا پاسخ‌های موردانتظار را انتخاب نماید. در این حالت، تعداد گزینه‌هایی که ارائه می‌شود، متفاوت است و ممکن است حداکثر تا ۲۵ گزینه را شامل شود. به منظور کاهش اثرات حدس زدن، مجموعه گزینه‌های ارائه شده باید شامل تصورات نادرست و اشتباهات رایج باشد.

آزمون «ویژگی‌های کلیدی» از نوع بسته‌پاسخ

موضوع: اسهال **محور: تشخیص** **گروه هدف: میانسالی** **موقعیت: وضعیت بحرانی و تهدید کننده حیات بیمار**

شما به عنوان پزشک عمومی در درمانگاه گوارش مشغول ویزیت بیماران هستید که خانم جوان ۲۹ ساله با شکایت اصلی ضعف و بی‌حالی مراجعه می‌کند. در شرح حال‌گیری از بیمار متوجه می‌شوید که وی مدت طولانی است که اسهال خونی دارد. در بررسی سابقه مصرف دارو، بیمار ذکر می‌کند در نتیجه مراجعه قبلی به پزشک در حال حاضر تحت درمان با مسالازین و آزیتوپرین و کلسیم خوراکی است. در معاینات فیزیکی مشخص می‌شود که وی تب دارد و دهیدارته است و در معاینه شکم تندرنس ژنرالیزه وجود دارد.
با توجه به موارد فوق ذکر کنید که کدام یک از موارد زیر برای تشخیص بیمار ضروری است (حداکثر ۵ مورد را انتخاب کنید).

۱. آندوسکوپی فوقانی	۷. شمارش گلبول‌های سفید خونی
۲. بررسی از نظر سلیاک	۸. کشت مدفوع
۳. بررسی آنتی ژن کلستریدیوم دیفیسیل در مدفوع	۹. کولونوسکوپی توتال
۴. بررسی گازهای خون شریانی	۱۰. گرافی ساده ایستاده شکم
۵. سونوگرافی کامل شکم و لگن	۱۱. معاینه رکتوم
۶. سی تی اسکن با کنتراست وریدی و خوراکی	

اگرچه، اجرای سؤالات نوع کوته‌پاسخ می‌تواند تا حدودی دشوار باشد، نتایج حاصل از مطالعات صورت‌گرفته نشان می‌دهد که سؤالات کوته‌پاسخ نسبت به نوع انتخاب آن، از قدرت تمیز بالاتری برخوردار هستند (پیج و همکاران ۲۰۰۰). به ویژه، در سیستم ارزیابی قبولی–ردی این موضوع بیشتر صدق می‌کند. از این رو بر اساس نتایج حاصل از مطالعات به نظر می‌رسد که سؤالات نوع بسته‌پاسخ در شناسایی فراگیران ضعیف موثرتر باشند (پیج و همکاران ۲۰۰۰).

همچنین توصیه می‌شود که بهتر است از مدل کوته‌پاسخ، در ارزیابی توانایی تشخیص، درمان و مداخلات آزمایشگاهی استفاده شود. در حالی که نوع بسته‌پاسخ معمولاً به منظور ارزیابی یافته‌های حاصل از شرح‌حال‌گیری و معاینات فیزیکی مناسب‌تر هستند. به علاوه، باید توجه داشت که سؤالات نوع کوته‌پاسخ تنها به شکل دستی قابل تصحیح هستند در حالی که سؤالات نوع بسته‌پاسخ ممکن است به وسیله رایانه نیز تصحیح شوند.

راتهف و همکاران[1] ۲۰۰۶

این پژوهشگران در مطالعه‌ای به مقایسه بین آزمون «ویژگی‌های کلیدی» از نوع کوته‌پاسخ با نوع بسته‌پاسخ پرداختند. ۱۴۶ دانشجوی سال چهار پزشکی به طور تصادفی به داخل دو گروه تقسیم شدند. در این مطالعه هفت سناریو با مجموع ۲۵ سؤال ویژگی‌های کلیدی طراحی گردید. همه سؤالات در هر دو گروه محتوای یکسانی را می‌سنجیدند و تنها نه سؤال از نظر شکل طراحی متفاوت بودند.
نتایج مطالعه نشان داد که بین متوسط تعداد پاسخ‌های صحیح در دو گروه هیچ تفاوت معناداری وجود نداشت (p= ۰/۹۳). زمان پاسخ‌دهی برای هر دو گروه نیز هیچ تفاوت معناداری را نشان نداد (p= ۰/۶۵). بر این اساس محققان خاطرنشان کردند که تفاوت معناداری بین این دو نوع سؤال در آزمون «ویژگی‌های کلیدی» وجود ندارد و همچنین ذکر کردند که در مقایسه با سؤالات چندگزینه‌ای، پاسخ‌دهی به این سؤالات به زمان بیشتری نیاز دارد و علت این امر را در ماهیت ارزیابی مهارت استدلال بالینی ذکر کردند.

1. Rotthoff et al.

طراحی گزینه‌ها در سؤال بسته‌پاسخ

اگر نوع کوته‌پاسخ برای آزمون «ویژگی‌های کلیدی» انتخاب شده باشد، این مرحله موضوعیتی ندارد و می‌توان به مرحله بعدی رفت. اما اگر نوع بسته‌پاسخ مدنظر است، گزینه‌های سؤال باید طراحی شوند. گزینه‌ها تمام پرسش‌هایی هستند که می‌توان به صورت بالقوه از دانشجو پرسید. گزینه‌های صحیح باید به وضوح، درست و بدون ابهام باشند و گزینه‌های انحرافی باید غلط، اما در عین حال جذاب و محتمل باشند، به نحوی که بتواند نظر دانشجویان ضعیف را به خود جلب کند. به عبارت دیگر می‌توان گفت که هدف از طرح گزینه‌های انحرافی این است که فراگیرانی که دانش ناکافی دارند، از جواب صحیح منحرف شوند.

برای طراحی گزینه‌های انحرافی می‌توان از اشتباهات شایع داوطلبان استفاده کرد. یک روش بسیار عالی برای تهیه فهرست پاسخ‌های محتمل در مدل بسته‌پاسخ این است که آزمون به شکل پایلوت و به صورت سؤالات کوته‌پاسخ طراحی شود و سپس از پاسخ‌های نادرست ارائه شده به عنوان منبعی برای طراحی گزینه‌های انحرافی استفاده گردد. همچنین گزینه انحرافی می‌تواند شامل پاسخ‌های صحیحی باشد که جزء ویژگی‌های کلیدی آن سؤال در نظر گرفته نشده است (گزینه انحرافی خنثی).

در مجموع، یک گزینه انحرافی خوب از ویژگی‌های زیر برخوردار است:

- از نظر محتوایی همگون با گزینه‌های صحیح است (به عنوان مثال همه گزینه‌ها باید در مورد مداخلات تشخیصی یا درمانی باشند).
- هیچ سرنخی به سمت پاسخ‌های صحیح نمی‌دهد.
- به اندازه کافی جذاب است که بتوانند باعث جلب نظر افراد تازه‌کار و غیرحرفه‌ای شود.
- از نظر ساختاری و طول گزینه‌ها با سایر گزینه‌ها همسان است.

در خصوص تعداد گزینه‌ها در آزمون «ویژگی‌های کلیدی» از نوع بسته‌پاسخ باید گفت که تعداد گزینه‌های ارائه شده باید به اندازه کافی باشد تا بتوان از حدس زدن پاسخ صحیح توسط فراگیران ضعیف ممانعت به عمل آورد. هیچ قانون ثابت‌شده‌ای برای تعداد بهینه گزینه‌ها وجود ندارد. با این حال همواره باید سعی شود تعداد گزینه‌ها متناسب با شرایط واقعی باشد که فراگیران در محیط کار با آن مواجهه می‌شوند. به علاوه، حداقل چهار یا پنج گزینه نادرست به ازای هر پاسخ درست باید در نظر گرفته شود. به عنوان یک قاعده کلی می‌توان گفت که معمولاً در آزمون «ویژگی‌های کلیدی» ۱۵ تا ۲۰ گزینه وجود دارد اما آزمون‌هایی با ۵ تا ۳۰ گزینه هم طراحی شده‌اند.

همچنین بهتر است که به منظور اجتناب از ارائه هرگونه سرنخ، گزینه‌ها به ترتیب حروف الفبا ارائه شوند.

هنگام طراحی گزینه‌ها باید موارد زیر را باید در نظر داشت:

- هر گزینه فقط در مورد یک داده باشد. مثلاً «معاینات قلب و ریه و شکم و اندام‌ها» گزینه مناسبی نیست بلکه باید آن‌ها را به گزینه‌هایی جداگانه تقسیم کرد و به صورت «معاینه قلب»، «معاینه ریه»، «معاینه شکم» و «معاینه اندام‌ها» نوشت. همان‌طور که در مثال زیر مشاهده می‌کنید، طرح دو ویژگی کلیدی در قالب یک گزینه اشتباه است و در هر گزینه تنها باید یک داده عملکردی مطرح گردد.

گزینه ضعیف

تزریق فوری یک دوز آمپی‌سیلین در بخش اورژانس و ترخیص بیمار با تجویز آمپی‌سیلین خوراکی

گزینه بهتر

تزریق فوری یک دوز آمپی‌سیلین در بخش اورژانس
ترخیص بیمار با تجویز آمپی‌سیلین خوراکی

- گزینه‌هایی که به صورت بدیهی کاملاً غلط هستند، را نباید در سؤال گنجاند چرا که در این شرایط داوطلب بدون نیاز به خواندن سناریو، به راحتی نادرست بودن آن را تشخیص می‌دهد.
- مسأله مهم دیگر در انتخاب گزینه‌ها، کلی یا جزئی بودن آن‌هاست. تصمیم در مورد جزئی یا کلی نوشتن گزینه‌ها بستگی به هدف سؤال دارد. به عنوان مثال می‌توان آزمایش کامل ادرار یا اندازه‌گیری میزان پروتئین در ادرار را به عنوان گزینه مطرح کرد. باید توجه داشت که آوردن هر دو گزینه در یک سؤال نابجاست و بسته به هدف مورد نظر یکی از این دو مناسب‌تر است.

نکته دیگر اینکه بدیهی است اگر سؤال به صورت بسته‌پاسخ طراحی شود، به پاسخنامه‌ای نیاز است که دانشجو جواب‌های درست را در آن علامت بزند. برخلاف پاسخنامه آزمون‌های چندگزینه‌ای معمول که در آنها تنها یک پاسخ صحیح وجود دارد، پاسخنامه آزمون «ویژگی‌های کلیدی» باید به شکلی باشد که امکان انتخاب تعداد بیشتری از گزینه‌ها در آن وجود داشته باشد. در اینجا نمونه‌ای از پاسخ‌نامه آزمون «ویژگی‌های کلیدی» نوع بسته‌پاسخ ارائه شده است. فراگیر باید روبروی هر سؤال شماره گزینه‌های صحیح را با علامت ضربدر مشخص نماید.

	گزینه ۱	گزینه ۲	گزینه ۳	گزینه ۴	گزینه ۵	گزینه ۶	گزینه ۷	گزینه ۸	گزینه ۹	گزینه۱۰
سؤال ۱		×	×				×			
سؤال ۲		×	×							
سؤال ۳				×	×	×			×	×
سؤال ۴			×	×						
سؤال ۵		×		×						

تصمیم‌گیری در خصوص نحوه نمره‌دهی سؤالات

یکی از بخش‌هایی که تفاوت عمده میان آزمون‌های استدلال بالینی و سایر آزمون‌های متداول را نشان می‌دهد، نحوه نمره‌دهی آن‌ها است. در این آزمون‌ها تنها یک جواب صحیح وجود ندارد، بلکه چندین گزینه می‌تواند به عنوان پاسخ صحیح سؤال در نظر گرفته شود. همچنین صحت جواب‌ها توسط گروه خبرگان تشخیص داده می‌شود (جهت کسب اطلاعات بیشتر در این خصوص به فصل یک همین بخش مراجعه نمایید). از این رو، در این آزمون‌ها، کلید آزمون بر اساس ارجاع به صفحه مشخصی از یک کتاب مرجع نخواهد بود.

به طور کلی، در آزمون «ویژگی‌های کلیدی» با توجه به تعداد زیاد سؤال و طبعاً پاسخ‌های متعدد صحیح، ضروری است که تصویر روشنی از شیوه نمره‌دهی و نحوه تخصیص نمره برای هر سؤال وجود داشته باشد. فرایند نمره‌دهی در آزمون «ویژگی‌های کلیدی» شامل تعیین تعداد پاسخ‌های صحیح مجاز و تخصیص نمره به پاسخ‌ها است.

مشخص نمودن تعداد پاسخ‌های مجاز

فارغ از اینکه سؤال از نوع بسته‌پاسخ یا کوته‌پاسخ است، در ادامه باید تعداد پاسخ‌هایی که فراگیر می‌تواند به عنوان پاسخ صحیح انتخاب یا ذکر نماید، مشخص کرد. پاسخ‌های صحیح تعیین شده باید مستقیماً منعکس‌کننده ویژگی‌های کلیدی مورد ارزیابی باشند. به عبارت دیگر، تمام پاسخ‌های صحیح در کلید باید بخشی از ویژگی‌های کلیدی مورد نظر باشند، نه بیشتر و نه کمتر. توجه داشته باشید که تعداد پاسخ‌های مجاز در هر سؤال متفاوت است و بستگی به هدف سؤال دارد. به علاوه، در صورتی که وزن و ارزش پاسخ‌های مختلف یکسان نیست، حتماً باید ذکر شود که فراگیران پاسخ‌های خود را اولویت‌بندی کنند.

تعیین تعداد مجاز پاسخ توسط گروه خبرگان معمولاً به یکی از سه شکل زیر است:

- انتخاب یا ذکر تنها یک پاسخ: از این نوع، در مواردی که یک پاسخ قطعی مشخص (از قبیل تشخیص اولیه یا مهم‌ترین اقدام درمانی) وجود دارد، می‌توان استفاده کرد.
- انتخاب یا ذکر حداکثر تا X پاسخ: این نوع محدودیت برای سؤالاتی مناسب است که یک یا تعداد بیشتری پاسخ صحیح دارند. تعیین تعداد X بر اساس تعداد پاسخ‌های صحیح قطعی و تعداد پاسخ‌های نادرستی که ممکن است برای داوطلبان با عملکرد ضعیف جذاب، باشد صورت می‌گیرد. در این شکل، در صورتی که داوطلب تعداد بیشتری از تعداد پاسخ‌های مجاز را ذکر نماید، هیچ نمره‌ای از آن سؤال کسب نخواهد کرد.
- بدون محدودیت: این نوع، در مواردی قابل کاربرد است که تعیین تعداد اقداماتی که یک فراگیر درخواست می‌دهد، مهم باشد. در این روش اگرچه حداکثر تعداد پاسخ‌ها در فرایند نمره‌دهی باید مشخص گردد، با این حال داوطلب از این محدودیت هیچ اطلاعی ندارد.

تهیه بارم هر پاسخ

پس از اینکه پاسخ‌های صحیح مورد نظر مشخص شدند، باید نمره‌ای که به جواب‌ها تعلق می‌گیرد، تعیین شود. تخصیص میزان نمره بین پاسخ‌های مختلف بسته به درجه اهمیت پاسخ‌ها دارد. در مواردی که هر یک از پاسخ‌های صحیح از اهمیت یکسانی برخوردار هستند، بارم نمره سؤال بین گزینه‌های مختلف که پاسخ صحیح تلقی می‌شوند، تقسیم می‌گردد اما در مواردی که ارزش و درجه اهمیت پاسخ‌های صحیح مربوط به یک سؤال یکسان نیست بارم سؤال بر اساس درجه اهمیت هر پاسخ تقسیم می‌شود. به طور کلی، استفاده از وزن‌های یکسان نمره بین گزینه‌های مختلف ترجیح داده می‌شود.

یه عنوان یک قاعده کلی، هر چه شیوه نمره‌دهی آسان‌تر باشد مطلوب‌تر است و در مقابل، هر چه نمره‌دهی آزمون پیچیده‌تر و متنوع‌تر باشد احتمال خطای تصحیح در آن بیشتر خواهد بود. مطالعات مختلف نشان می‌دهند که اختصاص وزن‌های متفاوت نمره‌دهی به بهبود پایایی آزمون کمکی نمی‌کند (پیج و همکاران ۱۹۹۵، نورسینی و همکاران ۱۹۸۳). بنابراین، انجام این کار صرفاً منجر به صرف زمان بیهوده در فرایند نمره‌دهی خواهد شد.

راتهف و همکاران[1] ۲۰۰۶

مثال ۱: داوطلب در یک بیمار با علایم ساق پای متورم و دردناک باید به ترومبوز وریدهای عمقی در فهرست تشخیص‌های افتراقی توجه نماید. بنابراین نمره‌ای که برای این سؤال لحاظ می‌شود، به شکل زیر است.

نمره	پاسخ صحیح	پاسخ‌های مترداف
۱	ترومبوز وریدهای عمقی	DVT, Deep Venous thrombosis
صفر	پاسخ‌های دیگر	-

مثال ۲: داوطلب در فرایند شرح‌حال گیری فرد مبتلا به تشنج مکرر و بدون هوشیاری باید به علل احتمالی آن شامل استفاده از الکل، داروها، مواد مخدر و دیابت توجه نماید.

نمره	پاسخ صحیح	پاسخ‌های مترداف
۰/۲۵	سابقه استفاده از الکل	اعتیاد به الکل
۰/۲۵	اعتیاد به کوکائین یا هروئین	-
۰/۲۵	سابقه ابتلا به دیابت	-
۰/۲۵	سابقه مصرف دارو	-
صفر	ذکر بیش از چهار پاسخ	-
صفر	ذکر پاسخ‌های دیگر	

همان‌طور که در مثال اول دیده می‌شود، ممکن است تنها یک پاسخ صحیح قابل قبول باشد یا مانند مثال دوم ممکن است چندین پاسخ صحیح ارائه شود.

باید به خاطر داشت که علاوه بر تعیین سهم نمره مربوط به پاسخ‌های صحیح در کلید پاسخ، باید تصمیماتی در خصوص مواردی که انتخاب یا نوشتن آن‌ها باعث می‌شود تا داوطلب نتواند هیچ نمره‌ای کسب کنند نیز اتخاذ شود. به این جواب‌ها، پاسخ‌های از بین‌برنده[1] نیز اطلاق می‌شود. علت این نامگذاری از آن جهت است که انتخاب یا ذکر این پاسخ‌ها از سوی داوطلب منجر به از دست رفتن نمره دانشجو در آن سؤال حتی با فرض اشاره به تعدادی از پاسخ‌های صحیح یا کل آن‌ها می‌شود. این کار در مواردِ انتخاب یا ذکر تعداد زیادی پاسخ یا در نظر گرفتن اقداماتی که می‌توانند برای بیمار آسیب رسان باشند (به عنوان مثال، درخواست آزمایش‌های تهاجمی غیرضروری یا اقدامات درمانی خطرناک) باید حتماً صورت گیرد. البته باید در نظر داشت که تعیین تعداد زیادی پاسخ به عنوان پاسخ از بین برنده، اثرات منفی بر قدرت تمیز سؤال خواهد داشت.

یکی دیگر از موارد مورد توجه در فرایند نمره‌دهی آزمون ویژگی‌های کلیدی، عدم ارائه نمره منفی به پاسخ‌های غلط فراگیران است. علت این امر، عدم تأثیر نمره منفی بر بهبود پایایی آزمون است. به علاوه، در این حالت ممکن است به دلیل متفاوت بودن میزان ریسک‌پذیری دانشجویان، نتایج حاصل دچار خطا شود (فارمر و پیج ۲۰۰۵).

در نظر داشته باشید که در هر سؤال، در نهایت جمع سهم نمره مربوط به هر پاسخ نباید از یک نمره تجاوز کند. استثنا در این مورد زمانی است که هر سؤال به ارزیابی بیش از یک ویژگی کلیدی می‌پردازد. در این حالت برای هر ویژگی کلیدی حداکثر میزان نمره در نظر گرفته شده برابر با یک و حداقل آن صفر خواهد بود. همان‌گونه که در بالا نیز در خصوص سهم نمره گزینه‌های صحیح اشاره شد، تخصیص سهم نمره در ویژگی‌های کلیدی مربوط به هر سؤال و حتی موارد بالینی در کل آزمون بستگی به درجه اهمیت آن‌ها دارد.

در مجموع نمره‌دهی این آزمون‌ها به دو بخش تقسیم می‌شود: نمره‌دهی به هر سؤال و نمره‌دهی به کل آزمون. این دو بخش می‌توانند به روش‌های متفاوتی انجام شوند که در فصل یک همین بخش به تفصیل در مورد روش‌های مختلف محاسبه آن در گروه خبرگان توضیحاتی ارائه شده است.

تدوین دستورالعمل کلی آزمون

در ابتدای دفترچه آزمون «ویژگی‌های کلیدی» لازم است که دستورالعمل آزمون درج شود. مخصوصاً اگر این شیوه ارزیابی برای بار اول مورد استفاده قرار می‌گیرد و دانشجویان با آن آشنایی ندارند، وجود آن ضروری است. برعکس

1. Killer responces

سناریوی سؤال که عمدتاً باید مبهم نوشته شود، دستورالعمل آزمون «ویژگی‌های کلیدی» را باید کاملاً واضح و گویا نوشت. در دستورالعمل کلی آزمون باید توجه دانشجویان را به مواردی از قبیل تعداد پاسخ‌های مجاز، شیوه پاسخ‌دهی به دو نوع سؤالات کوته‌پاسخ و بسته‌پاسخ، بارم سؤالات و مواردی از این دست جلب کرد.

به علاوه، در متن دستورالعمل باید به این نکته تاکید کرد که انتخاب یا ذکر پاسخ‌های بیشتر به معنای به دست آوردن نمره بیشتر نیست. همچنین در دستورالعمل آزمون باید به مدت زمان پاسخگویی به آزمون نیز اشاره نمود تا فراگیران بتوانند زمان آزمون را مدیریت نمایند. این موضوع از حیث عدم وجود تجربه کافی برای شرکت فراگیران در این آزمون‌ها اهمیت دارد. بنابراین ذکر زمان آزمون الزامی است.

نمونه‌ای از دستورالعمل آزمون «ویژگی‌های کلیدی»

داوطلب گرامی،
در هر ردیف مجاز هستید از میان گزینه‌های موجود حداکثر پنج گزینه را انتخاب نمایید.
در صورت انتخاب بیش از پنج گزینه نمره منفی محاسبه خواهد شد و انتخاب گزینه‌های بیشتر به معنای به دست آوردن نمره بیشتر در آزمون نیست.
بارم هر سؤال یک نمره است.
مدت زمان کل آزمون یک ساعت خواهد بود.

مرور سؤال

پس از اینکه سؤال خود را طراحی کردید، مهم است که یک بار آن را مرور کنید و ببینید که آیا نکات مطرح شده را در خصوص آن رعایت کرده‌اید یا خیر. سپس در صورت لزوم سؤال خود را اصلاح کنید. برای این کار می‌توانید از همکاران خود نیز کمک بگیرید. معمولاً دیگران بهتر متوجه اشتباهات ما می‌شوند. بدین ترتیب، با نظر خواستن از همکاران قبل از اینکه سؤال به دست دانشجو برسد، از اشتباهات احتمالی جلوگیری کنید. به علاوه در این مرحله باید از صحت پاسخ‌های سؤالات نیز اطمینان حاصل کنید. این کار را می‌توانید از طریق ارجاع به منابع معتبر یا پانل متخصصان انجام دهید.

آزمون «ویژگی‌های کلیدی» از نوع بسته‌پاسخ

مشکل بالینی:	محور:
گروه هدف:	موقعیت:

ویژگی‌های کلیدی

ویژگی کلیدی ۱:
ویژگی کلیدی ۲:
ویژگی کلیدی ۳:
ویژگی کلیدی ۴:

سناریو

	ویژگی کلیدی ۱	ویژگی کلیدی ۲	ویژگی کلیدی ۳	ویژگی کلیدی ۴
سؤال هدایت کننده ۱				
سؤال هدایت کننده ۲				
سؤال هدایت کننده ۳				

کلید پاسخ سؤال شماره یک		
نمره	پاسخ صحیح	پاسخ های مترادف

نمونه سؤال «ویژگی‌های کلیدی» از نوع کوته‌پاسخ

موضوع: حمله تشنج
گروه هدف: میانسالی
محور: تشخیص
موقعیت: وضعیت بحرانی و تهدید کننده حیات بیمار

ویژگی‌های کلیدی این مورد با پاسخ‌های پیشنهادی

ویژگی کلیدی ۱: پیشنهاد دادن تشخیص موقتی استاتوس اپیلپتیکوس
ویژگی کلیدی ۲: اطمینان و حفظ وضعیت قلبی تنفسی بیمار
ویژگی کلیدی ۳: شروع درمان اولیه با نرمال سالین، ویتامین ب، گلوکز، دیازپام و فنی‌توئین
ویژگی کلیدی ۴: اخذ شرح‌حال از بیمار در ارتباط با عللی مانند مصرف الکل، داروها، دیابت
ویژگی کلیدی ۵: دستور آزمایش‌ها به صورت فوری: الکترولیت‌ها، گلوکز، کلسیم، گازهای خون شریانی و سی‌تی‌اسکن

سناریو

شما به عنوان پزشک اورژانس مشغول طبابت هستید که آقای ۳۶ ساله‌ای توسط آمبولانس به واحد اورژانس بیمارستان انتقال می‌یابد. در شرح حال‌گیری ابتدایی از تیم فوریت‌های پزشکی متوجه می‌شوید که ایشان در پیاده‌روی خیابان در حالی که منتظر اتوبوس بوده است، هوشیاری خود را از دست داده و به زمین افتاده است. ناظران سریعاً با اورژانس تماس گرفته و به تکنسین‌های اورژانس گزارش داده‌اند که وی قبل از افتادن به زمین هوشیار، بی‌قرار و دچاراختلال دید شده است. بیمار بعد از از دست دادن هوشیاری برای مدت زمان کوتاهی دچار انقباض ناگهانی گردیده و رنگ صورت وی آبی شده و سپس حرکات تشنجی متناوب در تمام بدن برای حدود یک دقیقه بروز کرده است. وی هوشیاری خود را بعد از حمله به دست نیاورده است. در طول زمان ۱۰ دقیقه‌ای که در آمبولانس بوده است، او دو حمله مشابه را بدون اینکه هوشیاری خود را به دست آورد، تجربه کرده است و شما هنگام ورود، شاهد حمله سوم آن هستید. درجه حرارت وی ۳۷/۸ درجه سانتی‌گراد است و به نظر بیهوش می‌رسد. هیچ یک از اعضا خانواده یا نزدیکان همراه بیمار نیستند.

	ویژگی کلیدی ۱	ویژگی کلیدی ۲	ویژگی کلیدی ۳	ویژگی کلیدی ۴	ویژگی کلیدی ۵
سؤال ۱	×				
سؤال ۲		×	×		
سؤال ۳				×	
سؤال ۴					×

سؤال ۱: تشخیص یا تشخیص‌های شما تا به اینجا چیست؟
سؤال ۲: درمان فوری شما تا به اینجا چیست؟ هر چیزی که فکر می‌کنید مناسب است، فهرست نمایید.
سؤال ۳: ده دقیقه بعد از ورود، بیمار هنوز هوشیار نیست. پرستار در کیف وی یک شماره تلفن پیدا می‌کند و شما تصمیم می‌گیرید فوراً با آن شماره تماس بگیرید. به فرض اینکه فرد پشت خط بیمار را بشناسد، چه سؤالاتی را جهت پرسیدن از آن فرد در ذهن مرور می‌کنید؟ شش سؤال را ذکر کنید.
سؤال ۴: ۱۵ دقیقه از زمانی که بیمار وارد شده است، گذشته است. در این لحظه چه آزمایش‌هایی درخواست می‌دهید؟

کلید پاسخ سؤال

نمره	پاسخ صحیح	پاسخ‌های مترادف
۰/۲۵	استاتوس اپیلپتیکوس	صرع مداوم
۰/۲۵	اطمینان و حفظ وضعیت قلبی تنفسی بیمار شروع درمان اولیه با نرمال سالین، ویتامین ب، گلوکز، دیازپام و فنی‌توئین	
۰/۲۵	اخذ شرح‌حال از بیمار در ارتباط با عللی مانند مصرف الکل، داروها، دیابت، سابقه خانوادگی، سابقه بیماری قبلی	
۰/۲۵	دستور آزمایش‌ها به صورت فوری: الکترولیت‌ها، گلوکز، کلسیم، گازهای خون شریانی و سی‌تی‌اسکن	

نمونه سؤال «ویژگی‌های کلیدی» از نوع بسته‌پاسخ

موضوع: اسهال
گروه هدف: خانم میانسال
محور: درمان
موقعیت: وضعیت بحرانی و تهدید کننده حیات بیمار

ویژگی‌های کلیدی این مورد با پاسخ‌های پیشنهادی

ویژگی کلیدی ۱: منع تغذیه از راه دهان و تجویز سدیم کلراید ۰/۹ درصد وریدی
ویژگی کلیدی ۲: درخواست کشت مدفوع
ویژگی کلیدی ۳: مشاوره بیماری عفونی

سناریو

از شما درخواست می‌شود تا یک خانم ۳۵ ساله با شکایت از اسهال آبکی و شدید را در کلینیک‌تان ویزیت کنید. از سؤالاتی که مطرح می‌کنید، مشخص می‌شود ۲۴ ساعت است که دچار اسهال شده است. طی این مدت ۱۵ بار اسهال آبکی، تهوع بدون استفراغ داشته است. وی به عنوان آشپز در یک مرکز نگهداری بیماران مزمن کار می‌کند اما محل کار خود را ترک کرده و به مطب مراجعه کرده است. پرستار مطب در پرونده او فشارخون در حال استراحت را ۱۰۵/۵۰ میلی‌متر جیوه، در حال خوابیده را ۱۱۰ و در حالت ایستاده را ۹۰/۴۰ و درجه حرارت دهانی را ۳۶/۸ ثبت نموده است. در معاینه فیزیکی مخاطات خشک است و صداهای روده‌ای بیش‌فعال است. آزمایش ادرار، طبیعی است و وزن مخصوص ادرار ۱۰۳۰ است.

	ویژگی کلیدی ۱	ویژگی کلیدی ۲	ویژگی کلیدی ۳
سؤال ۱	×		
سؤال ۲		×	
سؤال ۳			×

سؤال ۱: فوری‌ترین مداخله درمانی شما برای این بیمار در حال حاضر چیست؟ (تا سه مورد را انتخاب نمایید)

۱. یودوکینول	۲. منع تغذیه از راه دهان
۳. متوکلوپرامید	۴. مایعات خوراکی
۵. سدیم کلراید ۰/۹ درصد وریدی	۶. سوند رکتال
۷. سرم قندی نمکی وریدی	۸. جنتامایسین وریدی
۹. مترونیدازول	۱۰. انتقال به بیمارستان
۱۱. ساکشن و گذاردن لوله معده	۱۲. مشاوره جراحی
۱۳. آمپی‌سیلین خوراکی	۱۴. کلرامفینکل خوراکی

سؤال ۲: بعد از مدیریت اولیه این بیمار چه اقدامات دیگری را انجام می‌دهید؟ تا چهار مورد را انتخاب کنید.

۱. اجتناب از فراورده‌های لبنی	۲. غیبت از کار به طور موقت
۳. احتیاطات روده‌ای	۴. کولونوسکوپی
۵. تجویز سرم ایمونوگلوبولین	۶. مشاوره گوارش
۷. ایزولاسیون شدید بیمار	۸. مشاوره عفونی
۹. کشت مدفوع	۱۰. اطلاع به اداره بهداشت عمومی

پاسخنامه

گزینه ۱۱	گزینه ۱۰	گزینه ۹	گزینه ۸	گزینه ۷	گزینه ۶	گزینه ۵	گزینه ۴	گزینه ۳	گزینه ۲	گزینه ۱	
						×			×		سوال ۱
		×	×								سوال ۱

کلید پاسخ سوال	
نمره	پاسخ صحیح
۰/۲۵	منع تغذیه از راه دهان
۰/۲۵	سدیم کلراید ۰/۹ درصد وریدی
۰/۲۵	کشت مدفوع
۰/۲۵	مشاوره بیماری عفونی

سودمندی آزمون «ویژگی‌های کلیدی»

روایی

امکان نمونه‌گیری وسیع‌تر از مسائل بالینی در آزمون ویژگی‌های کلیدی به بهبود روایی محتوایی این آزمون‌ها می‌انجامد. بر اساس شواهد موجود، در این نوع آزمون بر خلاف PMP، نمونه‌گیری کافی از محتوای دوره درسی امکان‌پذیر می‌شود (فارمر و هینچی ۲۰۰۵، پیج و همکاران ۱۹۹۵). یافته‌های سایکومتریک نشان می‌دهد که در صورتی که بلوپرینت یا جدول مشخصات آزمون به خوبی تنظیم و طراحی شود، این آزمون از روایی محتوای بالای برخوردار خواهد بود (پیج و بوردیج ۱۹۹۵).

امینی و همکاران ۱۳۹۲

پژوهشگران در مطالعه خود به بررسی نتایج آزمون «ویژگی‌های کلیدی» و آزمون چهارگزینه‌ای در پایان دوره کاروزی داخلی دانشگاه علوم پزشکی شیراز پرداختند. در این پژوهش شبه‌تجربی، تمام دانشجویان سال آخر پزشکی در بخش داخلی (۱۰۰ نفر) شرکت کردند. پس از برگزاری آزمون چهارگزینه‌ای، آزمون «ویژگی‌های کلیدی» برگزار شد. سپس همبستگی بین نمرات دو آزمون محاسبه شد. پایایی آزمون ویژگی‌های کلیدی با روش آلفای کرونباخ تعیین شد. ضریب دشواری سؤالات نیز با روش من‌ویتنی و سیبرز[1] محاسبه شد. همچنین در این مطالعه، ضریب همبستگی هر سؤال با کل آزمون به عنوان شاخص ضریب تمیز مورد محاسبه قرار گرفت.
در آزمون «ویژگی‌های کلیدی» ۴۴ نمره‌ای، حداقل نمره دانشجویان ۲۳/۷ و حداکثر ۴۱/۲ و میانگین نمرات آنان برابر با ۳۰/۹ بود. پایایی آزمون ویژگی‌های کلیدی برابر با ۰/۶۸، ضریب دشواری سؤالات بین ۰/۳ تا ۰/۸ و ضریب تمیز سؤالات بین ۰/۱۶ تا ۰/۵۴ به دست آمد. ضریب همبستگی بین آزمون «ویژگی‌های کلیدی» و چهار گزینه‌ای ۰/۲۵ به دست آمد که از نظر آماری معنادار بود. بر اساس نتایج حاصل محققان نتیجه گرفتند که سنجش مقوله استدلال بالینی توسط آزمون چهارجوابی امکان‌پذیر نبوده و استفاده از روش‌های ارزیابی جدید در حیطه بالینی از جمله آزمون نکات کلیدی را توصیه نمودند.

1. Mann-Whitney & Siber method

بوردیج و همکاران ۱۹۹۵

به منظور ارزیابی روایی محتوایی آزمون «ویژگی‌های کلیدی» طراحی شده توسط کمیته ارزیابی انجمن پزشکی کانادا[1] ، از ۹۹ پزشک عمومی خارج از این کمیته درخواست شد تا در سه پروژه‌ای که در سال ۱۹۹۱ انجام شد، شرکت نمایند. فاز اول که یک مطالعه گذشته‌نگر بود با هدف تعیین میزان موافقت و عدم موافقت پزشکان با سؤالات «ویژگی‌های کلیدی» که برای هر مشکل بالینی توسط اعضا کمیته ارزیابی طراحی شده بود، صورت گرفت. فاز دوم مطالعه که یک مطالعه آینده‌نگر بود، با هدف مقایسه سؤالات طراحی‌شده توسط شرکت‌کنندگان با سؤالات طراحی شده کمیته صورت گرفت. در فاز سوم مطالعه نیز از پزشکان درخواست شد تا تخمین بزنند که از نظر آنان به چه میزان دانشجویان با این موارد بالینی طراحی شده در واقعیت مواجه می‌شوند.
تقریباً تمام سؤالات طراحی شده توسط کمیته ارزیابی توسط پزشکان مورد تأیید قرار گرفت (۹۲ درصد در فاز اول و ۹۴ درصد در فاز دوم). بر اساس نظر پزشکان، دانشجویان حداقل یک بار یا دو بار در طول دوره کارآموزی با ۲۲ مورد بالینی (۳۷ درصد)، سه تا پنج بار با ۲۷ مورد بالینی (۴۶ درصد) و شش بار یا بیشتر با ۱۰ مورد بالینی طراحی شده مواجه می‌شوند. نتایج حاصل از سه فاز مطالعه نشان‌دهنده بالا بودن روایی محتوایی آزمون «ویژگی‌های کلیدی» طراحی‌شده توسط کمیته ارزیابی بود.

1. Test committee for the Medical Council of Canada

پایایی

در آزمون «ویژگی‌های کلیدی»، به دلیل طرح تعداد سؤالات بیشتر در مدت زمان کلی آزمون، امکان نمونه‌گیری وسیع‌تری از مسائل بالینی وجود دارد که این موضوع به بهبود پایایی نمرات این آزمون‌ها می‌انجامد. به طور متوسط هر داوطلب به منظور پاسخگویی به هر سؤال «ویژگی‌های کلیدی» حدود سه دقیقه زمان و برای هر سؤال PMP به حدود شش دقیقه و نیم زمان نیاز دارد. به این ترتیب در مدتی که می‌توان ۱۰ تا ۱۲ PMP برگزار کرد، ۳۰ تا ۴۰ سؤال از نوع ویژگی کلیدی قابل پاسخگویی است (بوردیج و همکاران ۱۹۹۵).

- انجمن پزشکی کانادا، پایایی آزمون «ویژگی‌های کلیدی» را در صورتی که ۴۰ سؤال در مدت زمان تقریبی ۳/۵ ساعت برگزار شود، در حدود ۰/۸ اعلام نموده است (بوردیج و پیج ۱۹۹۵).
- بر اساس نتایج حاصل از مطالعه دیگری برای رسیدن به حد مورد انتظار پایایی ۰/۸۰ آزمون ویژگی‌های کلیدی باید دارای ۴۰ مشکل بالینی باشد که حدود ۴/۱ ساعت طول می‌کشد (پیج و بوردیج (۱۹۹۵).
- از نظر هاتالا و نورمن[1] (۲۰۰۲)، یک آزمون ویژگی‌های کلیدی با ۱۵ مشکل بالینی، ضریب پایایی در حدود ۰/۵۰ خواهد داشت که برای آزمون‌های با درجه اهمیت متوسط مناسب است. اگر پایایی بالاتر از ۰/۵ نیاز است، آزمون «ویژگی‌های کلیدی» می‌تواند در مدت زمان طولانی‌تر و با طرح تعداد بیشتری مورد بالینی برگزار گردد. به عنوان مثال، برگزاری یک آزمون چهارساعته پایایی در حدود ۰/۶۷ خواهد داشت.
- در پژوهشی که توسط فیشر (۲۰۰۵) انجام شده است، پایایی آزمون ویژگی‌های کلیدی ۰/۶۵ گزارش شده است.
- نتایج حاصل از مطالعه نورمن نیز نشان داد که استفاده از یک مورد بالینی در آزمون «ویژگی‌های کلیدی» به پایایی پایین‌تر این آزمون‌ها منتهی می‌شود. از طرف دیگر، طرح چهار و یا تعداد بیشتر سؤال در هر مورد بالینی کار بیهوده‌ای خواهد بود و به بهبود پایایی نمرات آزمون کمکی نخواهد کرد (نورمن و همکاران ۲۰۰۶).

هاتالا و نورمن ۲۰۰۲

پژوهشگران آزمون «ویژگی‌های کلیدی» را در دوره کارآموزی بالینی طراحی و اجرا کردند. این مطالعه در فاصله سال‌های ۱۹۹۸ تا ۱۹۹۹ در بخش داخلی دانشگاه مک‌مستر انجام شد. آزمون «ویژگی‌های کلیدی» بر اساس بلوپرینت آزمون و اهداف آموزشی مورد انتظار طراحی شد و در نهایت ۸۲ مورد بالینی بر اساس تکرار بیماری‌ها، شدت و اهمیت آن‌ها اولویت‌بندی شدند. هشت عضو هیأت علمی و چهار دستیار به صورت مجزا در کارگاه طراحی سؤالات شرکت کردند. هر کدام از اعضای هیأت علمی مسؤول طراحی ۱۴ تا ۱۸ مورد بالینی بودند که هر مورد بالینی شامل یک تا چهار ویژگی کلیدی بود. مجموعه موارد بالینی طراحی شده در نهایت توسط سایر اعضای هیأت علمی مرور و بررسی گردید. سؤالات طراحی شده از نوع بازپاسخ بودند. در این مطالعه در صورت ارائه حداقل پاسخ قابل قبول نمره یک به دانشجو تعلق گرفت و در صورت عدم ارائه پاسخ‌های مورد انتظار هیچ نمره‌ای به دانشجو اختصاص نیافت. به علاوه، برای هر سؤال بر اساس تعداد ویژگی‌های کلیدی مورد اندازه‌گیری، بین یک تا چهار نمره در نظر گرفته شده بود. این آزمون در مدت زمان دو ساعت اجرا شد و شامل ۱۵ ویژگی‌های کلیدی بود که برای ۱۰۱ دانشجو در هشت گروه در چرخش بخش داخلی برگزار گردید. پایایی آزمون بوسیله آلفای کرونباخ محاسبه و ۰/۴۹ گزارش گردید. همبستگی بین نتایج حاصل از آزمون با سایر آزمون‌های شناختی و عملکردی در حد متوسط گزارش شد. محققان بر اساس نتایج حاصل از مطالعه نتیجه‌گیری کردند که آزمون «ویژگی کلیدی» از پایایی و قابلیت اجرای خوبی برای اندازه‌گیری مهارت استدلال بالینی فراگیران برخوردار است و پایین بودن همبستگی نتایج حاصل از این آزمون با سایر آزمون دال بر این نیست که این آزمون خوبی نیست، بلکه بدین معناست که این آزمون حوزه دیگری از استدلال بالینی را اندازه‌گیری می‌کند که به وسیله سایر آزمون قابل اندازه‌گیری نیست. به علاوه از نظر محققان اجرای آزمون «ویژگی‌های کلیدی» در بخش‌های بالینی مزایای زیادی را به دنبال خواهد داشت از جمله آنکه باعث تعامل بیشتر بین اعضای هیأت علمی در فرایند تدوین سؤالات امتحانی و فراهم سازی فرصتی برای بحث پیرامون مباحث آموزشی از قبیل اهداف آموزشی و روش‌های ازریابی می‌شود.

مقبولیت

از این سؤالات در سراسر دنیا برای ارزیابی استدلال بالینی فراگیران استفاده می‌شود که نشان‌دهنده مقبولیت بالای این آزمون‌هاست. از جمله مؤسسات آموزشی و کمیته‌های ارزیابی که از آزمون «ویژگی‌های کلیدی» در فرایند ارزیابی

1. Hatala & Norman

خود استفاده می‌کنند، می‌توان به دانشکده پزشکان و جراحان پاکستان (علی و بوردیج ۱۹۹۵)، بورد جراحی کولون و رکتال آمریکا[1] (ترودل و همکاران ۲۰۰۸)، کالج سلطنتی پزشکان عمومی استرالیا[2] (فارمر و پیج ۲۰۰۵، فارمر و هینچی ۲۰۰۵)، بورد امتحانات ملی سویس، دانشکده‌های پزشکی آلمان (فیشر و همکاران[3] ۲۰۰۵) اشاره نمود. همچنین دانشگاه تورنتو از آزمون ویژگی‌های کلیدی به عنوان بخشی از امتحانات داخلی دانشجویان پزشکی استفاده می‌کند (فارمر و پیج ۲۰۰۵).

کیوا و همکاران[4] ۲۰۰۷

در این مطالعه از آزمون «ویژگی‌های کلیدی» به منظور ارزیابی مهارت‌های استدلال بالینی در بخش پزشکی خانواده دانشگاه مالزی استفاده شد. در سال ۲۰۰۶، ۱۶ مورد بالینی در مدت زمان دو ساعت مورد ارزیابی قرار گرفت. هر مورد بالینی شامل چندین سؤال ویژگی‌های کلیدی (به طور معمول دو تا چهار سؤال) بود. سؤالات یا به صورت کوته‌پاسخ و یا بسته‌پاسخ (به طور معمول بین ۱۰ تا ۳۰ گزینه) طراحی شدند. بر اساس تجربه اجرای این آزمون در دانشگاه مالزی، محققان نتیجه گرفتند که آزمون «ویژگی‌های کلیدی» پوشش مناسبی از مشکلات موجود در حیطه پزشکی خانواده را فراهم می‌سازد.

4. Kwa et al.

هزینه

با در نظر گرفتن این موضوع که هزینه هرآزمون شامل هزینه‌های مربوط به طراحی سؤال، اجرای آن و تصحیح برگه‌های امتحانی است. واضح است که در هر مؤسسه آموزشی بخشی قابل توجهی از هزینه‌های آزمون «ویژگی‌های کلیدی» به دلیل جدید بودن این آزمون‌ها و عدم آشنایی استادان با نحوه طراحی آنان باید صرف توانمندسازی استادان شود.

علاوه بر صرف هزینه‌های مستقیم جهت برگزاری کارگاه‌های آموزشی توانمندسازی جهت آشناسازی اعضای هیأت علمی با شیوه طراحی آزمون ویژگی‌های کلیدی، بخشی از هزینه‌های آزمون به تدوین کلید پاسخ مشخص و نمره‌دهی آن مربوط می‌شود. معمولاً برای ساختن آزمون ویژگی‌های کلیدی که محتوای غنی دارند، زمان زیادی باید صرف شود. مشخص است که این زمان تا حدود زیادی به تجربه و مهارت طراح سؤال بستگی دارد. به عنوان مثال، برای طراح سؤال با تجربه حدود ۳۰ دقیقه اما برای کسانی که تجربه چندانی ندارند، چهار الی پنج ساعت ممکن است زمان نیاز باشد.

تأثیر آموزشی

به نظر می‌رسد نحوه درس خواندن دانشجویان متأثر از نوع آزمونی است که از آنان به عمل می‌آید. از آنجا که در آزمون «ویژگی‌های کلیدی»، امکان سنجش توانایی استدلال بالینی فراگیران به صورت طرح سناریو وجود دارد، انتظار می‌رود تأثیر آموزشی این دسته از آزمون‌ها مطلوب باشد. به همین دلیل در دانشکده‌های پزشکی به طور معمول از این شکل از آزمون‌ها با هدف یاددهی استفاده می‌شود (داست و همکاران[4] ۱۹۹۸، استارمبرگ و همکاران[5] ۲۰۰۳).

فیشر و همکاران ۲۰۰۵

این مطالعه با هدف ارزیابی مهارت استدلال بالینی دانشجویان پزشکی با استفاده از آزمون «ویژگی‌های کلیدی» انجام شد. ۱۵ سؤال بر اساس راهنمای طراحی آزمون «ویژگی‌های کلیدی» ارائه شده توسط انجمن پزشکی کانادا، طراحی و سپس توسط چهار پزشک که دارای سابقه بالا در طب داخلی بودند، بررسی شد. همچنین به منظور همبستگی نتایج حاصل از آزمون ویژگی‌های کلیدی با سؤالات چند گزینه‌ای (۴۰ سؤال) مورد بررسی قرار گرفت. ۳۷ دانشجوی پزشکی سال پنجم به صورت داوطلبانه در این مطالعه شرکت کردند. میزان پذیرش دانشجویان از این روش ارزیابی با استفاده از پرسشنامه‌ای (شامل دو سؤال تشریحی و ۲۲ سؤال با مقیاس لیکرت پنج‌تایی) بررسی شد.
بر اساس نتایج حاصل از مطالعه، پایایی آزمون ۰/۶۵ گزارش شد. سطح دشواری سؤالات بین ۰/۳ تا ۰/۸ بود. همبستگی بین نتایج حاصل از آزمون «ویژگی‌های کلیدی» با آزمون‌های چندگزینه‌ای بین ۰/۴۴ تا ۰/۴۷ گزارش گردید. همچنین پذیرش دانشجویان نسبت به این ابزار ارزیابی در حد متوسط بود. نویسندگان نتیجه گرفتند که آزمون «ویژگی‌های کلیدی» یک ابزار ارزیابی با قابلیت اجرای بالا است که از پایایی خوبی نیز برخوردار است.

1. American Society of Colon and Rectal Surgeons (ASCRS)
2. Royal Australian College General Practitioners (RACGP)
3. Fischer et al.
4. Doucet et al.
5. Sturmberg et al.

منابع

1. Ali SK, Bordage G. Validity of Key Features for a Family Medicine Pilot Exam at the College of Physicians and Surgeons Pakistan. J Coll Phys Surg Pakistan 1995; 5(6):256-60
2. Bordage G, Page G. An Alternative to PMPs: The "Key Features" Concept. Further Developments in Assessing Clinical Competence, 2nd Ottawa Conference 1987; 59-75.
3. Bordage G, Brailovsky C, Carretier H, Page G. Content Validation of Key Features on a National Examination of Clinical Decision-making Skills. Ac Med 1995; 70:276-81.
4. Bordage G, Carretier H, Bertrand R, Page G. Comparing Times and Performances of French and English speaking Candidates Taking a National Examination of Clinical Decision making Skills. Acad Med 1995; 70:359-365.
5. Charlin B, Desaulniers M, Gagnon R, Blouin D, Van der Vleuten C. Comparison of an aggregate scoring method with a consensus scoring method in a measure of clinical reasoning capacity. Teach Learn Med 2002, 14:150-156
6. Eva, KW, Wood, TJ, Riddle, J, Touchie, C, Bordage, G. How clinical features are presented matters to weaker diagnosticians. Med Educ 2010; 44: 775-85.
7. Farmer EA, Hinchy J. Assessing general practice clinical decision-making skills: the key features approach. Aust Fam Physician 2005; 34:1059-61.
8. Farmer EA, Page G. A Practical Guide to Assessing Clinical Decision-Making Skills using the Key Features approach. Medical Education 2005; 39: 1188-1194.
9. Fischer MR, Kopp V, Holzer M, Ruderich F, Junger J. A modified electronic key feature examination for undergraduate medical students: validation threats and opportunities. Med Teach 2005; 27:450 -55.
10. Hatala R, Norman GR. Adapting the key feature examination for a clinical clerkship. Med Educ 2002; 36: 160 -65.
11. Hurtz, Gregory M., et al. Measuring clinical decision making: do key features problems measure higher level cognitive processes?. Evaluation & the health professions 2012; 35(4) 396-415
12. Kwa SK, Sheikh Mohd A, AC Ang. Avoiding common errors in key feature problems. Malaysian family physician: the official journal of the Academy of Family Physicians of Malaysia 2007: 18.
13. Norman G, Bordage G, Page G, Keane D. How Specific is Case Specificity? Med Educ 2006; 40:618-23.
14. Page G, Bordage G. The Medical Council of Canada›s Key Feature Project: A More Valid Written Exam. of Clinical Decision-making Skills. Acad. Med 1995; 70: 104-110.
15. Page G, Bordage G, Allen T. Developing Key-Feature Problems and Examinations to Assess Clinical Decision-making Skills. Acad. Med 1995; 70: 194-201.

16. Page G, Boulais AP, Blackmore D, Dauphinee D. Justifying the Use of Short Answer Questions in the KF Problems of the MCCC's Qualifying Exam. In: Proceedings of the 9th Ottawa Conference, Cape Town. 2000.
17. Schaper E, Tipold A, Ehlers JP. Use of key feature questions in summative assessment of veterinary medicine students. Irish veterinary journal 2013; 66(1):1.
18. Trudel J, Bordage G, Downing S. Reliability and validity of Key Feature Cases for the Self-assessment of Colon and Rectal Surgeons. Ann Surg 2008; 248(2):252-8.
۱۹. امینیم، کاظم پور ر، مقدمیم، لطفی ف، ابوالفتحی ا. مقایسه نتایج آزمون نکات کلیدی با آزمون چهارجوابی پایان دوره کارورزی در بخش داخلی دانشگاه علوم پزشکی شیراز . مجله پزشکی هرمزگان، ۱۳۹۲؛ دوره ۱۷، شماره ۳: ۲۷۲-۲۵۶.
۲۰. منجمی ع. استدلال بالینی: مفاهیم، آموزش و ارزیابی. انتشارات دانشگاه علوم پزشکی اصفهان، ۱۳۸۹، صفحه ۱۰۸.

فصل ۱۵

آزمون «همخوانی با شرح‌نامه»

ساختار آزمون «همخوانی با شرح‌نامه»

آزمون «همخوانی با شرحنامه»، یک آزمون مبتنی بر مورد[1] است. مورد بالینی در این آزمون به صورت سناریوی کوتاهی مطرح می‌شود که مطرح‌کننده یک موقعیت پیچیده و مبهم است و بنابراین راه‌های متعددی را برای فرضیه‌های تشخیصی یا درمانی باز می‌گذارد. در ادامه، تعدادی سؤال پرسیده می‌شود که هر کدام سه بخش دارند (شکل ۱-۱۵):

- بخش اول «اگر شما به .... فکر کردید» است که شامل گزینه تشخیصی یا درمانی مرتبط است.
- سپس بخش دوم یعنی «و بعداً به این یافته در بیمار برمی‌خورید» می‌آید که به یافته‌ای در شرح‌حال، معاینه یا پاراکلینیک اشاره دارد.
- بخش سوم شامل مقیاس لیکرت پنج‌تایی است. وظیفه آزمون‌دهندگان آن است که تأثیر یافتهٔ جدید را بر گزینه تشخیصی یا درمانی (مثبت یا منفی) و شدت تأثیر (۰-۱-۲) را مشخص کنند.

در واقع در متن سؤالات این آزمون، فرضیه‌هایی ارائه می‌شوند و آزمون‌دهنده باید وزن و اهمیت داده‌ها را در شرحنامه بیماری بر اساس سناریو بسنجد. به همین دلیل این آزمون، روشی مناسب برای سنجش ارزیابی فرضیه‌ها است (منجمی ۱۳۸۹). در این آزمون معمولاً ۲۰ سناریو مطرح می‌شود که ذیل هر سناریو سه سؤال می‌آید یعنی نهایتاً آزمون‌دهنده باید ۶۰ سؤال را در مدت زمانی حدود یک ساعت پاسخ دهد.

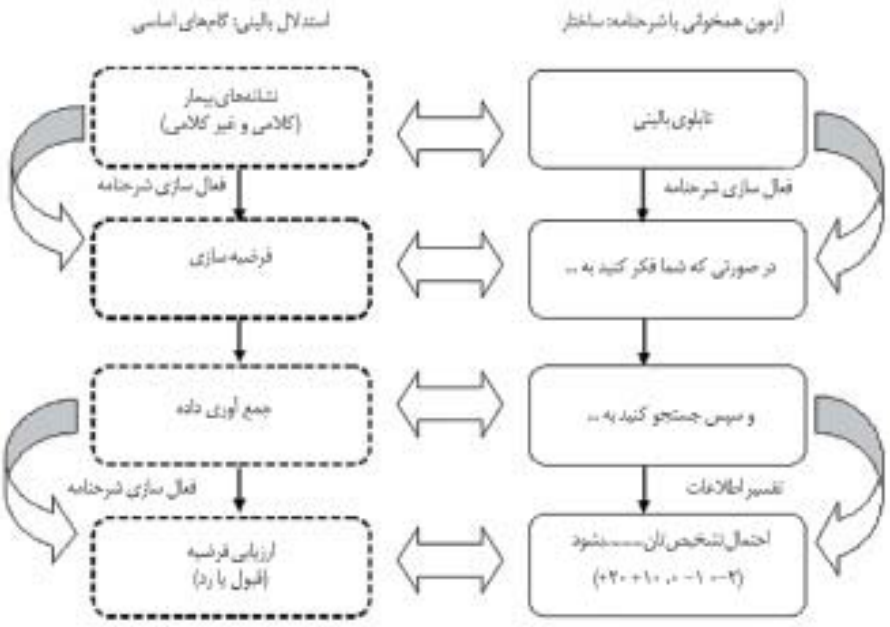
شکل ۱-۱۵: اجزای آزمون همخوانی با شرحنامه

1. Case-based

دوم آن که آزمون «همخوانی با شرحنامه» از سه ویژگی اساسی برخوردار است:

- در این آزمون داوطلبان با موقعیت‌های بالینی واقعی چالش‌زا مواجه می‌شوند که از طریق چندین گزینه امکان‌پذیر است. سؤال پرسیده شده باید با سناریو ارتباط داشته باشد و بدون آن قابل پاسخ‌دهی نباشد. به عبارت دیگر اگر بتوان بدون نیاز به خواندن سناریو به سؤال پاسخ داد، مشخص می‌شود که این سؤال معلومات را می‌سنجد، نه استدلال را.
- پاسخ‌ها بر اساس مقیاس لیکرت ارائه می‌شوند که منعکس‌کننده نظریه استدلال بالینی شرحنامه است. معنای اعداد ۲+ تا ۲– باید کاملاً برای آزمون‌شونده توضیح داده شود. این که این پرسش در مورد تشخیص است یا اداره بیمار باعث می‌شود که معنای این اعداد تغییر کند.
- شیوه تعیین نمره‌دهی به روش تجمیعی[1] است.

در شکل شماره ۲–۱۵ ساختار آزمون «همخوانی با شرحنامه» بر اساس فرایند استدلال بالینی به صورت شماتیک نشان داده شده است.



شکل ۲-۱۵. ارتباط بین مراحل اساسی استدلال بالینی با ساختار آزمون همخوانی با شرحنامه

مزایا و محدودیت‌های آزمون «همخوانی با شرح‌نامه»

مزایای «آزمون همخوانی با شرحنامه»

- **نمره‌دهی مناسب:** یکی از ویژگی‌های منحصر به فرد آزمون «همخوانی با شرحنامه» شیوه نمره‌دهی آن است که در آن طیف وسیعی از پاسخ‌های احتمالی در آن در نظر گرفته می‌شود.
- **امکان ارزیابی موقعیت‌های بالینی پیچیده:** یکی دیگر از ویژگی‌های آزمون «همخوانی با شرحنامه» توانایی این آزمون در سنجش موقعیت‌های پیچیده بالینی است. با استفاده از این آزمون می‌توان عملکرد داوطلبان را در مواجهه با

1. Aggregate method

موارد بالینی که یک تصمیم قطعی در مورد آن امکان‌پذیر نیست، سنجید. به علاوه شیوه نمره‌دهی این آزمون امکان ارائه پاسخ‌های متنوع را زمانی که پزشکان با موقعیت‌های پیچیده مواجه می‌شوند، فراهم می‌سازد.

- **امکان طراحی شکل‌های متنوعی از آزمون:** یکی از ویژگی‌های آزمون «همخوانی با شرحنامه» این است که علاوه بر روش کاغذی، امکان برگزاری آزمون به صورت الکترونیکی وجود دارد که می‌تواند قابلیت اجرای آن را بهبود بخشد.
- **امکان ارزیابی سازماندهی اطلاعات:** هر چند در مطالعات متعدد، آزمون «همخوانی با شرحنامه» به عنوان یک ابزار جهت ارزیابی مهارت استدلال بالینی معرفی شده است، با این حال بهتر است گفته شود که تمرکز این ابزار بر سنجش فرضیه‌های تشخیصی در زمینه استدلال بالینی است.

محدودیت‌های آزمون «همخوانی با شرحنامه»

- **طراحی دشوار:** طراحی آزمون همخوانی با شرحنامه به دلیل عدم آشنایی اعضای هیأت علمی با این آزمون، بسیار دشوار و فرایندی زمان‌بر است. طراحی این آزمون نیازمند تجربه و تمرین فراوان است.
- **نیاز به حضور تعداد قابل ملاحظه متخصصان در گروه خبرگان:** به دلیل پیچیدگی فرایند تعیین پاسخ‌های مجاز و حدنصاب قبولی آزمون همخوانی با شرحنامه، حضور تعداد زیادی از متخصصان ضروری است.

مراحل طراحی آزمون «همخوانی با شرح‌نامه»

همانند سایر ابزارهای ارزیابی، طراحی آزمون «همخوانی با شرحنامه»، با توجه به اهداف مورد انتظار (ارزیابی تکوینی یا آزمون‌های مهو سرنوشت‌ساز و ارائه گواهی و ...)، گروه‌های هدف (دانشجویان، دستیاران یا کارورزان، متخصصان و ...) و حیطه دانشی (جراحی، گوارش، پرستاری، اخلاق و ...) طراحی می‌شود. در ادامه سعی می‌شود تا بر اساس راهنماهای موجود، گام‌های طراحی آزمون «همخوانی با شرحنامه» با ذکر مثال‌های مرتبط در هر مرحله به طور مفصل توضیح داده شود (جدول ۱-۱۵).

جدول ۱-۱۵: خلاصه مراحل طراحی آزمون «همخوانی با شرحنامه»

ردیف	عنوان مرحله	توضیح
۱	انتخاب مورد بالینی	هنگام انتخاب موضوع، روی یک مشکل مهم بالینی باید متمرکز شد. برای این کار می‌توان فهرستی از مشکلات بالینی مهمی که فراگیران در آینده با آن‌ها مواجه خواهند شد، تهیه کرد.
۲	انتخاب محور مناسب	محور موقعیتی است که تعیین می‌کند چه اطلاعاتی (تشخیص، مداخله، گزینه‌های درمانی) باید در سؤال ارائه شود.
۳	نگارش سناریو (تابلوی بیماری)	سناریوی طراحی شده باید به گونه‌ای تنظیم شود که هیچ پاسخ مشخص و مستقیمی به آن مترتب نباشد و برای پاسخ‌دادن به آن فراگیر باید خود را در شرایط عدم قطعیت تلقی کند.
۴	تدوین سؤالات	هر سؤال شامل سه بخش است. در بخش اول این عبارت که «اگر شما به .... فکر کردید»، ارائه می‌شود که شامل یک گزینه تشخیصی یا درمانی مرتبط است. بخش دوم این عبارت که «و بعدا به این یافته در بیمار برمی‌خورید» می‌آید که به یک یافته جدید در شرح‌حال، معاینه یا پاراکلینیک اشاره دارد. در نهایت بخش سوم، مقیاس لیکرت پنج‌تایی است.
۵	تعیین شیوه نمره‌دهی	از آنجا که یک پاسخ صحیح منفرد وجود ندارد، روش نمره‌دهی در این آزمون از نوع تجمیعی است. به گونه‌ای که نظرات متفاوت متخصصان در مواجه با یک موقعیت بالینی مشخص پوشش داده می‌شود. بنابراین در این آزمون، نمره‌دهی توسط تیمی از متخصصان انجام می‌شود.
۶	مرور سؤال	

انتخاب مورد بالینی

اولین نکته‌ای که باید در طراحی آزمون «همخوانی با شرحنامه» در نظر گرفته شود، تعیین شرحنامه‌ای است که قصد ارزیابی آن وجود دارد. همان‌طور که در فصل پیشین نیز اشاره شد، جهت تعیین تابلوی بالینی موردنظر می‌توان از فهرست مشکلات بالینی مهمی که فراگیران در آینده با آن‌ها مواجه می‌شوند و بلوپرینت دوره استفاده نمود. به این ترتیب می‌توان سؤالاتی را طراحی نمود که از لحاظ بالینی مهم و مرتبط با آینده حرفه‌ای فراگیران است.

انتخاب محور مناسب

محور سؤال تعیین کننده دامنه‌ای از موضوع است که قصد ارزیابی آن وجود دارد. محور سؤال در آزمون «همخوانی با شرحنامه» می‌تواند پرسش در مورد تشخیص یک بیماری، مداخلات آزمایشگاهی یا سودمندی برنامه درمانی باشد. در زیر مثال‌هایی از سؤالات در محورهای مختلف ارائه شده است.

موضوع: سرگیجه — محور: تشخیص
گروه هدف: میانسالی

وخانم ۵۸ ساله‌ای با سابقه سرگیجه متناوب طی دو هفته اخیر به بخش اورژانس بیمارستان مراجعه می‌کند. او در شرح‌حال ذکر می‌کند که در بین سرگیجه‌ها احساس بهتری دارد.

اگر به تشخیص .... فکر کنید	و یافته ... وجود داشته باشد	احتمال تشخیص .... می‌شود
حمله‌های خوش‌خیم موضعی سرگیجه	مدت زمان آخرین سرگیجه ۳۰ دقیقه	۲+ ۱+ ۰ ۱- ۲-
حمله ایسکیمیک گذرا	سابقه فشارخون	۲+ ۱+ ۰ ۱- ۲-
سندرم منییر	جراحی اخیر ضایعات پوستی	۲+ ۱+ ۰ ۱- ۲-

موضوع: سر درد — محور: مداخلات آزمایشگاهی
گروه هدف: میانسالی

خانم ۳۳ ساله‌ای با سندرم تخمدان پلی‌کیستیک و سابقه افزایش فشار خون در حاملگی قبلی برای چک سردرد پس از زایمان ارجاع داده می‌شود. در معاینه بالینی، اختلالات بینایی و پارستزی بازوها وجود دارد. فشارخون وی در مطب ۱۸۰/۱۰۰ میلی‌متر است.

اگر به مداخله .... فکر کنید	و یافته ... وجود داشته باشد	احتمال سودمندی مداخله .... می‌شود
ونوگرافی رزونانس مغناطیسی	شدت یافتن سردرد بیمار با دراز کشیدن	۲+ ۱+ ۰ ۱- ۲-
آنالیز پروتیئن ۲۴ ساعته ادرار	سابقه زایمان طبیعی چهار هفته پیش	۲+ ۱+ ۰ ۱- ۲-
پونکسیون کمر	سابقه سزارین یک هفته پیش	۲+ ۱+ ۰ ۱- ۲-

موضوع: تنگی شریان کاروتید — محور: درمان
گروه هدف: میانسالی

خانم ۷۴ ساله با سابقه فشارخون بالا و مصرف هیدروکلروتیازید و آسپرین ۸۰ میلی‌گرم به صورت روزانه و تجربه لکنت زبان و بی‌حسی در دست چپ به مطب مراجعه می‌کند. داپلر کاروتید نشان‌دهنده ۹۰ درصد تنگی شریان کاروتید داخلی راست است.

اگر به گزینه درمانی .... فکر کنید	و یافته ... وجود داشته باشد	احتمال سودمندی درمان .... می‌شود
حمله‌های خوش خیم موضعی سرگیجه	مدت زمان آخرین سرگیجه ۳۰ دقیقه	۲+ ۱+ ۰ ۱- ۲-
حمله ایسکیمیک گذرا	سابقه فشارخون	۲+ ۱+ ۰ ۱- ۲-
سندرم منییر	جراحی اخیر ضایعات پوستی	۲+ ۱+ ۰ ۱- ۲-

نگارش سناریو (تابلوی بیماری)

همان‌طور که در مبحث آزمون ویژگی‌های کلیدی اشاره شد، ارائه سناریو موجب تقویت توانایی داوطلب در حل مسائل بالینی می‌شود. جهت نگارش تابلوی بالینی در آزمون طراح می‌تواند متداول‌ترین موقعیت بالینی را که اخیراً در عملکرد بالینی با آن مواجه شده است، انتخاب نماید. سناریو معمولاً در چند جمله، یک موقعیت بالینی چالشی را به تصویر می‌کشد، طوری که پزشکان باتجربه معمولاً نمی‌توانند صرفاً یک پاسخ صحیح و مشخص به آن بدهند؛ چرا که اطلاعات کافی در اختیارشان نیست یا نسبت به راه‌های درمانی و اقدامات تشخیصی بین افراد مختلف توافق‌نظر وجود ندارد. نکته حائز اهمیت آن است که موارد بالینی انتخابی نباید به شکلی ارائه گردند که صرفاً توانایی آزمون‌شونده را در توصیف ویژگی‌های مربوط به یک بیماری بسنجند. جهت کسب اطلاعات بیشتر در این خصوص به فصل دوم همین بخش مراجعه نمایید.

موضوع: درد شکم محور: تشخیص
گروه هدف: میانسالی

خانم ۵۴ ساله‌ای ساعت چهار صبح با شکایت درد شکم به اورژانس مراجعه کرده است. او چند سال پیش به خاطر دردهای شکمی متناوب و مشابه، ولی کمی خفیف‌تر در بیمارستان بستری گردیده بود. اخیراً میزان این تناوب به سه بار در هفته رسیده است که ماهیتی ناپیوسته داشته و به نیمه فوقانی شکم محدود بوده است و گه‌گاه به پشت انتشار داشته است. درد فعلی بیمار به مدت دو روز ادامه داشته که بر شدت آن نیز افزوده شده است و با تهوع و استفراغ همراه است. بیمار روزانه ۶ نخ سیگار مصرف می‌کند و شرح‌حالی از مصرف الکل نمی‌دهد. در معاینه بیمار چاق است و شدیداً عرق کرده است. ضربان قلب ۹۶ در دقیقه و فشار خون ۱۱۰/۷۵۰ میلی‌متر جیوه است. کبد پنج سانتی‌متر زیر لبه دنده ملموس است. حساسیت قابل ملاحظه‌ای در لمس شکم وجود دارد که همراه با گاردینگ در ناحیه اپی‌گاستر است. صداهای روده‌ای قابل شنیدن نیست. معاینه رکتوم طبیعی است.

به طور کلی هنگام طراحی سناریو باید به موارد زیر توجه داشت:

- یکی از ویژگی‌های آزمون «همخوانی با شرحنامه»، ارزیابی عملکرد داوطلبان در موقعیت‌های چالش‌زای بالینی است. سناریو باید به گونه‌ای تنظیم شود که هیچ پاسخ مشخص و مستقیمی به آن مترتب نباشد و برای پاسخ‌دادن به آن فراگیر باید خود را در شرایط عدم قطعیت تلقی کند.
- هنگام نگارش سناریوی بالینی، بهتر است تصور شود که یک بیمار حقیقی معرفی می‌شود. بدین ترتیب سناریوی طراحی شده، به موقعیت‌های واقعی عملکرد داوطلبان نزدیک‌تر خواهد شد.
- باید مطمئن شد که اطلاعات مندرج در سناریو، واضح، روان و بدون ابهام است. وجود ابهام در سناریو علاوه بر آن که فراگیر را از مسیر صحیح پاسخ‌دهی به سؤال منحرف می‌کند، منجر به صرف وقت بیشتر برای پاسخ‌دهی به آن خواهد شد که این موضوع بر پایایی آزمون تاثیرگذار است.
- همواره باید سعی شود که زبان مورد استفاده در طرح سناریو، متناسب با گروه داوطلبان موردنظر باشد.
- نکته آخر اینکه هنگام نوشتن سناریو در عین اینکه اطلاعات ضروری شرح داده می‌شوند، باید از ارائه مطالب غیرضروری خودداری کرد.

تدوین سؤالات

همان‌طور که پیشتر اشاره شد، در آزمون «همخوانی با شرحنامه» برای هر مشکل بالینی، معمولاً سه سؤال طرح می‌شود و هر سؤال، خود، شامل سه بخش است که در ادامه جزئیات مربوط به هر بخش به تفصیل شرح داده می‌شود.

- **بخش اول: تدوین گزینه‌ها (تشخیصی یا درمانی):** این مرحله اطلاعات مربوط به ستون اول را فراهم می‌سازد و به شکل «اگر شما به .... فکر کردید» نوشته می‌شود. با توجه به محور انتخابی، این گزینه‌ها ممکن است تشخیصی یا درمانی باشند. گزینه‌ها در واقع فرضیه‌هایی مرتبط با موقعیت بالینی موردنظر هستند که دانشجو باید صحت آنها را بررسی کند. این اطلاعات به فعال‌سازی شرحنامه بیماری در ذهن داوطلب کمک می‌کند. نکته‌ای که باید در این

مرحله در نظر داشت این است که فرضیه‌های تشخیصی/ درمانی باید منطقی انتخاب شوند (به عنوان مثال، داوطلب باید احساس کند در آن موقعیت بالینی مشخص چنین ملاحظات درمانی/ تشخیصی وجود دارد). به عنوان مثال فرضیه‌های تشخیصی برای سناریوی مثال بالا می‌تواند به این صورت طراحی شود.

فرضیه‌های تشخیصی
اگر به گزینه تشخیصی .... فکر کنید
انفارکتوس میوکارد
کبد چرب
کولیت اولسروز

- **بخش دوم: تعیین اطلاعات بالینی که به تصمیم‌گیری بهتر کمک می‌کنند:** این بخش اطلاعات ستون دوم جدول را فراهم می‌سازد و منعکس‌کننده عملکرد داوطلب در فرایند جمع‌آوری اطلاعات است. این بخش به صورت «و بعداً به این یافته در بیمار برمی‌خورید» نوشته می‌شود و به یک یافته جدید در شرح‌حال، معاینه یا پاراکلینیک اشاره دارد. توجه داشته باشید اطلاعاتی که در این بخش قرار می‌گیرد، باید به تأیید یا رد اطلاعات ستون قبلی کمک کند. در جدول زیر مثال‌هایی از یافته‌ها برای تشخیص‌های افتراقی ارائه شده در بالا آورده شده است.

جمع‌آوری اطلاعات	
اگر به گزینه تشخیصی ... فکر کنید	**و یافته ... وجود داشته باشد**
انفارکتوس میوکارد	تهوع و استفراغ
کبد چرب	کبد بزرگ
کولیت اولسروز	ماهیت ناپیوسته درد

- **بخش سوم: تصمیم‌گیری در خصوص مقیاس لیکرت:** بخش سوم شامل مقیاس لیکرت است. وظیفه آزمون‌دهندگان آن است که تأثیر یافتهٔ جدید را بر گزینه تشخیصی یا درمانی (مثبت یا منفی) و شدت تأثیر (۰-۱-۲) را مشخص کنند.

جمع‌آوری اطلاعات		
اگر به گزینه تشخیصی ... فکر کنید	**و یافته ... وجود داشته باشد**	**احتمال تشخیص درمان ... می‌شود.**
انفارکتوس میوکارد	تهوع و استفراغ	۲+ ۱+ ۰ -۱ -۲
کبد چرب	کبد بزرگ	۲+ ۱+ ۰ -۱ -۲
کولیت اولسروز	ماهیت ناپیوسته درد	۲+ ۱+ ۰ -۱ -۲

در آزمون «همخوانی با شرحنامه» معمولاً استفاده از لیکرت پنج‌تایی متداول است. هر چند در این خصوص اختلاف نظرهایی در بین متخصصان وجود دارد. دو مطالعه به مقایسه استفاده از لیکرت پنج‌تایی و سه‌تایی پرداختند که نتایج متناقضی در ارتباط با تأثیر انتخاب لیکرت بر پایایی آزمون به دست آوردند (بلاند و همکاران[1] ۲۰۰۵، رامیکرز و همکاران[2]

1. Bland et al.
2. Ramaekers

۲۰۱۰). در مورد استفاده از مقیاس‌های لیکرتی، اشاره به این نکته ضروری است که لیکرت هفت‌تایی (۳- تا ۳+) برای موقعیت‌هایی مناسب است که آزمون رقابتی است و میزان تفکیک و تمایز در آزمون اهمیت دارد. در مقابل، وقتی آزمون را با اهداف آموزشی طراحی می‌کنیم، لیکرت سه‌تایی (۱- تا ۱+) نیز کفایت می‌کند. به عنوان مثال، از لیکرت سه‌تایی برای موضوعاتی پیرامون آموزش مداوم حرف سلامتی می‌توان استفاده نمود (پترلا و دیویس[1] ۲۰۰۷). همچنین استفاده از لیکرت سه‌تایی برای ارزیابی مهارت دانشجویانی که در سطح مبتدی قرار دارند، منطقی است (کلی و همکاران[2] ۲۰۱۲). نکته حائز اهمیت در این خصوص آن است که برای هر کدام از سؤالات آزمون «همخوانی با شرحنامه»، توصیفات لیکرت بر اساس نوع محور سؤال متفاوت خواهد بود. در جدول ۲-۱۵ سه دسته کلی (تشخیص، مداخله و درمان) و توصیفات مربوط به آن‌ها ارائه شده است.

جدول ۲-۱۵: محورهای سؤالات در آزمون «همخوانی با شرحنامه» و تفاسیر مربوطه

تفسیر	عبارت توصیفی	محور سؤال
تشخیص	اگر به تشخیص ... فکر کنید و یافتهٔ ... در بیمار وجود داشته باشد، تشخیص شما ...	۲+ قویاً تأیید می‌شود ۱+ تأیید می‌شود ۰ تقویت یا تضعیف نمی‌شود ۱- تضعیف می‌شود ۲- قویاً تضعیف می‌شود
سودمندی مداخله	اگر به اقدام .... فکر کنید و یافتهٔ ... در بیمار وجود داشته باشد، این اقدام	۲+ بسیار سودمند است. ۱+ سودمند است. ۰ کمابیش سودمند نیست. ۱- سودمند نیست. ۲- قویاً سودمند نیست.
خطر/ منفعت اقدام درمانی	اگر به درمان با .... فکر کنید و یافته ... در بیمار وجود داشته باشد، آن اقدام درمانی ...	۲+ قویاً کنتراندیکاسیون دارد. ۱+ اندیکاسیون دارد ۰ کمابیش اندیکسیون ندارد ۱- کنتراندیکاسیون دارد ۲- کنتراندکاسیون مطلق دارد.

باید توجه شود که گاهی پیش می‌آید طراحان سؤال براساس اهداف ارزیابی و گروه‌های هدف تغییراتی را در ساختار سؤالات آزمون «همخوانی با شرحنامه» ایجاد می‌کنند. در هر حال، اصول زیربنایی شرح داده شده یکسان خواهد بود. در ادامه، نمونه‌ای از سؤالات طراحی شده در رشته رادیولوژی دانشگاه شربروک [3] ارائه شده است.

موضوع: ندول ریوی محور: تشخیص

در رادیوگرافی قفسه سینه، ندول ریوی در بخش قدامی لوب فوقانی چپ دیده شده است. فهرست تشخیص‌های احتمالی شامل موارد زیر است:

- کانسر ریه
- آبسه ریوی
- هیستوپلاسموز
- ندول روماتوئید
- متاستاز منفرد ریوی

در تفسیر رادیوگرافی قفسه سینه، علایم زیر شرح داده شده است:

1. Petrella & Davis
2. Kelley et al.
3. University of Sherbrooke

سؤال ۱: کلسیفیکاسیون در ندول مشاهده نمی‌شود. این یافته چه تاثیری در تأیید یا رد تشخیص‌های احتمالی دارد؟							
کانسر اولیه ریه	-۳	-۲	-۱	۰	+۱	+۲	+۳
هیستوپلاسموز	-۳	-۲	-۱	۰	+۱	+۲	+۳
متاستاز منفرد ریوی	-۳	-۲	-۱	۰	+۱	+۲	-۳
سؤال ۲: ندول حالت حفره مانند نیست. این یافته چه تاثیری در تأیید یا رد تشخیص‌های احتمالی دارد؟							
متاستاز منفرد ریوی	-۳	-۲	-۱	۰	+۱	+۲	+۳
ندول روماتوئید	-۳	-۲	-۱	۰	+۱	+۲	+۳
آبسه ریوی	-۳	-۲	-۱	۰	+۱	+۲	-۳
برگرفته از برازیو- لامونتاگن (۲۰۰۴) دانشگاه شربروک							

به طور کلی هنگام طراحی این سؤالات باید به موارد زیر توجه داشت:

- هر سؤال فقط در مورد یک داده باشد (جهت کسب اطلاعات بیشتر به بحث طراحی گزینه در آزمون «ویژگی‌های کلیدی» مراجعه کنید).
- سؤالات باید به سناریو متصل باشند و بدون خواندن سناریو نشود به آن‌ها پاسخ داد.
- گزینه‌های انتخاب شده، حتماً نکات کلیدی آن مورد بالینی باشند. به عبارت دیگر می‌توان گفت، یافته‌ها و تشخیص‌هایی که در بدنهٔ سؤال مورد استفاده قرار می‌گیرند، همگی باید از اهمیت کلیدی برخوردار باشند. ذکر اطلاعات جزئی و فارغ از اهمیت، علاوه بر آنکه اثرات آموزشی منفی بر نحوه یادگیری دانشجویان به دنبال خواهد داشت، تخمین درستی از عملکرد داوطلبان نیز فراهم نخواهد شد.
- سؤالات باید گستره زیادی از پاسخ‌های درست احتمالی را پوشش دهند. سؤالتی که در آن توافق و اجماع روی یکی از گزینه‌ها وجود دارد، قدرت تمایز و تفکیک پایینی دارد. اگر سؤالی طراحی شد که همهٔ اعضای گروه خبرگان یک گزینه را انتخاب کردند (از میان ۲- تا ۲+) این سؤال باید از آزمون حذف شود، چرا که احتمالاً سؤالی است که تنها سطح دانشی داوطلبان را می‌سنجد. از سوی دیگر، سؤالی که گسترهٔ پاسخ‌ها در آن بسیار زیاد است مثلاً تمام پنج گزینهٔ ممکن برای آن انتخاب شده بود (از ۲- تا ۲+)، نیز سودمند نیست و بهتر است حذف شود. بنابراین باید توجه شود که یافته‌های بدیهی کاملاً غلط یا درست را در سؤال نگنجاند چرا که در این شرایط داوطلب بدون نیاز به خواندن سناریو، به راحتی نادرست بودن یا صحیح بودن آن را تشخیص می‌دهد.

تعیین شیوه نمره‌دهی

به طور کلی در آزمون «همخوانی با شرحنامه» پاسخ‌های صحیح متعددی برای یک سؤال وجود دارد که شناسایی آنها توسط تیمی از متخصصان صورت می‌گیرد. انتخاب گروه خبرگان یک مرحله اساسی در طراحی این آزمون به شمار می‌رود. جهت کسب اطلاعات بیشتر در این خصوص به فصل یک همین بخش مطالعه نمایید.

در سیستم نمره‌دهی آزمون «همخوانی با شرحنامه»، برای تک‌تک سؤالات، پاسخ‌ها توسط اعضای گروه خبرگان تعیین می‌شود. بنابراین روش متداول نمره‌دهی در این آزمون از نوع تجمیعی است که در آن نظرات متفاوت متخصصان در مواجه با یک موقعیت بالینی پوشش داده می‌شود (نورمن ۱۹۸۵، نورسینی و همکاران ۱۹۹۰). گزینه‌هایی که حداکثر تعداد اعضای گروه خبرگان آن را انتخاب نموده‌اند، به عنوان استاندارد طلایی[1] آن سؤال در نظر گرفته می‌شود و نمره کامل سؤال (یک نمره) به

1. Gold standard

آن گزینه تعلق می‌گیرد. از آنجا که سایر گزینه‌های انتخاب شده توسط دیگر اعضای گروه خبرگان نیز منعکس‌کننده تفاوت در تفسیرهای افراد از یک موقعیت بالینی است، آن‌ها نیز ارزشمند تلقی می‌شوند و بنابراین، بر اساس تعداد متخصصانی که آن گزینه را انتخاب نموده‌اند، بخشی از نمره به آن گزینه تعلق می‌گیرد (لابارسکی و همکاران ۲۰۱۳). به گزینه‌هایی که هیچ یک از متخصصان انتخاب ننموده‌اند، هیچ نمره‌ای تعلق نمی‌گیرد. به عنوان مثال، در صورتی که از تیم ۲۲ نفره متخصصان، ۱۷ نفر برای یک سؤال امتیاز ۱+ در نظر بگیرند، به این معناست که ۱+ جوابی است که حداکثر نمره یعنی یک نمره ($\frac{۱۷}{۱۷}$) را به خود اختصاص می‌دهد. در صورتی که داوطلب ۱+ را انتخاب کرده باشد، نمره یک به او تعلق می‌گیرد. را انتخاب نماید، نمره یک برای آن سؤال دریافت می‌کند. همچنین اگر پنج نفر از متخصصان امتیاز ۲+ را انتخاب نمایند، در صورت انتخاب ۲+ توسط دانشجو ۰/۲۹ نمره برای وی منظور می‌شود ($\frac{۵}{۱۷}$). به همین منوال به انتخاب‌های ۱–، ۲– و صفر داوطلبان هیچ نمره‌ای تعلق نمی‌گیرد. جدول شماره ۳–۱۵ نحوه نمره‌دهی را به صورت روشن‌تری نشان می‌دهد.

جدول ۳–۱۵: نمونه‌ای از سیستم نمره‌دهی آزمون «همخوانی با شرحنامه»

پاسخ	۲–	۱–	۰	۱+	۲+
تعداد متخصصانی که پاسخ را انتخاب کردند	۰	۰	۱	۵	۴
نمره	۰	۰	$\frac{۱}{۱۰}$	$\frac{۵}{۱۰}$	$\frac{۴}{۱۰}$
نمره تبدیلی	۰	۰	$\frac{۱}{۵}$	$\frac{۵}{۵}$	$\frac{۴}{۵}$
امتیاز هر سؤال	۰	۰	۰/۲	۱	۰/۸

با گذشت زمان، در مطالعات گوناگون سعی شد تا تغییراتی در شیوه نمره‌دهی ایجاد شود به این امید که پایایی نمرات این نوع آزمون بهبود یابد. به این ترتیب روش‌های جایگزین متعددی از قبیل تعیین وزن گزینه‌ها بر اساس تعداد کل انتخاب‌های گروه خبرگان، استفاده از میانگین نمرات و فاصله از پاسخ طلایی معرفی شد که در ادامه به شرح مختصر این روش‌ها خواهیم پرداخت. هرچند نتایج حاصل از برخی مطالعات نشان می‌دهد زمانی که از روش سنتی نمره‌دهی استفاده می‌شود، پایایی و شاخص تمیز آزمون مشابه و حتی در مواردی بهتر خواهد شد (رامیکرز و همکاران ۲۰۱۰)

- در «روش پرت»[1] ابتدا بر اساس پاسخ‌هایی که هریک از اعضای گروه خبرگان به سؤالات داده‌اند، نمره کل تک‌تک اعضا محاسبه می‌شود. سپس میانگین نمرات اعضای گروه محاسبه می‌شود. متخصصانی که نمره کل آنها از دو انحراف معیار نسبت به میانگین نمرات گروه خبرگان کمتر یا بیشتر باشد، از فهرست حذف می‌شوند و نظرات آنان در تهیه کلید پاسخ لحاظ نمی‌شود. بنابراین گزینه‌های صحیح مجاز در این روش بر اساس نظرات متخصصانی که نمره کل آنان در دامنه میانگین گروه خبرگان قرار دارد تعیین می‌شود.
- روش «فاصله از مد»[2] بر اساس شناسایی و حذف پاسخ‌هایی است که خارج از پاسخ متداول انتخاب شده توسط گروه خبرگان قرار دارند. به عنوان مثال، برای یک سؤال مشخص اگر استاندارد طلایی ۱+ باشد، متخصصانی که گزینه‌های ۲– یا ۲– و ۱– را انتخاب کرده باشند از فرایند نمره‌دهی آن سؤال حذف می‌شوند. بنابراین گزینه‌های صحیح مجاز در این روش بر اساس در نظر گرفتن نظرات متخصصانی که پاسخ متعارف به سؤال داشتند، تعیین می‌شود.
- در روش «قضاوت توسط متخصصان»[3] تمام متخصصان به صورت مستقل از یکدیگر تمام سؤالات را بررسی

1. Outlier method
2. Distance-from-mode method
3. Judgment-by-experts method

می‌کنند تا پاسخ‌هایی را که غیرقابل قبول هستند، شناسایی کنند. در این روش پاسخ‌هایی که حداقل توسط دو یا سه متخصص به عنوان غیرقابل قبول انتخاب شده بودند از فهرست کلید پاسخ خارج می‌شوند.

گانون و همکاران (۲۰۱۱) در پژوهشی به بررسی اثرات سه روش فوق در گروه خبرگان پرداختند و بر اساس نتایج حاصل از پژوهش به این نتیجه رسیدند که در پانل ۴۵ نفره متخصصان هیچ کدام از این سه روش بر پایایی نمرات آزمون و همبستگی بین میانگین نمرات تاثیری نداشت. همچنین تفاوت نمره بین پانل خبرگان و گروه دستیاران، بین پانل خبرگان و گروه دانشجویان و بین گروه دانشجویان و گروه دستیاران دیده نشد. به علاوه، هنگامی که تعداد اعضای پانل کاهش یافت (۱۵ نفر)، تنوع در روش‌های تصحیح کماکان موجب تفاوت معناداری در پایایی نمرات یا در همبستگی نمرات نشد اما تفاوت نمرات بین گروه‌ها معنادار بود. بر اساس نتایج حاصل، نویسندگان تعداد اعضای پانل نمره‌دهی را بیش از ۱۵ نفر توصیه کردند. به علاوه ذکر کردند که روش «فاصله از مد» بهتر می‌تواند نوسانات عملکردی بین نمرات پنل و دستیاران را مشخص نماید و به عنوان موثرترین استراتژی توصیه شد.

به طور معمول دو ویژگی کلیدی بر فرایند نمره‌دهی آزمون «همخوانی با شرحنامه» مترتب است. اولاً توجه به این نکته ضروری است که نمره حاصل از هر آزمون مستقیماً با سایر آزمون‌های مرتبط قابل مقایسه نیست. این موضوع به دلیل آن است که هر آزمون به سنجش عملکرد داوطلبان در انجام یک وظیفه خاص، در یک موقعیت مشخص و با حضور پانلی از متخصصان صورت می‌گیرد. تفاوت در آزمون‌های مختلف ممکن است به دلیل تأثیر عواملی از جمله سطح دشواری مشکل بالینی، ویژگی‌های داوطلبان و مشخصات گروه خبرگان باشد. دومین ویژگی در ارتباط با تعداد اعضای گروه خبرگان است. نتایج حاصل از مطالعات نشان می‌دهد که حضور حداقل ۱۵ متخصص در گروه خبرگان برای رسیدن به سطح پایایی قابل قبول نیاز است (گانون و همکاران ۲۰۰۵). جهت کسب اطلاعات بیشتر به فصل یک همین بخش مراجعه نمایید.

در نمره‌دهی این آزمون باید موارد زیر در نظر گرفته شوند:

- در این آزمون هیچ پاسخ صحیح یا نادرست مطلقی وجود ندارد و هر گزینه می‌تواند توسط متخصصان قابل قبول باشد و این در حالی است که در سؤالات چندگزینه‌ای، تنوع در پاسخ‌ها مطرح نیست و یک پاسخ صحیح منحصر به فرد وجود دارد (جهت کسب اطلاعات بیشتر به بخش دوم کتاب مراجعه کنید).
- نمره هر سؤال بر اساس مجموع نمرات ارائه شده توسط هر یک از اعضای گروه خبرگان تعیین خواهد شد.
- سهم نمره داوطلب از هر سؤال بر اساس نسبت بین پاسخ ارائه شده توسط وی با پاسخ‌های اعضای گروه خبرگان تعیین می‌شود.
- نمره کل آزمون داوطلب از جمع نمرات تک‌تک سؤالات فراگیر حاصل می‌شود.

پارک و همکاران[1] ۲۰۱۰

در این مطالعه نمره‌دهی سؤالات با استفاده از مجموع نظرات متخصصان زنان و زایمان با استفاده از مقیاس لیکرت صورت گرفت. نمره‌کلی آزمون نیز از جمع نمره هر سؤال حاصل شد. به منظور سهولت در تفسیر نتایج حاصل از آزمون، سطح نمرات تغییر یافت به گونه‌ای که حداکثر نمره حاصل از آزمون ۱۰۰ در نظر گرفته شد. به طور کلی آزمون نهایی مشتمل بر ۴۲ سؤال همخوانی با شرحنامه برای ۷۵ نفر از دستیاران اجرا شد. همسانی درونی (آلفای کرونباخ) ۰/۷۳ و پایایی بازآزمایی ۰/۷۶ گزارش شد.

1. Park et al.

مرور و بازبینی سؤال

پس از اینکه سؤال طراحی شد، مهم است که یک بار آن مرور شود تا مشخص گردد آیا نکات مطرح شده را در خصوص آن رعایت شده‌اند یا خیر، تادر صورت لزوم سؤال اصلاح شود. همچنین می‌توان از همکاران در خصوص سؤال طرح‌شده نظر خواست. علاوه بر آن در مورد کیفیت محتوایی سؤالات آزمون می‌توان از نظرات متخصصان کمیته ارزیابی نیز استفاده نمود.

همچنین در صورت اجرای آزمون به شکل پایلوت یا حتی دوره‌های قبلی اجرای آن، از طریق آنالیز پاسخ‌های داوطلبان به

سؤالات نیز می‌توان به تحلیل کیفیت سؤالات آزمون‌های موجود پرداخت. محاسبه ضریب همبستگی کل سؤالات، تخمینی از ظرفیت تمیز سؤالات آزمون فراهم می‌سازد و می‌توان از آن به عنوان گامی برای شناسایی سؤالات مشکل‌دار استفاده نمود. به عنوان یک قاعده کلی، سؤالاتی که ضریب همبستگی آن‌ها با کل آزمون منفی است و همچنین سؤالاتی که ضریب همبستگی آن‌ها با کل آزمون مثبت است اما میزان آن کمتر از ۰/۰۵ است، به این دلیل که به هیچ وجه یا به میزان کمی در پایایی نمرات آزمون تاثیرگذار هستند بهتر است از کل آزمون حذف گردند (کریر و همکاران[1] ۲۰۰۸). البته باید توجه داشت که در برخی موارد، همبستگی پایین دال بر وجود خطا در طراحی سؤالات آزمون نیست و می‌تواند معرف ناهمگونی صلاحیت بالینی اعضا گروه خبرگان یا تنوع موجود در حیطه‌های مورد ارزیابی باشد. قاعدتاً در چنین مواردی طراحان آزمون باید در خصوص حذف یا عدم حذف آن تصمیم بگیرد.

سودمندی آزمون همخوانی با شرحنامه

همان‌طور که به کرّات در این کتاب اشاره شده است، به منظور کسب شاخص سودمندی ابزار باید به چند شاخص در ابزار اندازه گیری توجه نمود: روایی، پایایی، تاثیرآموزشی، هزینه و مقبولیت روش ارزیابی (شوورث و ون‌درولوتن ۲۰۰۴). به عنوان مثال، در یک آزمون پذیرش فراگیر در دانشکده پزشکی، دغدغه اساسی مدیران داشتن یک ابزار با پایایی بسیار بالا می‌باشد و تأثیر آموزشی ابزار کمتر نگران کننده است اما در مقابل، ابزارهایی که برای ارزیابی کلاسی به کار می‌روند باید ارزش و اثر آموزشی بالایی در حمایت از یادگیری فراگیران داشته باشند و قاعدتاً در چنین موقعیت‌هایی ممکن است به پایایی چندان توجهی نشود.

روایی

یکی از مهمترین مباحث در هر ابزار ارزیابی، تعیین روایی محتوایی آن است. روایی محتوای نشان‌دهنده قابلیت ابزار در پوشش دادن محتوای دوره به شکل مناسب است و اغلب از طریق بلوپرینت تعیین می‌شود. در حقیقت بلوپرینت، اهدافی را که در یک آزمون مشخص باید مورد ارزیابی قرار گیرند و همین‌طور وزن نسبی آن‌ها را در آزمون مشخص می‌نماید. همان‌طور که پیشتر نیز اشاره شد، طراحی بلوپرینت مناسب، اولین گام ضروری در ایجاد یک ابزار روا و معتبر است و نباید نادیده انگاشته شود. آزمون «همخوانی با شرحنامه» نیز از این قاعده مستثنی نیست. به طور کلی روایی بالاتر سؤالات از طریق مواردی از قبیل میزان مرتبط بودن با اهداف آموزشی دوره، کیفیت ساختاری سؤالات طرح شده در آزمون، نگارش واضح سؤالات، مقبولیت گزینه‌های طراحی شده حاصل می‌شود. پس از برگزاری آزمون نیز باید روایی ابزار مورد سنجش قرار گیرد تا از میزان روایی آن اطمینان حاصل شود. مطالعات متعددی به بررسی روایی این سؤالات پرداخته‌اند که در ادامه به تعدادی از آن‌ها اشاره خواهیم نمود.

بررسی شواهد مرتبط با روایی سازه «آزمون همخوانی با شرحنامه» نشان می‌دهد که پژوهشگران در پی پاسخ به این سؤال بوده‌اند که آیا با استفاده از آزمون «همخوانی با شرحنامه» می‌توان بین عملکرد داوطلبان در سطوح مختلف آموزشی تفاوتی قائل شد. پیش‌فرض اکثر این مطالعات این بوده است که به دلیل تفاوت در توانایی مهارت‌های استدلال بالینی این ابزار باید بتواند بین عملکرد دانشجویان، دستیاران و متخصصان به خوبی تمیز قائل شود:

- در مطالعه گانون و همکاران (۲۰۰۵) رگرسیون خطی نمرات حاصل از اجرای آزمون در سطوح مختلف آموزشی، نشان‌دهنده روایی سازه ابزار مورد ارزیابی بود.
- ماری و همکاران[2] (۲۰۰۵) نیز پس از اجرای آزمون ۹۵ سؤالی برای دانشجویان پزشکی (۱۷ نفر)، دستیاران پزشک خانواده (۹ نفر)، دستیاران تخصص داخلی (۵ نفر) و متخصصان داخلی (۷ نفر) به این نتیجه رسیدند که تفاوت نمرات حاصل از اجرای آزمون بین گروه‌های مختلف از نظر آماری معنادار بود (p=/۰۰۰۱).

1. Carrière et al.
2. Marie et al.

- رویز و همکاران[1] (۲۰۱۰) نیز در مطالعه خود نشان دادند که بین عملکرد متخصصان طب سالمندان، فلوشیپ‌های سال آخر و اول، دستیاران و دانشجویان پزشکی در آزمون «همخوانی با شرحنامه» تفاوت معنادار وجود دارد. محققان اینگونه نتیجه گرفتند که آزمون «همخوانی با شرحنامه»، یک ابزار ارزیابی با روایی محتوایی و روایی سازه بالا است که به خوبی می‌تواند بین عملکرد پزشکان با تجربه و فاقد تجربه کافی تمیز قائل شود.
- نوح و همکاران[2] (۲۰۱۲)، بر اساس یافته‌های حاصل از اجرای آزمون «همخوانی با شرحنامه» روی ۲۰۲ دستیار در ۹ برنامه آموزشی جراحی عمومی در کانادا به این نتیجه رسیدند که این آزمون از توانایی تمیز بین عملکرد دستیاران در سال‌های مختلف آموزشی برخوردار است.

به صورت کلی، مطالعات متعدد صورت گرفته در رشته‌های مختلف از قبیل مامایی، پرستاری، جراحی، رادیولوژی، کاردرمانی و پزشک خانواده نشان‌دهنده بالابودن روایی سازه نمرات حاصل از این آزمون است (چارلین و همکاران ۱۹۹۸، برایلفسکی و همکاران[3] ۲۰۰۱، سیبرت و همکاران ۲۰۰۲، برازیو و همکاران[4] ۲۰۰۴، چارلین و ون‌درولوتن ۲۰۰۴، گانون ۲۰۰۵، گانون و همکاران ۲۰۰۶، کوهن و همکاران ۲۰۰۹، دیشنس و همکاران[5] ۲۰۱۱).

سیرت و همکاران ۲۰۰۵

در این مطالعه دو سری متوالی از آزمون «همخوانی با شرحنامه» به شکل الکترونیکی طراحی شد. یکی برای اعضای گروه خبرگان به منظور تعیین روایی محتوایی سؤالات آزمون و تدوین سیستم نمره‌دهی و دومی برای شرکت فراگیران با سطوح مختلف تخصصی شامل دانشجویان سال پنجم و ششم پزشکی، دستیاران و متخصصان اورولوژی. در این پژوهش ابتدا از دو عضو هیأت علمی خواسته شد تا شایعترین موقعیت‌های بالینی را که در اورلوژی وجود دارد مشخص نمایند و از آنان خواسته شد تا برای هر موقعیت: ۱). فرضیه‌های مرتبط، استراتژی‌های تشخیصی و گزینه‌های درمانی ۲). معاینات فیزیکی و آزمون‌های آزمایشگاهی، ۳). اطلاعات بالینی تقویت‌کننده یا ردکننده را مشخص نمایند.
سپس برای هر یک از داوطلبان به شکل انفرادی یک کد امنیتی در نظر گرفته شد که پیش از ورود به وب‌سایت باید آن را وارد می‌کردند. هر یک از داوطلبان ممکن بود در طول دوره آموزشی چندین بار در آزمون «همخوانی با شرحنامه» شرکت کنند. در طول شش ماه، ۸۰ درصد از اورولوژیست‌ها، ۶۸ درصد از دستیاران و ۲۰ درصد از دانشجویان پزشکی این آزمون را پشت سر گذاشتند. میانگین نمرات آزمون برای دانشجویان ۴/۱۸±۵۳/۱۷، برای دستیاران ۴/۷۸ ± ۵۷/۴۲ و برای متخصصان اورلوژی ۶/۶۸ ± ۶۲/۵۲ بود. نتایج حاصل از مطالعه نشان‌دهنده افزایش میانگین نمرات آزمون با افزایش سطح آموزشی است.
محققان نتیجه گرفتند که اجرای شکل الکترونیکی آزمون «همخوانی با شرحنامه» از قابلیت اجرای بالای برخوردار است به گونه‌ای که حدود دو سوم جمعیت مورد هدف در طول دوره شش ماهه توانستند در این آزمون شرکت نمایند. به علاوه در این مطالعه ویژگی‌های سایکومتریک آزمون در مقیاس وسیع بررسی شد.

متریسین و همکاران[1] ۲۰۰۷

این پژوهشگران در مطالعه خود بررسی کردند که آیا آزمون «همخوانی با شرحنامه» یک ابزار روا برای ارزیابی مهارت‌های تصمیم-گیری عملکردی فراگیران است یا خیر؟ در مرحله اولیه ۱۹۶ سؤال بر اساس اهداف بورد جراحی آمریکا[2] در خصوص آموزش دستیاری طراحی شد. سپس برای سنجش روایی سه نفر از اعضای بورد جراحان عمومی، به بررسی سؤالات پرداختند تا مشخص کنند آیا واقعاً سؤالات طراحی شده به ارزیابی عملکرد واقعی می‌پردازند و مهارت‌های تصمیم‌گیری بالینی را می‌سنجند یا خیر. همچنین روایی آزمون از طریق مقایسه محتوای سؤالات با اهداف آموزش دستیاری بورد جراحان عمومی آمریکا نیز بررسی شد. در نهایت بعد از تعیین روایی سؤالات ۱۰۰ سؤال انتخاب و برای ۳۶ نفر از دستیاران جراحی عمومی از سال اول تا پنج برگزار گردید. جهت تعیین روبریک نمره‌دهی، آزمون به ۱۰ نفر از اعضای بورد جراحان ارائه شد و از ایشان در خواست شد تا به شکل انفرادی به سؤالات پاسخ دهند. بعد از حذف و اصلاح سؤالات، در نهایت ۶۲ سؤال جهت آنالیز آماری مورد استفاده قرار گرفت.
نتایج حاصل از آزمون نشان‌دهنده پایایی بالای این آزمون (آلفای کرونباخ ۰/۸۵) بود. همچنین به جز یک کاهش جزئی که در نمرات دانشجویان سال پنجم مشاهده شد، نمرات آزمون با افزایش سطح آموزشی دستیاران افزایش یافت. به طور کلی تفاوت معناداری بین عملکرد دانشجویان سال پایین (سال اول و دوم) و سال بالا (سال سوم، چهارم و پنجم) مشاهده شد (p = ۰/۰۰۰۱). بر اساس نتایج حاصل از آزمون «همخوانی با شرحنامه» می‌تواند به طور دقیقی بین عملکرد دانشجویان سال پایین و بالای تمیز قائل شود.

1. Meterissian et al
2. American Board of Surgery

1. Riuz et al.
2. Nouh et al.
3. Brailovsky et al.
4. Brazeau et al.
5. Deschênes et al.

بیولفا و همکاران [1] ۲۰۱۰

در این مطالعه آزمون «همخوانی با شرحنامه» برای ۶۸ نفر از دانشجویان داروسازی برگزار شد. تعدادی از سؤالات در خصوص فرضیه‌های تشخیصی و برخی دیگر در مورد ارائه توصیه‌های درمانی بودند. در این مطالعه برای تعیین پاسخ‌های احتمالی، یک پانل نه نفره از داروسازان با تجربه حضور داشتند و برای تصحیح برگه‌های امتحانی از روش تجمیعی استفاده شد. بعد از تعیین کیفیت سؤالات آزمون، ۱۱ سؤال حذف شد و در نهایت آزمون نهایی به ۵۵ سؤال کاهش یافت. دانشجویان برای پاسخگویی به سؤالات ۹۰ دقیقه زمان داشتند. پاسخ‌های داوطلبان به هر سؤال با پاسخ‌های متخصصان مقایسه شد.
پایایی آزمون با استفاده از آلفای کرونباخ ۰/۶۸ گزارش شد. میانگین نمرات دانشجویان ۶۸/۵ درصد (انحراف معیار=۹/۸ درصد) و میانگین نمرات اعضای پانل ۸۶/۵ درصد (انحراف معیار=۴/۲ درصد) بود. محققان همچنین نتایج حاصل از اجرای آزمون را با سایر آزمون‌های کتبی متداول از جمله آزمون چندگزینه‌ای و آزمون کوته‌پاسخ مورد ارزیابی قرار دادند. نتایج نشان داد که آزمون چندگزینه‌ای برای سنجش حقایق ابزاری مفید است اما نمی‌تواند مهارت استدلال بالینی را به خوبی ارزیابی نماید و به‌همین دلیل ممکن است فراگیران را به مطالعه سطحی مطالب تشویق کند به جای آن که توانایی حل مسأله را در آنان تقویت کند.

1. Boulouffe et al

هامبرت و همکاران [1] ۲۰۱۱

این مطالعه با هدف ارزیابی مهارت استدلال بالینی دانشجویان دوره پیش بالینی [2] توسط آزمون «همخوانی با شرحنامه» انجام شد. ۷۵ سؤال در حوزه‌های آناتومی، بیوشیمی، فیزیولوژی و بافت شناسی طراحی شد. چارچوب نمره‌دهی به وسیله یک پانل ۳۰ نفره از اعضای هیأت علمی تعیین شد. در این پژوهش تعیین ویژگی‌های سایکومتریک ابزار ارزیابی بر محاسبه همسانی درونی و توانایی آزمون در ایجاد تمایز در عملکرد داوطلبان متمرکز شد. آزمون برای ۴۱۱ نفر از دانشجویان سال دوم پزشکی و ۷۰ نفر از دانشجویان سال چهارم برگزار شد.
پایایی ابزار قابل قبول گزارش شد (آلفای کرونباخ برابر با ۰/۷۳). بر اساس نتایج حاصل از مطالعه مشخص شد که این آزمون به خوبی می‌توان میان عملکرد دانشجویان سال دوم و چهارم و متخصصان تفاوت قائل شود.

1. Humbert et al.
2. Pre-clinical

نکته دیگر در زمینه بررسی روایی آزمون «همخوانی با شرحنامه»، اندازه‌گیری ارتباط بین نتایج حاصل از آن با نتایج حاصل از سایر ابزارهای اندازه‌گیری است:

- در مطالعه‌ای پژوهشگران ارتباط بین آزمون شفاهی ساختارمند را با آزمون «همخوانی با شرحنامه» بررسی کردند و نتیجه گرفتند که آزمون شفاهی می‌تواند اجزای استدلال بالینی را ارزیابی نماید اما وجود محدودیت‌هایی از قبیل دشواری استانداردسازی، ذهنی بودن فرایند نمره‌دهی و قابلیت اجرای پایین برای تعداد زیادی از داوطلبان مواردی هستند که اجرای این آزمون را تا حدودی دچار چالش‌هایی نموده است (گانون و همکاران ۲۰۰۵).
- گانون و همکاران (۲۰۰۶) در این مورد مطرح کردند که در مقایسه با سؤالات چندگزینه‌ای که به ارزیابی توانایی دانشجویان در به کارگیری دانش تاکید دارند، آزمون «همخوانی با شرحنامه» برای سنجش توانایی استدلال بالینی در موقعیت‌های مبهم و پیچیده طراحی شده است.
- در مطالعه دیگری نتایج حاصل از اجرای آزمون «همخوانی با شرحنامه» با سایر آزمون کتبی متداول از جمله سؤال چندگزینه‌ای و کوته‌پاسخ نشان داد که این آزمون‌ها برای سنجش حقایق، ابزاری مفید هستند اما نمی‌توانند مهارت استدلال بالینی را به خوبی ارزیابی نمایند (بیولفا و همکاران ۲۰۱۰).
- لابارسکی و همکاران (۲۰۱۱) در پژوهش خود به ارزیابی روایی آزمون «همخوانی با شرحنامه» بر اساس ۳۷ مطالعه منتشرشده پرداختند و شواهدی از روایی این آزمون در ارتباط با سایر آزمون‌ها ارائه نمودند که از آن میان می‌توان به روایی همگرا در ارتباط با سایر آزمون‌های استدلال بالینی و روایی واگرا در ارتباط با سایر آزمون‌های کتبی و آزمون‌های سنجش عملکرد اشاره کرد.

برایلفسکی و همکاران ۲۰۰۱

پژوهشگران در این مطالعه به ارزیابی توانمندی استدلال بالینی فراگیران در ابتدای دوره آموزشی به عنوان عاملی برای پیش‌بینی استدلال بالینی در انتهای دوره دستیاری پرداختند. این مطالعه به طور تجربی و با استفاده از آزمون «همخوانی با شرحنامه» صورت گرفت. نمرات حاصل از آزمون «همخوانی با شرحنامه» در انتهای دوره کارآموزی، دو سال بعد با نمرات حاصل از آزمون‌های OSCE، امتحانات شفاهی و آزمون PMP در انتهای دوره دستیاری مقایسه شد. تعداد دانشجویان شرکت‌کننده در این مطالعه ۲۴ دانشجو بود. از ضریب همبستگی پیرسون برای تعیین روایی ملاکی بین نمرات آزمون‌ها استفاده شد.
ضریب همبستگی بین نمره آزمون «همخوانی با شرحنامه» با آزمون PMP و آزمون شفاهی به ترتیب برابر با ۰/۴۵ و ۰/۴۴ بود که از نظر آماری معنادار بود. به علاوه نتایج حاصل از مطالعه نشان‌داد که نمرات حاصل از آزمون «همخوانی با شرحنامه» با آزمون OSCE از نظر آماری معنادار نبود (۰/۳۴۰، ۰۵۲/=p). نویسندگان بر اساس نتایج حاصل فرض کردند که نمرات آزمون «همخوانی با شرحنامه» با بخشی از مهارت استدلال بالینی که به وسیله آزمون PMP و آزمون شفاهی سنجیده می‌شود در ارتباط بود اما با نتایج حاصل از آزمون OSCE که به اندازه‌گیری مهارت بالینی و استدلال بالینی به صورت توام می‌پردازد، همبستگی نداشت.

گوله و همکاران [1] ۲۰۱۰

در این مطالعه، در طول یک دوره دو ساله، از ۲۰ نفر از پزشکان خانواده خواسته شد تا علاوه بر آزمون ساختارمند شفاهی در آزمون «همخوانی با شرحنامه» شرکت نمایند. از سه ارزیاب خواسته شد تا با توجه به اهداف مورد انتظار، کیفیت استدلال بالینی داوطلبان را تعیین نمایند. به منظور تعیین پایایی بین نمرات ارزیابان، همبستگی بین طبقات محاسبه گردید.
ضریب همبستگی بین نمرات آزمون شفاهی ساختارمند و آزمون «همخوانی با شرحنامه» برابر با ۰/۸۹ و ضریب آلفای کرونباخ در آزمون «همخوانی با شرحنامه» برابر با ۰/۹۰ گزارش شد. ضریب توافق کاپا بین دو روش ارزیابی استدلال بالینی برای ۱۳ نفر از داوطلبان ۰/۳۰ (p = ۰/۱۸) گزارش شد. محققان بر اساس نتایج حاصل از مطالعه پیشنهاد کردند که آزمون «همخوانی با شرحنامه» به همراه آزمون شفاهی ساختارمند، یک ابزار مفید در ارزیابی استدلال بالینی پزشکان با عملکرد ضعیف است.

1. Goulet et al

پایایی

همسانی درونی آزمون «همخوانی با شرحنامه» در مطالعات متعدد با استفاده از آلفای کرونباخ، به عنوان شاخصی از پایایی، اندازه‌گیری شده است.

- چارلین و همکاران (۲۰۰۰) بر اساس نتایج حاصل از مطالعه خود پیشنهاد کردند که حدود ۵۰ تا ۶۰ سؤال برای کسب همسانی درونی ۰/۸۰ با استفاده از روش آلفای کرونباخ کافی خواهد بود.
- در مطالعه‌ای دیگری که توسط فورنیر و همکاران[1] (۲۰۰۸) انجام شد، مشخص شد که در آزمون «همخوانی با شرحنامه» برای دستیابی به ضریب پایایی بالاتر از ۰/۷۵، یک آزمون یک‌ساعته شامل ۲۰ سناریو و ۶۰ سؤال نیاز خواهد بود.
- کریر و همکاران[2] (۲۰۰۸)، یک آزمون یک‌ساعته مشتمل بر ۶۰ سؤال را در قالب ۳۸ مورد بالینی برای ۵۳ دستیار در انتهای دوره آموزشی‌شان در بخش اورژانس کودکان اجرا نمودند و پایایی ۰/۷۷ برای آزمون را گزارش کردند.
- نتایج حاصل از مطالعه دیگر نشان داد که بهترین ضریب پایایی از طریق طرح سه سؤال برای هر سناریو کسب می‌گردد. همچنین افزودن تعداد بیشتری سؤال به هر سناریو، در مقایسه با زمانی که تعداد سناریوهای آزمون افزایش داده می‌شود، افزایش کمتری در پایایی آزمون ایجاد می‌کند، (گانون و همکاران ۲۰۰۹).
- دوری و همکاران[3] (۲۰۱۲) بر اساس مرور نظام‌مند انجام شده نتیجه گرفتند که آزمون «همخوانی با شرحنامه» می‌تواند در مدت زمان کوتاه (حدود ۹۰–۶۰ دقیقه) ارزیابی پایا و قابل اطمینانی از عملکرد داوطلبان فراهم سازد. به علاوه یافته‌ها مطالعه نشان داد که در آزمون‌هایی با حدود ۱۰۰ سؤال (۲۵ تا ۳۰ تابلوی بالینی)، که ۲۵ درصد از آن‌ها در مرحله تحلیل سؤالات کنار گذاشته شدند، نمرات از پایایی نسبتاً قابل قبولی برخوردار هستند (دوری و همکاران ۲۰۱۲).

1. Fournier et al.
2. Carrière
3. Dory et al.

پارک و همکاران ۲۰۱۰

پژوهشگران در این مطالعه به ارزیابی استدلال دستیاران در بخش جراحی زنان با استفاده از آزمون «همخوانی با شرحنامه» پرداختند. مطالعه به روش آینده‌نگر در پنج برنامه دستیاری زنان و زایمان (با شرکت ۷۵ دستیار) انجام شد. این پژوهش در دو فاز صورت گرفت: فاز اول شامل انتخاب و طراحی سؤالات آزمون (شامل ۴۲ سناریوی بالینی) و فاز دوم شامل تعین روایی و پایایی آزمون بود. سناریوهای بیماری بر اساس شایع‌ترین پروسیجرهای بیماری‌های زنان توسط تیمی از جراحان تدوین شد. روایی سازه ابزار از طریق تعیین ارتباط و همبستگی نمرات حاصل از آزمون با سطح شرکت‌کنندگان، آزمون‌های درون بخشی و ارزیابی مهارت‌های جراحی با استفاده از مقیاس نمره‌دهی گلوبال محاسبه شد.
همسانی درونی ابزار (آلفای کرونباخ= ۰/۷۳) و پایایی به روش بازآزمایی (ضریب همبستگی= ۷۶/۰) شد. نتایج حاصل از مطالعه نشان‌دهنده تفاوت معنادار بین نمرات حاصل از آزمون با سطح دستیاران بود (p=۰/۰۰۲). همچنین نتایج نشان داد که نمرات حاصل از آزمون «همخوانی با شرحنامه» با نمرات آزمون درون بخشی ارتباط معنادار دارد (۰/۰۰۱ = p، ۰/۳۸ = r). محققان بر اساس نتایج حاصل از مطالعه پیشنهاد کردند که این آزمون می‌تواند یک ابزار پایا و روا در ارزیابی مهارت تصمیم‌گیری دستیاران در دوره آموزشی جراحی زنان و زایمان باشد.

دیشنس و همکاران ۲۰۱۱

یک مطالعه متدولوژیک با هدف طراحی آزمون «همخوانی با شرحنامه» و تعیین اعتبار اولیه آن انجام شد. نمره‌دهی این آزمون به روش تجمیعی و به وسیله ۱۵ نفر از متخصصان صورت گرفت. ۳۰ نفر از دانشجویان پرستاری سال اول در این آزمون شرکت کردند. مقایسه نمرات متخصصان و دانشجویان با استفاده از آزمون t-test و پایایی نمرات آزمون با استفاده ضریب آلفای کرونباخ اندازه‌گیری شد. تفاوت معنادار آماری بین نمرات متخصصان و دانشجویان مشاهده شد. پایایی نمرات حاصل از اجرای آزمون بالا بود. محققان بر اساس نتایج حاصل از آزمون پیشنهاد کردند که آزمون «همخوانی با شرحنامه» یک روش استاندارد برای ارزیابی استدلال بالینی پرستاران فراهم می‌سازد.

مقبولیت

فورنیر و همکاران (۲۰۰۸) بر اساس نتایج حاصل از مطالعه خود بیان کردند که از آنجا که ممکن است تعداد کثیری از داوطلبان با این شکل از سؤالات آشنایی نداشته باشند، پیش از اجرا باید آموزش‌های واضح به همراه ارائه تعدادی نمونه سؤال جهت مثال و تمرین به فراگیران داده شود.

بیولفا و همکاران (۲۰۱۰) چهار ماه بعد از اجرای آزمون «همخوانی با شرحنامه»، برداشت‌های ۲۰ نفر از دانشجویان داروسازی را که به شکل تصادفی انتخاب شده بودند، با استفاده از مقیاس لیکرت پنج‌تایی بررسی نمودند. ۷۰ درصد از دانشجویان با این ابزار ارزیابی موافق یا کاملاً موافق بودند و ۹۰ درصد دانشجویان شرکت‌کننده در نظر سنجی اظهار کردند که آزمون «همخوانی با شرحنامه» توانسته موقعیت‌های واقعی مرتبط با داروسازی را بسنجد. ۷۵ درصد از دانشجویان نیز خاطر نشان کردند که این آزمون به آماده‌سازی آن‌ها برای آینده کمک نموده است.

استیوز و همکاران [1] ۲۰۱۲

در این مطالعه از دانشجویان در خصوص سودمندی و اثربخشی ابزار جدید اجرا شده در زمینه ارزیابی مهارت استدلال بالینی، از طریق توزیع پرسشنامه سه سؤالی نظرسنجی صورت گرفت. سؤالات پرسشنامه شامل موارد زیر بودند:

- آیا آزمون «همخوانی با شرحنامه» یک ابزار مفید در ارزیابی استدلال بالینی است؟
- آیا استفاده از نظرات ۲۰ متخصص به جای یک نفر در فرایند نمره‌دهی، فرایند ارزیابی را به شرایط واقعی عملکرد در بالین نزدیک‌تر می‌کند؟
- من احساس می‌کنم که آزمون «همخوانی با شرحنامه» در ارزیابی استدلال بالینی در استئوپاتی مفید است.

پرسشنامه برای تمام ۱۷ نفر از دانشجویان شرکت‌کننده ارسال شد که ۱۳ نفر (۷۶ درصد) پرسشنامه را تکمیل و بازگرداندند. نتایج حاصل نشان داد که ۳۱ درصددانشجویان کاملاً موافق و ۶۹ درصدموافق بودند که آزمون «همخوانی با شرحنامه» یک ابزار مفید در ارزیابی استدلال بالینی است. در ارتباط با استفاده از نظرات ۲۰ متخصص در گروه خبرگان، ۳۸ درصد کاملاً موافق و ۶۲ درصد موافق با این موضوع بودند. در نهایت اینکه ۳۱ درصد کاملاً موافق و ۶۲ درصد موافق بودند که آزمون «همخوانی با شرحنامه» یک ابزار مفید در ارزیابی مهارت استدلال بالینی در استئوپاتی به شمار می‌رود.

1. Esteves et al.

تأثیر آموزشی

همان‌طور که به کرات در سراسر این کتاب اشاره شده است، هر آزمون فارغ از سنجش عملکرد داوطلبان بر نحوه یادگیری آنان نیز تاثیرگذار است. برای رسیدن به نتایج مطلوب و ایجاد اثرات مثبت بر یادگیری فراگیران باید طراحی آزمون با توجه به دستورالعمل‌های ارائه شده صورت گیرد. تنها در این شرایط است که آزمون طراحی شده منجر به تغییرات رفتار مطلوب در یادگیری آن‌ها خواهد شد. در مطالعه بیولفا و همکاران (۲۰۱۰)، دانشجویان داروسازی بیان کردند که آزمون «همخوانی با شرحنامه» می‌تواند هم به آن‌ها در آمادگی برای دوره‌های آموزش‌شان و هم عملکرد حرفه‌ای آتی‌شان کمک کند.

هزینه

انجام ارزیابی ایده‌آل همیشه به علت محدودیت منابع امکان‌پذیر نیست. تعدادی از محدودیت‌های منابع که بسیار مرتبط با آموزش پزشکی است و باید مورد ملاحظه قرار گیرد، به شرح زیر است:

- میزان در دسترس بودن آزمونگران
- زمان طراحی آزمون
- زمان اجرای آزمون
- زمان تصحیح و تجزیه و تحلیل اوراق امتحانی
- هزینه مرتبط با اجرای آزمون
- آموزش اعضاء هیأت علمی

بنابراین هنگام طراحی آزمون «همخوانی با شرحنامه» نیز مانند سایر ابزارهای ارزیابی تمام این موارد در نظر گرفته شوند.

مسائل چالشی آزمون «همخوانی با شرحنامه»

طرح چه تعداد مورد بالینی و سؤال در آزمون «همخوانی با شرحنامه» کفایت می‌کند؟

در مطالعات مختلف از نظریه کلاسیک آزمون و نظریه تعمیم پذیری برای پاسخ به این سؤال که طرح چه تعداد مورد بالینی و سؤال در این نوع آزمون کفایت خواهد کرد، استفاده شده است.

گانون و همکاران (۲۰۰۹) بر اساس نتایج مطالعه خود توصیه کردند که طرح حداقل ۵۴ سؤال در پرستاری و ۴۸ سؤال در انکولوژی و ۱۰۲ سؤال در کودکان در این نوع آزمون ضروری است.

در یک مرور نظام‌مند انجام شده به منظور تعیین تعداد بهینه سؤالات و موارد بالینی در آزمون «همخوانی با شرحنامه»، نتایج حاصل از سه مطالعه بررسی شد (دوری و همکاران ۲۰۱۲). در یکی از مطالعات از فرمول اسپیرمن–براون[1] برای تعیین تعداد موارد بالینی مورد نیاز برای رسیدن به پایایی حدود ۰/۸۰ استفاده شده بود که نتایج نشان داد تعداد بهینه موارد بالینی ۳۶ تا (هر مورد بالینی با یک یا دو سؤال) خواهد بود (کریر و همکاران ۲۰۰۹). دو مطالعه دیگر که از نظریه تعمیم‌پذیری و D study استفاده کرده بودند، تعداد سؤالات آزمون را برای رسیدن به پایایی حدود ۰/۸۰، به ترتیب ۱۳۰ و ۱۰۲– ۴۸ برآورد کردند (رامیکرز و همکاران ۲۰۱۰، گانون و همکاران ۲۰۰۹).

به طور کلی در این مرور نظام‌مند محققان چنین نتیجه گرفتند که برای کسب پایایی قابل قبول (آلفای کرونباخ بین ۰/۷۵ تا ۰/۸۰)، آزمون «همخوانی با شرحنامه» باید حداقل شامل ۲۵ مورد بالینی باشد که در هر کدام از موارد بالینی نیز سه سؤال طرح شده باشد.

1. Spearman–Brown formula

روش نمره‌دهی تجمیعی یا اجماع‌نظر، کدام یک برای نمره‌دهی آزمون «همخوانی با شرحنامه» مناسب‌تر است؟

همان‌طور که در قسمت تصحیح آزمون «همخوانی با شرحنامه» اشاره شد، روش متداول نمره‌دهی در این آزمون از نوع تجمیعی است. در این روش، تنوع پاسخ‌های متخصصان در زمان مواجه با سؤالات پیچیده مربوط به حیطه استدلال بالینی پوشش داده می‌شود و نمره هر سؤال بر اساس پاسخ‌های انفرادی پانل متخصصان تعیین می‌شود. اصل زیربنایی این روش نمره‌دهی آن است که پاسخ‌های هر کدام از اعضای گروه خبرگان منعکس‌کننده نظرات ارزشمندی است. به این ترتیب کلیه پاسخ‌های ارائه شده در فرایند نمره‌دهی لحاظ می‌شوند. بر پایه این اصل، پاسخ‌ها با ضریب توافق پایین از پانل متخصصان نباید از فرایند نمره‌دهی کنار گذاشته شوند. در مقابل در روش نمره‌دهی اجماع‌نظر، پانل متخصصان به صورت گروهی در خصوص پاسخ‌های احتمالی به توافق نظر می‌رسند. با توجه به موارد فوق، یکی از سؤالات اساسی این است که کدام یک از این روش‌های نمره‌دهی به بهبود پایایی نمرات آزمون می‌انجامد؟

برای پاسخ به این سؤال، مطالعه‌ای توسط چارلین و همکاران در سال ۲۰۰۲ انجام شد. در این مطالعه در انتهای دوره آموزشی زنان و زایمان از ۱۵۰ نفر دانشجویان درخواست شد تا در صورت تمایل در آزمون «همخوانی با شرحنامه» شرکت نمایند. جهت تهیه کلید پاسخ از هفت عضو هیأت علمی گروه زنان و زایمان در خواست شد تا به صورت انفرادی پاسخ‌های محتمل را مشخص نمایند (روش تجمیعی). در مرحله بعد، از پانل متخصصان درخواست شد تا بهترین پاسخ را برای هر سؤال به صورت گروهی مشخص نمایند (روش اجماع‌نظر). آنالیز آماری نشان داد در روش تجمیعی، ۵۹ درصد پاسخ‌های ارائه‌شده متفاوت از پاسخ‌هایی بود که از روش اجماع‌نظر به دست آمد. به علاوه، تحلیل توصیفی عملکرد داوطلبان در آزمون نشان داد که میانگین نمرات در روش تجمیعی بالاتر از روش اجماع‌نظر متخصصان بود و در مقابل، دامنه توزیع و تغییرپذیری نمرات در روش اجماع‌نظر بیشتر از روش دیگر بود. همچنین ضریب آلفای کرونباخ در دو روش تجمیعی و اجماع‌نظر به ترتیب برابر با ۰/۵۲ و ۰/۶۳ بود. محققان افزایش میانگین نمرات در روش تجمیعی را ناشی از تأثیر نمره‌دهی نسبی پاسخ‌های احتمالی آزمون توسط گروه خبرگان دانستند زیرا در این روش گاهی اوقات برای پاسخ‌های نادرست نیز نمره‌ای تعلق می‌گیرد. در حالی که در روش اجماع‌نظر صرفاً به بهترین پاسخ تعیین شده نمره تعلق می‌گیرد. نتایج حاصل از مطالعه نشان داد که روش تجمیعی بهتر می‌تواند بین عملکرد دانشجویان و متخصصان تمیز قائل شود. به این ترتیب، محققان پیشنهاد کردند که روش نمره‌دهی تجمیعی در مقایسه با روش نمره‌دهی اجماع‌نظر از روایی سازه بالاتری برخوردار است.

مطالعه دیگری با استفاده از نظریه سؤال پاسخ نشان داد که در هر دو روش نمره‌دهی ویژگی‌های سایکومتریک آزمون مشابه می‌شود (کرایتر و همکاران ۲۰۰۵).

در مطالعات گوناگون روش‌های جایگزین متعددی برای افزایش پایای نمره‌دهی «آزمون همخوانی با شرحنامه» بررسی شده است، با این حال نتایج یک مرور نظام‌مند نشان داد که این روش‌ها بی‌تأثیر بوده‌اند و نتایج حاصل از روش سنتی (روش تجمیعی) روی پایایی و همبستگی نمرات آزمون رضایت‌بخش‌تر است (دوری و همکاران، ۲۰۱۲). البته باید در نظر داشت همان‌طور که ذکر شد یکی از انتقادهای جدی وارده شده به روش نمره‌دهی تجمیعی آن است که در این روش گاهی اوقات برای پاسخ‌های نادرست نیز نمره‌ای تعلق می‌گیرد که این امر منجر می‌شود تا سطح نمره داوطلبان از آنچه در واقعیت وجود دارد، قدری افزایش یابد.

منابع

1. Brazeau-Lamontagne L, Charlin B, Gagnon R, et al. Measurement of perception and interpretation skills during radiology training: utility of the script concordance approach. Med Teach 2004;26: 326-332.
2. Boulouffe C, Charlin B, Vanpee D.Evaluation of Clinical Reasoning in Basic Emergencies Using a Script Concordance Test. American Journal of Pharmaceutical Education 2010; 74 (10): 1-6
3. Brailovsky C, Charlin B, Beausoleil S, et al. Measurement of clinical reflective capacity early in training as a predictor of clinical reasoning performance at the end of residency: an experimental study on the script concordance test. Med Educ 2001;35:430-436
4. Carrière B, Gagnon R, Charlin B, Downing S. Georges Bordage,Assessing Clinical Reasoning in Pediatric Emergency Medicine: Validity Evidence for a Script Concordance Test. Annals of Emergency Medicine 2009; 53(5):647-52
5. Charlin B, Roy L, Brailovsky C, Goulet F, van der Vleuten C. The script concordance test: a tool to assess the reflective clinician. Teach Learn Med 2000;12:189-95.
6. Charlin B, van der Vleuten C. Standardized assessment of reasoning in contexts of uncertainty: The script concordance approach. Evaluation and the Health Professions2004 ; 27, 304-319.
7. Charlin B, Desaulniers M, Gagnon R, Blouin D, van der Vleuten C. Comparison of an aggregate scoring method with a consensus scoring method in a measure of clinical reasoning capacity. Teach Learn Med 2002;14 (3):150–6
8. Charlin B, Brailovsky CA, Leduc C, Blouin D. The diagnostic script questionnaire: A new tool to assess a specific dimension of clinical competence. Advances in Health Sciences Education 1998;51:3-8
9. Charlin B, Brailovsky CA, Brazeau-Lamontagne L, Samson L, Leduc C. Script questionnaires: Their use for assessment of diagnostic knowledge in radiology. Medical Teacher 1998;20:567–71.
10. Cohen LJ, Fitzgerald SG, Lane S, Boninger ML. Development of the seating and mobility script concordance test for spinal cord injury: obtaining content validity evidence. Assist Technol 2005;17:122–32
11. Deschênes MF, Charlin B, Gagnon R, Goudreau J, use of a script concordance test to Assess Development of clinical reasoning in Nursing students. Journal of Nursing Education 2011; 50(7):381-7
12. Dory V, Gagnon R, Vanpee D, Charlin B. How to construct and implement script concordance tests: insights from a systematic review. Medical Education 2012; 46: 552–563
13. Esteves JE, Bennison M, Thomson OP. Script concordance test: Insights from the literature and early stages of its implementation in osteopathy. International Journal of Osteopathic Medicine 2013; 1-9

14. Fournier P, Demeester A, Charlin B. Script concordance tests: guidleines for construction. BMC medical informatics and decision making 2008;8(18): 1.
15. Gagnon R, Charlin B, Lambert C, et al. Script concordance testing: more cases or more questions? Adv Health Sci Educ Theory Pract75–14:367;2009
16. Gagnon R, Charlin B, Coletti M, Sauvé E, van der Vleuten C. Assessment in the context of uncertainty: How many members are needed on the panel of reference of a script concordance test? Medical Education 2005; 39, 284-291
17. Gagnon R, Lubarsky S, Lambert C, Charlin B. Optimization of answer keys for script concordance testing: should we exclude deviant panelists, deviant responses, or neither? Adv in Health Sci Educ 2011; 16:601–608
18. Grant J, Marsden P. Primary knowledge, medical education and consultant expertise. Med Educ 1988;22:9–173 .
19. Goulet F, Jacoues A, Gagnon R, Charin B, Shabah A. Poorly Performing Physicians: Does the Script Concordance Test Detect Bad Clinical Reasoning? Journal of continuing education the health profession 2010; 30(3):161–166
20. Humbert A, Johnson M, Miech E, Friedberg F, Grackin J, Seidman P. Assessment of clinical reasoning: A Script Concordance test designed for pre-clinical medical students. Medical Teacher 2011; 33: 472–477.
21. Kelly, William, Steven Durning, and Gerald Denton. Comparing a script concordance examination to a multiple-choice examination on a core internal medicine clerkship. Teaching and learning in medicine 2012;24(3):187-193.
22. Kreiter CD, Bland AC, Gordon JA. Comparing Two Methods of Scoring a Script Concordance Test. Central Group on Educational Affairs, Association of American Medical Colleges, Spring Conference,2005 , Madison, WI
23. Lubarsky S, Chalk C, Kazitani D, Gagnon R, Charlin B. The Script Concordance Test: A New Tool Assessing Clinical Judgement in Neurology. The Canadian journal of neurological sciences 2009; 36(3): 326-331
24. Lubarsky S, Dory V, Duggan P, Gagnon R, Charlin B. Script concordance testing: From theory to practice: AMEE Guide No. 75. Medical Teacher 2013; 35: 184–193.
25. Marie I, Sibert L, Roussel F, Hellot MF, Lechevallier J, Weber J. The script concordance test: a new evaluation method of both clinical reasoning and skills in internal medicine. Rev Med Interne 2005;26(6):501-7
26. Meterissian S, Zabolotny B, Gagnon R, et al. Is the script concordance test a valid instrument for assessment of intraoperative decision-making skills? Am J Surg 2007;193:248–51
27. Norcini JJ, Shea JA, Day SC. The use of the aggregate scoring for a recertification examination. Eval Health Prof 1990;13:241–51

28. Norman GR, Tugwell P, Feightner JW, Muzzin LJ, Jacoby LL. Knowledge and clinical problem solving. Medical Education 1985;19:344–536.
29. Nouh T, Boutros M, Reid S, Pace D, Walker R, MacLean A, Hameed M, Charlin B, Meterissian SH. The script concordance test as a measure of clinical reasoning: a national validation study. The American Journal of Surgery 2012;203: 530–534
30. Park A, Barber M, Bant A, Dooley Y, Dancz C, Sutkin G, Jelovsek E. Assessment of intraoperative judgment during gynecologic surgery using the Script Concordance Test. American Journal of Obstetrics & Gynecology 2010; 240:1-6
31. Petrella, Robert J., and Paul Davis. Improving management of musculoskeletal disorders in primary care: the Joint Adventures Program. Clinical rheumatology 2007;26(7): 1061-1066.
32. Ramaekers S, Kremer W, Pilot A, van Beukelen P, van Keulen H. Assessment of competence in clinical reasoning and decision making under uncertainty: the script concordance test method. Assess Eval High Educ35;2010 (6):661–73.
33. Ruiz JG, Tunuguntla R, Charlin B, Ouslander JG, Symes SN, Gagnon R, Phancao F, Roos BA. The script concordance test as a measure of clinical reasoning skills in geriatric urinary incontinence. J Am Geriatr Soc 2010;58(11):2178-84
34. Schuwirth LW, van der Vleuten CP. Different written assessment methods: what can be said about their strengths and weaknesses? Medical Education 2004; 38:974–9
35. Sibert L, Charlin B, Corcos J, et al. Stability of clinical reasoning assessment results with the Script Concordance Test across two different linguistic, cultural and learning environments. Med Teach 2002;24:522-527
36. Sibert L, Darmoni SJ, Dahamna B, Weber J, Charlin B.Online clinical reasoning assessment with the script concordance test: a feasibility study. BMC Med Inform Decis Mak 2005;5:18.
۳۷. منجمی ع. استدلال بالینی: مفاهیم، آموزش و ارزیابی. انتشارات دانشگاه علوم پزشکی اصفهان، ۱۳۸۹، صفحه ۱۰۸.

فصل ۱۶

خانواده آزمون‌های استدلال بالینی

آزمون سناریونویسی یا ساختن فرضیه

در این آزمون توانایی آزمون‌دهندگان در حیطهٔ ساختن فرضیه مورد سنجش قرار می‌گیرد. این فرضیه‌ها می‌توانند تشخیصی[1] یا درمانی[2] باشند. به نوعی در این آزمون، تفکر همگرا مورد ارزیابی قرار می‌گیرد، به این شکل که فرد با خواندن اطلاعات بیمار، به یک جمع‌بندی از آن می‌رسد که می‌تواند به صورت فرضیه یا فرضیه‌هایی باشد و این جمع‌بندی، نوعی تفکر همگرا را می‌طلبد. توانایی در این حوزه، از حوزه جمع‌آوری اطلاعات مهم‌تر است اما باید توجه کرد که این دو مهارت (جمع‌آوری اطلاعات و جمع‌بندی اطلاعات) در ارتباط با یکدیگر هستند.

توانایی ساختن فرضیه‌ها به دو اصل اساسی وابسته است: یکی استدلال رو به جلو و دیگری قاعدهٔ امساک[3]. اولی به این کار می‌آید که بر مبنای اطلاعات موجود بتوانیم فرضیه بسازیم و دومی تضمین‌کننده ساخت فرضیه‌هایی با کیفیت است با کمک قاعده امساک می‌توان با بیشترین علایم و نشانه‌ها، کمترین تعداد فرضیه‌های تشخیصی را ساخت. به عبارت دیگر قاعده امساک به معنای تلاش برای ساختن حداقل فرضیه‌های تشخیصی با حدکثر علایم بیمار است. در آزمون‌های این حیطه باید به این نکته توجه داشت که آزمون‌دهنده چه ارتباطی میان یافته‌های موجود و فرضیه‌هایی که می‌سازد، برقرار می‌کند و دیگر آن که آیا به رعایت قاعدهٔ امساک توجه ویژه دارد یا خیر. بنابراین در طراحی سؤال و تهیه کلید این دو اصل باید مورد توجه قرار گیرند.

در این آزمون تعدادی علامت و نشانه به داوطلب داده می‌شود و از او خواسته می‌شود بر اساس این علائم و نشانه‌ها سناریوی یک بیمار را بنویسد؛ به گونه‌ای که حداکثر تعداد علائم و نشانه‌ها در این سناریو گنجانده شود. در ضمن فراگیر باید تشخیص نهایی سناریو را نیز بنویسد. نکته مهم در طراحی این آزمون این است که علائم و نشانه‌ها، کلی و بدون توضیح نوشته شوند تا جا برای سناریوهای مختلف باز باشد. مثلاً اگر به درد قفسه سینه اشاره شده است، لازم نیست که مشخص شود قلبی، ریوی، یا در اثر درگیری استخوان‌ها و عضلات است. این داوطلب است که باید بتواند این علائم را به گونه‌ای تعریف و توصیف کند که در آن درد قفسه‌سینه شکل خاصی به خود بگیرد. یعنی اگر درد قفسه سینه را به سمت یک درد رترواسترنال که با فعالیت بدتر می‌شود هدایت کرده است، بداند که با نفس کشیدن نباید بدتر شود و بداند که باید در ادامهٔ این سناریو عوامل خطر قلبی را بنویسد. در این میان، قاعدهٔ امساک حتماً باید رعایت شود یعنی آزمون‌شونده باید بتواند تمام علائم و نشانه‌ها را با یک بیماری توجیه کند و تا حد امکان از طرح چند بیماری همزمان پرهیز کند. در نوشتن سناریو باید به

1. Diagnostic
2. Therapeutic
3. Priniciple of parsimony

انسجام[1]، استفادهٔ بجا از اطلاعات زمینه‌ای[2] (مثل سن، جنس، شغل، رژیم غذایی و ...) و صحت[3] آنها توجه ویژه شود. انسجام در نوشتن سناریو مشخص خواهد کرد که آیا آزمون‌دهنده بر اساس یک شرحنامه مشخص سناریو را بنا کرده است یا خیر. این مساله در کنار استفاده از اطلاعات زمینه‌ای و بررسی صحت آن‌ها از ویژگی‌های اثبات شده خبرگان و پزشکان باتجربه است. همین موارد هستند که مبنای نمره‌دهی به این آزمون قرار خواهند گرفت (منجمی ۱۳۸۹).

باید توجه داشت که بهتر است برای طرح سؤال مجموعه‌ای هم از یافته‌های عینی (علایم[4]) و هم ذهنی (نشانه‌ها[5]) مورد استفاده قرار گیرند. برای انتخاب علایم و نشانه‌ها می‌توان یک سناریو نوشت که به چند تشخیص افتراقی منتهی شود و سپس علایم را از داخل همین سناریو برداشت. در غیر این صورت یعنی اگر علایم و نشانه‌ها را بدون توجه به سناریو انتخاب کنیم، این خطر وجود دارد که نتوان با همگی آن‌ها یک سناریو نوشت. به علاوه، باید توجه داشت که هر چه تعداد علایم و نشانه‌ها زیادتر باشد، تعداد تشخیص‌های محتمل کاهش پیدا می‌کند و بالعکس. در ضمن در این آزمون چون میزان انعطاف‌پذیری و خلاقیت داوطلبان مورد سنجش قرار می‌گیرد، بهتر است تعداد علایم و نشانه‌ها خیلی زیاد نباشد. می‌توان از علائم متداول مثل تب، کاهش وزن و ... هم استفاده کرد تا هم سناریو به شرایط واقعی نزیک‌تر شود و هم تعداد انتخاب‌ها افزایش پیدا کند.

برای تهیه کلید، از طراحان خواسته می‌شود تا سناریوهای مدنظر خود را بنویسند اما باید دقت کرد که این سناریوها قاعدتاً تنها جواب‌های ممکن نخواهند بود و برای پیدا کردن تمام پاسخ‌های ممکن باید از گروه خبرگان یاری گرفت. با این حال، گاهی در فرایند تصحیح به جواب‌هایی از دانشجویان بر می‌خوریم که در کلید موجود نیستند، اما سناریوی مناسبی به نظر می‌آیند. در این شرایط می‌توان با اخذ نظر موافق گروه خبرگان نمره را لحاظ نمود.

به علاوه تفاوت دیگر روش نمره‌دهی این آزمون در مقایسه با سایر آزمون در این است که تصحیح آن بر اساس نظرات تمامی اعضای هیات مجرب انجام نمی‌شود، بلکه هر سناریو نوشته شده توسط داوطلبان بوسیله دو نفر از اعضا به صورت جداگانه خوانده می‌شود، سپس بر اساس چک لست استاندارد به آن‌ها نمره داده می‌شود. در این مرحله در صورتی که بین ارزیابان اختلاف نظرهایی وجود داشته باشد، در جلسه حضوری هر یک از ارزیابان نظرات خود را به بحث گذاشته تا به اجماع نظر در مورد عملکرد داوطلب و نمره وی برسند. در این آزمون به استفاده از قاعده امساک (جهت کسب اطلاعات بیشتر در این خصوص به فصل یک همین بخش مراجعه نمایید)، استفاده مناسب از علایم بالینی و فاکتورهای زمینه‌ای (از قبیل سن، جنس، شغل و ...)، تناسب تشخیص مطرح شده با سناریوی نوشته شده نمره داده می‌شود (بر[6] ۲۰۰۳).

نمونه سؤال سناریونویسی

لطفاً با علایم و نشانه‌های زیر دو فرضیه تشخیصی بسازید، به گونه‌ای که تمام علایم و نشانه‌ها را پوشش دهد. در متن سناریوی نوشته شده، زیر علایم و نشانه‌ها خط بکشید. هر سناریوی بالینی، حداکثر ۲۰۰ کلمه باشد. در ضمن تشخیص بیمار را در هر مورد جداگانه ذکر کنید.
پرفشاری خون، تنگی نفس، سرفه، تب، افزایش Tactile fremitus، افزایش Vocal fremitus

در زیر نمونه‌ای از سناریوهای طراحی شده به عنوان پاسخ به سؤال بالا ارائه شده است.

1. Coherency
2. Contextual information
3. Accuracy
4. Signs
5. Symptoms
6. Ber

نمونه سناریوی اول

مرد ۶۸ ساله، کارگر بازنشسته کارخانه نساجی شکایت اصلی: افزایش شدت سرفه خلط دار
بیمار از حدود ۱۰ سال پیش تاکنون در بیشتر ماه‌های سال سرفه خلط دار دارد. از دو روز پیش دچار افزایش شدت و فرکانس سرفه، افزایش دفع خلط به رنگ سبز شده است. وی ظرف این دو روز با کمترین فعالیت دچار تنگی‌نفس می‌شود. از تب خفیف هم شکایت دارد. ورم پاهایش از دو سال پیش شروع شده که ظرف این دو روز تشدید شده است. ۴۵ سال است که روزی یک پاکت سیگار می‌کشد و سه سال است که سیگار را ترک کرده است. ۱۵ سال است پرفشاری خون دارد. سابقه مصرف کورتیکواستروئید استنشاقی، اسپری آتروونت، آنالاپریل، آسپرین و استازولامید را می‌دهد. بیمار در معاینه فیزیکی، صورت بر افروخته دارد، سیانوتیک، چاق و ادماتو است. فشار ورید ژیگولار وی بالا است. صدای اول قلبی تشدید یافته است. Tactile fremitus و Vocal fremitus در قاعده ریه راست تشدید یافته هستند. کراکل خشن در هر دو ریه شنیده می‌شود. هپاتومگالی و shifting dullness دارد. در اندام‌ها آدم گودهگذار ۲+ وجود دارد.

نمونه سناریوی دوم

خانم ۴۲ ساله، خانه دار شکایت اصلی: تنگی‌نفس ناگهانی و شدید از روز گذشته
بیمار از دیروز صبح دچار تنگی‌نفس پیشرونده شده است. این تنگی‌نفس در حالت استراحت، فعالیت، خوابیده یا نشسته تفاوتی نمی‌کند. همراه تنگی نفس از احساس ناراحتی مبهم در سرتاسر قفسه سینه شکایت دارد که به جایی تیر نمیکشد و بیشتر حالت سنگینی دارد. چندین بار سرفه همراه دفع خلط خونی داشته است. از هفت سال پیش سابقه پرفشاری خون و از دو سال پیش دیابت و دیس‌لیپیدمی دارد. داروی ضدچربی و ضددیابت خوراکی مصرف می‌کند. تا دوسال پیش قرص‌های ضدبارداری مصرف می‌کرده است.

Bw= 65kg Height=16- cm

RR=24 regular PR=120 regular صبح T oral=37/8
BP=175/100

خانمی است که آژیته است و دیسترس تنفسی دارد. در سمع ریه ویزینگ موضعی در ۱/۳ تحتانی ریه چپ دارد و در همین ناحیه tactile fremitus و vocal fremitus افزایش یافته است. قلب تاکی کارد است و s۴ هم شنیده می‌شود.

نمونه سؤال سناریونویسی

دانشجویان گرامی؛
لطفاً با علایم و نشانه‌های زیر دو سناریو بسازید. سناریوها باید به گونه‌ای باشد که تمام داده‌های موجود در باکس در آن گنجانده شود و شکایت اصلی بیمار هم حتماً یکی یا تعداد بیشتری از داده‌های داخل باکس باشد. حداکثر تعداد کلمه در مورد هر سناریو باید ۱۰۰ کلمه باشد. در ضمن تشخیص بیمار را در هر مورد جداگانه بنویسید.

مصرف سیگار – هموپتیزی – کراکل – درد ساق پا

تشخیص سناریوی اول:
متن سناریوی اول

تشخیص سناریوی دوم:
متن سناریوی دوم

آزمون غربالگری استدلال بالینی دانشگاه علوم پزشکی تهران، خرداد ۹۱

آزمون استدلال بالینی

این آزمون برای ارزیابی مهارت ساختن فرضیه‌های تشخیصی کاربرد دارد. به طور معمول در این امتحان، تعداد ۱۰ سناریو در آزمون گنجانده می‌شود و مدت زمان ۱۲۰ تا ۱۵۰ دقیقه برای آن در نظر گرفته می‌شود.

در این آزمون ابتدا یک سناریو و تعدادی تشخیص افتراقی مرتبط با آن مطرح می‌شود. از داوطلبان درخواست می‌شود بر اساس سناریو از بین مجموعه تشخیص‌های افتراقی مطرح شده، تنها یک گزینه را به عنوان پاسخ صحیح انتخاب نماید. سپس فهرستی از یافته‌های حاصل از شرح‌حال، معاینه فیزیکی، آزمایش‌های پاراکلینیک و ... به صورت گزینه طرح می‌شود. داوطلب باید یافته‌هایی را که مرتبط با تشخیص موردنظر وی است، مشخص نماید و بر اساس تشخیص مورد نظر به آنها + یا — بدهند. مثبت به معنای این است که این یافته تشخیص مورد نظر را تأیید می‌کند و منفی به این معناست که یافته، تضعیف‌کننده تشخیص انتخاب‌شده است. بدیهی است تمام یافته‌ها نمی‌توانند با منفی مشخص شوند اما آمیزه‌ای از — و + یا تماماً + قابل‌قبول است. در ادامه همین سناریو، سؤال دیگری ارائه می‌شود که در آن از فراگیر پرسیده می‌شود با فرض نادرست بودن تشخیص اولیه، چه تشخیص دیگری برای بیمار مطرح است. مجدداً مراحل قبلی تکرار می‌شوند یعنی دانشجو باید یافته‌های مرتبط با تشخیص را انتخاب کند و به آنها علامت مثبت یا منفی بدهد. باید توجه داشت که اولویت تشخیص‌ها در این آزمون مهم نیست. مثلاً اگر در مثال زیر دو تشخیص انفارکتوس میوکارد و دیسکسیون آئورت مورد نظر باشد، تفاوتی نمی‌کند که کدام یک ابتدا انتخاب شود. این روند می‌تواند با فرض نادرست بودن تشخیص قبلی تا چندین مرحله طی شود که معمولاً تا دو سطح صورت می‌گیرد. بدین ترتیب با طرح چنین سؤالاتی، امکان سنجش تعداد بیشتری از تشخیص‌های افتراقی و یافته‌های مرتبط برای فراگیران فراهم می‌شود.

نمونه سؤال «استدلال بالینی»

موضوع: درد شکم **محور: تشخیص** **گروه هدف: میانسالی**

شما در اورژانس بیمارستان کشیک هستید که خانم ۵۴ ساله‌ای در ساعت چهار صبح با شکایت از درد شکم مراجعه می‌کند. او چند سال پیش به خاطر دردهای شکمی متناوب و مشابه، ولی کمی خفیف‌تر در بیمارستان بستری گردیده بود. اخیراً میزان این تناوب به سه بار در هفته رسیده است که ماهیتی ناپیوسته داشته و به نیمه فوقانی شکم محدود بوده است و گهگاه به پشت انتشار داشته است. درد فعلی بیمار به مدت دو روز ادامه داشته که بر شدت آن نیز افزوده شده است که با تهوع و استفراغ همراه است. بیمار روزانه شش نخ سیگار مصرف می‌کند و شرح‌حالی از مصرف الکل نمی‌دهد. در معاینه بیمار چاق است و شدیداً عرق کرده است. ضربان قلب ۹۶ در دقیقه و فشار خون ۱۱۰/۷۵۰ میلی متر جیوه است. کبد پنج سانتی‌متر زیر لبه دنده ملموس است. حساسیت قابل ملاحظه‌ای در لمس شکم وجود دارد که همراه با گاردینگ در ناحیه اپی‌گاستر بود. صداهای روده‌ای قابل شنیدن نیست. معاینه رکتوم طبیعی است.

الف- کدام تشخیص، تابلوی بالینی فعلی در بیمار را توجیه می‌کند؟ از مجموعه تشخیص‌های زیر فقط یک تشخیص را انتخاب کنید.

☐ انفارکتوس میوکارد
☐ کوله سیستیت
☐ کبد چرب
☐ توکسیک مگاکولون
☐ پانکراتیت
☐ کولیت اولسروز

ب- از بین یافته‌های بیمار که در مجموعه زیر گرد آمده‌اند، حداکثر ۵ یافته را که مرتبط با تشخیص شماست، انتخاب کنید. یافته‌هایی که به نفع تشخیص با (+) و یافته‌هایی که به ضرر تشخیص شما هستند، با (-) مشخص نمائید.

☐ خانم ۵۴ ساله
☐ سابقه دردهای شکمی
☐ افزایش تناوب دردها
☐ سابقه مصرف سیگار
☐ عدم مصرف الکل
☐ چاق
☐ تعریق
☐ کبد بزرگ
☐ گاردینگ اپی گاستر
☐ درد ناگهانی شکم
☐ حساسیت در لمس
☐ نبود صداهای روده‌ای
☐ درد در نیمه فوقانی شکم
☐ انتشار درد به پشت
☐ ماهیت ناپیوسته درد
☐ تهوع و استفراغ

ج- اگر اثبات شود تشخیص اول شما غلط است، تشخیص دوم شما چیست؟ از میان تشخیص‌های زیر فقط یکی را انتخاب کنید.

☐ انفارکتوس میوکارد
☐ کوله سیستیت
☐ کبد چرب
☐ توکسیک مگاکولون
☐ پانکراتیت
☐ کولیت اولسروز

د- از بین یافته‌های بیمار که در مجموعه زیر گرد آمده‌اند، حداکثر ۵ یافته را که مرتبط با تشخیص شماست، انتخاب کنید. یافته‌هایی که به نفع تشیخص با (+) و یافته‌هایی که به ضرر تشخیص شما هستند، با (-) مشخص نمائید.

☐ خانم ۵۴ ساله
☐ سابقه دردهای شکمی
☐ افزایش تناوب دردها
☐ سابقه مصرف سیگار
☐ عدم مصرف الکل
☐ چاق
☐ تعریق
☐ کبد بزرگ
☐ گاردینگ اپی گاستر
☐ درد ناگهانی شکم
☐ حساسیت در لمس
☐ نبود صداهای روده‌ای
☐ درد در نیمه فوقانی شکم
☐ انتشار درد به پشت
☐ ماهیت ناپیوسته درد
☐ تهوع و استفراغ

به صورت کلی می‌توان گفت گام اول در طراحی آزمون «استدلال بالینی» نوشتن سناریو است. اندازهٔ سناریو نه باید به اندازه‌ای کوتاه باشد که نتوان تشخیصی را مطرح کرد و نه آن قدر اطلاعات وجود داشته باشد که فقط به یک تشخیص اشاره کند. یعنی سناریو نباید به گونه‌ای طراحی شود که فقط یک پاسخ درست داشته باشد. این مسأله مستلزم آن است که سناریو تا حدی مبهم باشد اما این ابهام به معنای پیچیده بودن یا نادر بودن مورد نیست بلکه ابهام به واسطهٔ آن است که اطلاعات برای تشخیص نهایی کافی نیست (دقیقاً شبیه همان اتفاقی که در شرایط واقعی ویزیت بیمار می‌افتد). بهتر است سناریو به گونه‌ای نوشته شود که حداقل دو تشخیص افتراقی مطرح باشد. تشخیص‌های دیگری که به عنوان گزینه طرح می‌شوند، می‌توانند کاملاً نادرست باشند اما نباید به گونه‌ای نوشته شوند که بدون خواندن سؤال بتوان حدس زد که آنها بی‌ربط هستند. مثلاً اگر در سؤال بالا که مشخصاً یک بیمار قلبی است، تشخیص‌هایی مثل پیلونفریت، اسکیزوفرنی یا سلیاک در گزینه‌ها نوشته شوند، بدون آن که لازم باشد سناریو را بخوانیم و بفهمیم، جواب‌های درست را می‌توان حدس زد. به علاوه توصیه می‌شود تعداد شش تشخیص به عنوان گزینه نوشته شوند.

در مورد یافته‌ها، توجه به این نکته ضروری است که لازم نیست از همه یافته‌هایی که در سناریو آمده است، استفاده شود. در ضمن لزومی ندارد که برای هر کدام از تشخیص‌ها حتماً یافته منفی وجود داشته باشد اما هنگامی که سناریو طراحی می‌شود، باید متوجه باشیم که تعدادی یافته اختصاصاً برای هر کدام از تشخیص‌ها در نظر گرفته شود وگرنه انتخاب یافته‌ها و کسب امتیاز به تشخیص انتخاب‌شده ارتباطی پیدا نمی‌کند. مثلاً در مثال بالا A2 تشدید یافته، به نفع آمبولی ریه است اما ربطی به تشخیص انفارکتوس میوکارد ندارد، بنابراین اگر تشخیص انفارکتوس میوکارد انتخاب شود، این یافته جزء گزینه‌های انتخاب شده نیست. نوشتن گزینه‌ها باید به صورتی باشد که هر کدام، شامل یک یافته در سناریو باشند. گاهی می‌توان چند علامت مربوط به هم را با هم جمع کرد و نامی به آن داد که البته تحت شرایطی، با دقت و ملاحظه باید صورت گیرد. مثلاً در مثال بالا می‌توان از گزینهٔ ویژگی‌های درد استفاده کرد که شامل محل، مدت و انتشار آن است. مهم است که برای دانشجو شفاف شود که چه تعداد یافته را باید انتخاب کند.

نمونه سؤال «استدلال بالینی»

خانم ۷۰ ساله‌ای که به علت درد ناگهانی رترواسترنال که به گردن و پشت تیر می‌کشد، مراجعه کرده است. درد بیمار به مدت ۴۵ دقیقه طول کشیده است. سابقه هیپرلیپیدمی و دیابت از چندین سال قبل را می‌دهد. در معاینه، بیمار مضطرب و عرق کرده است و ضربان قلب وی ۹۶ و منظم و تعداد تنفس ۲۴ عدد در دقیقه است. فشار خون او ۱۶۰/۹۰ میلی‌متر جیوه است (از دست راست) در دق و سمع ریه مشکل ندارد. در سمع قلب، S1 نرمال و A2 تشدید یافته است و S4 دارد.

۱- کدام تشخیص، تابلوی بالینی فعلی در بیمار را توجیه می‌کند؟ از مجموعه تشخیص‌های زیر فقط یک تشخیص را انتخاب کنید.

- ☐ پریکاردیت
- ☐ آمبولی ریه
- ☐ اندوکاردیت
- ☐ انفارکتوس میوکارد
- ☐ دیسکسیون آئورت
- ☐ آشالازی

۲- از بین یافته‌های بیمار که در مجموعه زیر گرد آمده‌اند، حداکثر ۵ یافته را انتخاب و طبق دستورالعمل زیر عمل کنید. یافته بیمار که به نفع تشخیص (+) یا به ضرر (-) آن است، مشخص نموده سپس آنها را به این ترتیب ارزش‌گذاری کنید.

- ☐ خانم ۷۰ ساله ...
- ☐ انتشار درد به گردن و پشت ...
- ☐ افزایش تناوب دردها
- ☐ سابقه هیپرلیپیدمی ...
- ☐ بیمار مضطرب ...
- ☐ BP=160/90
- ☐ A2 تشدید یافته
- ☐ کبد بزرگ
- ☐ درد ناگهانی رترواسترنال ...
- ☐ مدت ۴۵ دقیقه‌ای درد ...
- ☐ حساسیت در لمس
- ☐ سابقه دیابت ...
- ☐ تعریق ...

۳- اگر اثبات شود که تشخیص شما غلط است، تشخیص بعدی شما چیست؟ از مجموعه تشخیص‌های زیر فقط یک تشخیص را انتخاب کنید.

- ☐ انفارکتوس میوکارد
- ☐ کوله سیستیت
- ☐ کبد چرب
- ☐ توکسیک مگاکولون
- ☐ پانکراتیت
- ☐ کولیت اولسروز

۴- مثل تشخیص قبلی، یافته‌هائی که به نفع یا به ضرر تشخیص شماست در جدول زیر انتخاب کرده و طبق دستورالعمل بالا آنها را ارزش‌گذاری کنید.

- ☐ خانم ۷۰ ساله ...
- ☐ انتشار درد به گردن و پشت ...
- ☐ افزایش تناوب دردها
- ☐ سابقه هیپرلیپیدمی ....
- ☐ بیمار مضطرب ...
- ☐ BP=160/90
- ☐ A2 تشدید یافته
- ☐ کبد بزرگ
- ☐ درد ناگهانی رترواسترنال ...
- ☐ مدت ۴۵ دقیقه‌ای درد ...
- ☐ حساسیت در لمس
- ☐ سابقه دیابت ...
- ☐ تعریق ...

نمونه سؤال «استدلال بالینی»

موضوع: تنگی نفس **محور: تشخیص** **گروه هدف: کنهسال**

آقای ۶۰ ساله‌ای به علتِ تنگی‌نفس متعاقب فعالیت متوسط به شما مراجعه کرده است. شرح‌حالی از سرفه‌های صبحگاهی به مدت چند سال را می‌دهد که اخیراً تشدید شده است. اخیراً شب‌ها دچار تنگی‌نفس می‌شود و چندین بار برای دفع ادرار به دستشویی می‌رود. سیگار نمی‌کشد ولی سابقه مصرف الکل دارد. حدود ۱۸ سال پیش عمل جراحی روی فتق اینگوئینال وی صورت گرفته است. در معاینه فشار خون ۱۵۰/۹۰ میلیمتر جیوه و تعداد تنفس ۲۰ در دقیقه است. در سمع ریه ویز بازدمی منتشر شنیده می‌شود. ضربان قلب وی ۹۰ در دقیقه است و اپکس قلب کمی جابجا شده است. پالمار اریتما هم دارد.

الف -کدام تشخیص، تابلوی بالینی فعلی در بیمار را توجیه می‌کند؟ از مجموعه تشخیص‌های زیر فقط یک تشخیص را انتخاب کنید.

- ☐ آدم حاد ریوی
- ☐ نارسایی کلیه
- ☐ سیروز
- ☐ هیپوتیروئیدی
- ☐ نارسایی بطن چپ
- ☐ آسم

ب- از بین یافته‌های بیمار که در مجموعه زیر گرد آمده‌اند، حداکثر ۵ یافته را که مرتبط با تشخیص شماست، انتخاب کنید. یافته‌هایی که به نفع تشیخص با (+) و یافته‌هایی که به ضرر تشخیص شما هستند، با (-) مشخص نمائید.

☐ تنگی‌نفس فعالیتی
☐ سرفه
☐ ناکچوری
☐ سابقه مصرف الکل
☐ عدم مصرف سیگار
☐ سابقه تنگی‌نفس شبانه
☐ فشار خون ۱۵۰/۹۰
☐ ویز بازدمی
☐ RR=20/m
☐ پالمار اریتما
☐ اپکس قلب جابجا شده
☐ RR=90/m

ج -اگر اثبات شود تشخیص اول شما غلط است، تشخیص دوم شما چیست؟ از میان تشخیص‌های زیر فقط یکی را انتخاب کنید.

☐ آدم حاد ریوی
☐ نارسایی کلیه
☐ سیروز
☐ هیپوتیروئیدی
☐ نارسایی بطن چپ
☐ آسم

د- از بین یافته‌های بیمار که در مجموعه زیر گرد آمده‌اند، حداکثر ۵ یافته را که مرتبط با تشخیص شماست، انتخاب کنید. یافته‌هایی که به نفع تشیخص با (+) و یافته‌هایی که به ضرر تشخیص شما هستند، با (-) مشخص نمائید.

☐ تنگی‌نفس فعالیتی
☐ سرفه
☐ ناکچوری
☐ سابقه مصرف الکل
☐ عدم مصرف سیگار
☐ سابقه تنگی‌نفس شبانه
☐ فشار خون ۱۵۰/۹۰
☐ ویز بازدمی
☐ RR=20/m
☐ پالمار اریتما
☐ اپکس قلب جابجا شده
☐ RR=90/m

آزمون غربالگری حیطهٔ استدلال بالینی المپیاد علمی دانشجویان علوم پزشکی خرداد ۹۱

در آزمون «استدلال بالینی» نیز همانند سایر آزمون‌های این حوزه، پاسخ‌های صحیح متعددی برای یک سؤال وجود دارد که شناسایی آنها توسط تیمی از متخصصان صورت می‌گیرد. در سیستم نمره‌دهی آزمون «استدلال بالینی»، برای تک‌تک سؤالات، گزینه تشخیصی و یافته‌های مرتبط توسط تمامی اعضای گروه خبرگان به صورت انفرادی تعیین می‌شود. سپس فراوانی گزینه‌های انتخابی توسط اعضای گروه خبرگان برای هر سؤال محاسبه می‌شود. بر این اساس، گزینه‌هایی که حداقل توسط دو سوم متخصصان به عنوان پاسخ صحیح انتخاب شده باشند به عنوان کلید سؤال انتخاب می‌شوند (امینی و همکاران ۲۰۱۱).

شیوه نمره‌دهی آزمون «استدلال بالینی» نیز به این صورت محاسبه می‌شود که امتیاز مربوط به تشخیص درست و یافته‌های مرتبط با آن، مساوی هم در نظر گرفته می‌شود. به عنوان مثال اگر به قسمت تشخیص‌گذاری و انتخاب یافته‌ها ۱/۲ نمره به طور کل تعلق گیرد، ۰/۲ نمره برای انتخاب تشخیص صحیح و یک نمره (۰/۲ برای انتخاب هر یافته صحیح) در نظر گرفته می‌شود. در این روش، در صورت انتخاب بیش از ۵ یافته به ازای انتخاب هر گزینه اضافی، یکی از گزینه‌های پنجگانه حذف می‌شود. به علاوه در صورتی که تشخیص انتخاب شده نادرست باشد، داوطلب کل نمره آن سؤال را از دست داده و هیچ امتیازی به یافته‌های انتخابی نیز تعلق نمی‌گیرد. اما اگر تشخیص درست باشد، در حالی که یافته‌های نادرستی انتخاب شده باشند تنها امتیاز تشخیص درست (۰/۲ نمره) به داوطلب داده می‌شود. نمره هر سؤال بر پایه مجموع گزینه‌های صحیح انتخاب شده و نمره نهایی آزمون بر اساس مجموع نمرات سؤالات محاسبه می‌شود (گروز و همکاران[1] ۲۰۰۲).

1. Groves et al.

در زیر نمونه‌ای از برگه پاسخنامه آزمون استدلال بالینی آورده شده است.

در پاسخنامهٔ زیر قسمت الف و ج مربوط به تشخیص‌هاست بنابراین از ۱ تا ۶ فقط یک عدد را انتخاب کنید. قسمت‌های ب و د مربوط به یافته‌های تشخیص است بنابراین از ۱ تا ۱۶ فقط ۵ گزینه را انتخاب کنید.

سؤال اول	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰	۱۱	۱۲	۱۳	۱۴	۱۵	۱۶
الف																
ب																
ج																
د																

آزمون غربالگری حیطهٔ استدلال بالینی المپیاد علمی دانشجویان علوم پزشکی خرداد ۹۱

آزمون پازل ادغام‌یافته

آزمون پازل ادغام‌یافته با این هدف طراحی شده است که مهارت داوطلب در شناسایی شرحنامه یا الگو را ارزیابی کند. این آزمون هم بر اساس نظریه شرحنامه بنا شده است. همان‌طور که پیشتر نیز اشاره شد، وقتی پزشک باتجربه با بیماری برخورد می‌کند که در حیطهٔ تجربه و تخصص او است، بلافاصله الگویی (یا شرحنامه‌ای) از آن بیماری در ذهنش فعال می‌شود و او بلافاصله بیماری را تشخیص می‌دهد. در این آزمون این روش شناسایی الگو ارزیابی خواهد شد، بنابراین برعکس بسیاری از آزمون‌های استدلال بالینی، در این آزمون برای هر سناریو، فقط باید یک تشخیص مطرح کرد و این امر مستلزم بازشناسی الگوی بیماری‌هاست.

این روش مانند آن است که پروندهٔ تعدادی از بیماران در هم ریخته شده است و ما تلاش می‌کنیم تا با خواندن مجدد اطلاعات مربوط به پرونده‌ها، آن‌ها را مجدداً مرتب کنیم. در این آزمون، پروندهٔ بیماران به چند قسمت شامل شکایت اصلی و بیماری فعلی، سابقهٔ پزشکی قبلی، معاینه و نتایج پاراکلینیک تفکیک شده است. دانشجویان باید قطعات درهم‌ریخته را جور کنند و برای هر تابلوی بالینی فقط یک تشخیص بگذارند تا موفق شوند قطعات مختلف پرونده را با هم جور کنند (بر[1] ۱۹۹۸، بر و همکاران ۲۰۰۰). می‌توان گفت که در این آزمون، استدلال غیرتحلیلی به خوبی ارزیابی می‌شود و همان طور که اشاره شد، از آنجا که استدلال غیرتحلیلی یکی از ویژگی‌های خبرگان در پزشکی است، این آزمون ارزش بسیار زیادی برای تفکیک پزشکان باتجربه از نوآموزان خواهد داشت.

نمونهٔ آزمون پازل بیماری‌ها

داوطلب گرامی؛
در این آزمون شما با اطلاعات پروندهٔ چهار بیمار که شامل بیماری فعلی، سابقهٔ پزشکی قبلی، معاینه و پاراکلینیک است مواجه می‌شوید. اطلاعات پرونده‌های این بیماران با هم مخلوط شده است. شما باید قطعات مرتبط با یکدیگر را در کنار هم قرار داده و پروندهٔ چهار بیمار را جداگانه بازسازی کنید. شمارهٔ هر کدام از قطعات شکایت اصلی و بیماری فعلی، سابقهٔ پزشکی قبلی، معاینه و نتایج پاراکلینیک را در مورد هر بیمار در یک ردیف وارد کنید.

1. Ber

الف- شکایت اصلی و بیماری فعلی	
الف ۱- آقای ۶۰ ساله با درد قفسه سینه در ناحیه رترواسترنال مراجعه کرده است. همراه با درد، تعریق سرد فراوان و تهوع و دو بار استفراغ حاوی مواد غذایی نیز دارد.	الف ۲- آقای ۲۰ ساله با تورم و آدم صورت و اندام‌ها مراجعه کرده است. شرح‌حالی از کاهش حجم ادرار و تیره شدن رنگ ادرار را نیز از چند روز قبل می‌دهد. بی اشتهایی، تهوع، سردرد و درد پهلوها را نیز متذکر است.
الف ۳- خانم ۱۸ ساله با ضعف و بی‌حالی شدید که حتی قادر به ایستادن نیست مراجعه نموده است. بیمار دچار خونریزی از بینی و دهان شده و ظاهر کاملاً Pale دارد، از لکه و خونمردگی‌هایی بر روی اندام‌ها نیز شکایت دارد و از دو سه روز قبل دچار تب ۳۹/۵ درجه شده است.	الف ۴-آقای ۶۵ ساله با تشدید سرفه و خلط و تنگی‌نفس مراجعه نموده است. نامبرده هر چند ماه یک بار دچار تشدید سرفه، خلط و تنگی‌نفس شده و به بیمارستان مراجعه می‌کند. درد قفسه سینه ندارد.
ب- سابقهٔ پزشکی قبلی	
ب ۱ - سابقه فشار خون بالا و لیپید بالا را متذکر است، مصرف سیگار روزانه یک پاکت به مدت ۲۰ سال داشته است.	ب ۲ - سابقه گلودرد چرکی را حدود دو هفته قبل بیان می‌کند. مصرف مسکن نداشته است. سابقه بیماری دیگری نداشته است.
ب ۳- سابقه مصرف ۴۰ ساله سیگار دارد HTN و HLP ندارد. سابقه یک بار نیز بستری در CCU را دارد شرح‌حالی از تنگی‌نفس و سرفه خلط را که به گفته خودش در زمستان و پائیز بدتر می‌شود را دارد.	ب ۴- سابقه هیچ بیماری را نداشته، شرح‌حال از خستگی، کاهش وزن و ابتلا به گلودردهای مکرر را از ۲ ماه قبل متذکر بوده است.
ج- معاینه	
ج ۱- اسکلرا کاملاً Pale، علایم کبود شدگی متعدد در اندام‌ها و تنه دارد. طحال و کبد در معاینه بزرگ است. تب نیز دارد. از ۳-۲ روز قبل تب داشته است.	ج ۲- آدم +۲ اندام تحتانی دارد.
ج ۳- دیسترس تنفسی و سیانوز مرکزی و محیطی دارد. سمع ریه‌ها رال های Coarse پراکنده در سرتاسر ریه‌ها دارد.	ج ۴- بیمار به شدت مضطرب به نظر می‌رسد اندام‌های بیمار سرد و Pale می‌باشد تعریق فراوان بر روی صورت و پیشانی دارد.
د- پاراکلینیک	
د ۱-	د ۲-
U/A: NL WBC=11000 PMN=65% TG=220 LDL=272 HDL=30 BS=176 Hb=16 ECG: ST(V4 تا V1 در	U/A: Pro=3+ WBC=1-2 BUN=33 Cr=3.6 WBC=9000 PMN=55% Hb=10.6 ECG→PVC
د ۳-	د ۴-
U/A=NL WBC=11000 PMN=80% BS=70 ABG: PH=7.56 PaCo2=65 HCo3-=34	U/A→NL Hb=2.9 WBC=800 Plt=1000 PT=27 PTT=56 INR=2.8

به منظور طراحی آزمون پازل بیماری‌ها به راحتی می‌توان تعدادی سناریو را انتخاب کرد و به قطعات مساوی (با توجه به تقسیم‌بندی بالا) تقسیم کرد و سپس قطعات به دست آمده را با هم مخلوط کرد. باید توجه داشت که تا حد امکان اجزای مختلف سناریوها با هم شباهت داشته باشند تا نتوان از شکل صوری آنها پی به تشخیص برد. مثلاً اگر یکی از سناریوها در مورد آنمی فقر آهن است، گنجاندن آزمایش CBC در صورتی که در آزمایش‌های پاراکلینیک بقیه قطعات این آزمایش نیامده باشد، به صورت خودکار ما را به این سمت می‌برد که این دو به هم مربوط هستند. در صورتی که اگر همه سناریوها در قسمت پاراکلینیک، CBC داشته باشند، با این روش نمی‌توان جواب را حدس زد.

از سوی دیگر اگر سناریوها بسیار به هم نزدیک باشند و شامل ویژگی‌های تیپیک بیماری موردنظر نباشند، استدلال تحلیلی فعال خواهد شد و ما را از منظورمان دور خواهد کرد. بنابراین در طراحی این گونه از آزمون‌ها باید توجه کرد که طراحی قسمت اول سناریو بسیار حیاتی است و اطلاعات باید به اندازه‌ای باشد که بتوان یک الگو را تشخیص داد.

در این آزمون نیز پاسخ‌ها بدون وزن ارزشی در نظر گرفته می‌شوند، به این معنا که سهم امتیاز هر گزینه مساوی است. نمره مربوط به هر سؤال از ترکیب امتیاز چهار بخش شرح حال، معاینه فیزیکی، پاراکلینیکی و تدبیر بالینی به عنوان جواب درست محاسبه می‌شود. در مواردی که دو یا سه قطعه پازل با هم جور باشند قسمتی از نمره کامل ارائه خواهد شد. به عنوان مثال اگر داوطلب به دو قطعه درست اشاره کند نمره ۰/۳ و اگر به سه قطعه اشاره کند نمره ۰/۶ به وی تعلق می‌گیرد. البته ذکر این نکته ضروری است که نمره تنها در شرایطی به داوطلب داده خواهد شد که حتماً یکی از قطعات پازل مربوط به شرح حال مورد بالینی باشد (بر ۲۰۰۳).

در زیر نمونه‌ای از فرم پاسخنامه آزمون پازل بیماری‌ها ارائه شده است.

در پاسخنامهٔ زیر عدد متناسب با هر سؤال در قطعات الف، ب، ج و د را وارد کنید.			
	الف	ب	ج	د
سؤال اول				
سؤال دوم				
سؤال سوم				
سؤال چهارم				

دانشجویان گرامی؛
در آزمون پازل شما با ۴۰ سؤال مواجه خواهید شد که در ۱۰ گروه (از A تا J) چهارتایی مرتب شده‌اند. هر گروه چهارتایی در پاسخنامه بخش مجزایی را به خود اختصاص داده است و شما در هر گروه فقط باید گزینه‌های همان قسمت را به هم جور کنید. به بیان دیگر گزینه‌های گروه A باید با هم جور شوند و گزینه‌های گروه D با همدیگر. هر سؤال مثال پروندهٔ یک بیمار است که در بین سه پروندهٔ دیگر گم شده و شما باید چهار پروندهٔ بیمار را از نوع بازیابی کنید. جواب تکراری وجود ندارد و هر گزینه فقط و فقط به یک بیمار اختصاص دارد. به این ترتیب شما اگر در هر گروه از سؤالات سه سؤال را درست پاسخ دهید، سؤال آخر به طور خودکار جواب داده می‌شود.
فرض کنید شما می‌خواهید به سؤال زیر پاسخ دهید. ابتدا گروه سؤال را با گروه روی پاسخنامه تطبیق دهید. به هر سؤال در پاسخنامه یک باکس اختصاص یافته است که شمارهٔ سؤال مورد نظر روی آن نوشته شده است، پس قدم دوم تطبیق شماره سؤال با شمارهٔ آن در پاسخنامه است.
گروه A

سؤال ۱-	سؤال ۲-
سؤال ۳- مرد سیگاری با حملهٔ تنگی نفس شدید ساعت ۲ بعد از نیمه شب به اورژانس آورده‌اند. خلط خونی کف‌آلود هم دفع می‌کند.	سؤال ۴-

در اِدامه کاری که باید انجام دهید این است که بر اساس تشخیصی که از سناریوی بیمار داده‌اید، قطعات الف، ب و ج را با آن جور کنید. مثلاً در بیمار بالا که تشخیص آن آدم حاد ریوی است باید در سابقهٔ قبلی به دنبال عوامل خطر قلبی بگردیم. پس در قطعات الف شمارهٔ ۳ را بر می‌گزینیم که به این شرح است:

الف ۱-

الف ۲-

الف ۳- سابقهٔ چربی خون بالا و فشار خون بالا را از چندین سال قبل می‌دهد. سابقهٔ بیماری عروق کرونری و بای‌پس کرونر ۳ سال قبل را دارد.

الف ۴-

به این ترتیب قسمتی که با فلش در پاسخنامه سؤال ۳ آمده است باید پر شود.
قسمت بعدی مربوط به معاینهٔ بیمار است. با توجه به تشخیص آدم حاد ریوی در بخش ب، شمارهٔ ۴ گزینه مناسب به نظر می‌رسد.

ب ۱-

ب ۲-

ب ۳-

ب ۴- در معاینه قادر به خوابیدن نیست. تعداد تنفس ۳۰ در دقیقه و فشار خون ۱۸۰/۹۵ است. در سمع ریه رال سرتاسری و در سمع قلب گالوپ S4-S3 دارد.

پس باید در ستون ب گزینه‌ای که با فلش نشان داده شده است، علامت زده شود. قسمت آخر، بخش گزینه‌های ج است که مربوط به پاراکلینیک است. با تشخیص آدم حاد ریوی که در سر داریم، گزینهٔ ج ۲ را مناسب تشخیص می‌دهیم که به این شرح است:

ج ۱-

ج ۲- در گرافی سایز قلب بزرگتر از عادی است و کدورت منتشر دو طرفه و بیشتر اطراف ناف ریه‌ها دارد.

ج ۳-

ج ۴-

به این ترتیب پاسخدهی به سؤال ۳ پایان می‌یابد. لازم به ذکر است که در ستون الف یا ب یا ج فقط می‌توانید یک گزینه را انتخاب کنید و در صورت انتخاب بیش از یک گزینه هیچ نمره‌ای به شما تعلق نمی‌گیرد. اما در هر ردیف هر سه گزینه در یک ردیف می‌توانند انتخاب شوند. در صورت انتخاب غلط نمرهٔ منفی تعلق نمی‌گیرد و در صورتی که موفق نشدید تمام گزینه‌ها را دست انتخاب کنید به ازاء انتخاب هر گزینهٔ درست بخشی از نمره به شما تعلق خواهد گرفت.

آزمون غربالگری حیطه استدلال بالینی دانشگاه علوم پزشکی تهران، خرداد ۹۱

آزمون PMP

نگاهی اجمالی به ابزارهای کتبی مرسوم در حیطه علوم پزشکی نشان‌دهنده این واقعیت است که الگوی رایج در ارزیابی مشتمل بر الگوی کمی‌نگر است و این در حالی است که به سبب ماهیت پویا و فرایندی استدلال بالینی، این رویکرد در ارزیابی مهارت استدلال بالینی چندان کارساز نیست. بنابراین توجه به این نکته ضروری است که ارزیابی استدلال بالینی نوعی ارزیابی کیفی‌نگر است و که اگر با دیدگاه کمی‌نگر به سراغ ارزیابی این گونه آزمون‌ها برویم، در سنجش این مهارت‌ها دچار مشکل خواهیم شد (ون‌درولوتن و نیوبل ۱۹۹۵).

این نگرانی‌ها منجر به ارائه شکل جدیدی از آزمون‌های کتبی به نام آزمون تدبیر مشکل بیمار (PMP) شد. آزمون PMP معمولاً با یک عبارت بالینی در ارتباط با مشکل موجود بیمار، همراه با خلاصه‌ای از شرح حال شروع می‌شود. بعد از آن چند مرحله در ادامه به صورت پی در پی مطرح می‌گردد و در هر مرحله از داوطلب خواسته می‌شود تا در مورد اداره و ارزیابی بیمار تصمیم بگیرد. بنابراین در این نوع ارزیابی آزمون‌شونده قادر است مشکل بیمار را درک کرده، اطلاعات لازم برای حل مسأله را جمع‌آوری کند، اطلاعات جمع‌آوری شده را تجزیه و تحلیل نماید و از اطلاعات برای حل مسأله استفاده کند. بدین ترتیب در آزمون PMP داوطلب با یک بیمار که اطلاعات محدودی از او در دسترس است، مواجهه می‌شود و بعد از مطالعه اطلاعات باید تصمیم‌گیری نماید که چه اقدامی برای بیمار باید صورت گیرد که ممکن است شامل درخواست یک سری آزمایش‌های پاراکلینیکی یا سایر روش‌های تشخیصی باشد. نهایتاً داوطلب باید در رابطه با درمان و مدیریت روند بیماری تصمیماتی را اتخاذ نماید.

زمان زیادی از ارائه این آزمون‌ها نگذشت که کشف برخی محدودیت‌های آن در مطالعات گوناگون، استفاده از این آزمون‌ها را با اما و اگرهایی همراه ساخت. در ادامه به تعدادی از این موارد اشاره‌ای خواهیم داشت.

یکی از مهمترین انتقادهای وارده به آزمون PMP، بحث ویژگی مورد بالینی بود. در اواخر دهۀ هفتاد میلادی این گونه تصور می‌شد که حل مسأله یک مهارت عمومی است که می‌تواند فراگرفته شود. مطالعات بعدی نشان داد که حل موفق هر مسأله، تضمین کننده حل موفق مسأله دیگری نیست (الستین[1] ۲۰۰۲). این پدیده نشان از آن دارد که راه حل هر مسأله، امری یکتا و منحصر به فرد است و به این شکل ارزیابی استدلال بالینی به مثابه ارزیابی مهارت‌هایی کلی و عمومی که نتایج حاصل از آن قابل تعمیم به سایر موقعیت‌هاست، تصوری نادرست است. پایایی پایین این آزمون‌ها یکی دیگر از نگرانی‌های مربوط به این شکل از آزمون بود (نورمن و همکاران ۱۹۸۵).

از آن گذشته، نتایج حاصل از مطالعات صورت گرفته نیز مشخص کرد که PMP به جای آنکه توانایی تصمیم‌گیری مناسب را مورد ارزیابی قرار دهد، در اغلب اوقات توانایی جمع‌آوری اطلاعات را در داوطلبان می‌سنجد (پیج و همکاران ۱۹۹۵). حتی از بُعد نتیجه، هیچ تفاوتی بین عملکرد دانشجویان و افراد متخصص در این نوع آزمون‌ها مشاهده نشد. همان‌طور که پیشتر اشاره شد، آزمون‌های استدلال بالینی بر اساس فرایند استدلال بالینی ارائه شده‌اند، اما باید به گونه‌ای طراحی شوند که به خوبی میان نوآموزان و خبرگان تمایز قائل شوند.

اگر آزمونی طراحی شود و معیار آن صرفاً تعداد سؤال پرسیده شده از بیمار باشد، احتمالاً نمرات دانشجویان و متخصصان دور از واقعیت می‌شود. چرا که در واقعیت متخصصان به دلیل آنکه تنها به دنبال جستجوی نکات کلیدی و مهم هستند، سؤالات کمتری از بیمار می‌پرسند. در حالی که دانشجویان مبتدی به دلیل ناکافی بودن اطلاعاتشان و ناتوانایی در افتراق مطالب ضروری از مطالب غیرضروری در یک موقعیت بالینی، سؤالات زیادی به ذهنشان می‌رسد. بنابراین یکی از مشکلاتی که موجب شد PMP به کنار گذاشته شود این بود که این آزمون نمی‌تواند به خوبی بین عملکرد دانشجویان مبتدی و خبرگان تمایز ایجاد نماید.

1. Elstein

در نهایت اینکه، مطالعات صورت گرفته نشان‌دهنده همبستگی بسیار بالای بین نمرات این آزمون‌ها با نمرات حاصل از سایر آزمون‌های کتبی بود که این موضوع را مطرح ساخت که احتمالاً این آزمون‌ها هیچ‌گونه اطلاعات بیشتری در خصوص عملکرد داوطلبان فراهم نمی‌سازد (ون‌درولوتن و نیوبل ۱۹۹۵، پیج و بوردیج ۱۹۹۵).

به طور خلاصه امروزه از این آزمون به دلیل روایی و پایایی پایین، زمان‌بر بودن و مشکلات اجرایی کمتر استفاده می‌شود (نورمن و همکاران ۱۹۸۵). معایب زیاد آزمون PMP، پژوهشگران علم سنجش و اندازه‌گیری در علوم پزشکی را بر آن داشت تا به فکر آزمون‌های جایگزین در زمینه استدلال بالینی باشند.

آزمون جامع استدلال بالینی

امروزه در منابع مکتوب ارزیابی استدلال بالینی، گرایش به سمت ارزیابی یک مهارت از طریق سنجش چندجانبه آن مورد توجه قرار گرفته است. به عبارت دیگر می‌توان گفت رویکرد یک آزمون برای سنجش یک ویژگی، جای خود را به رویکرد آزمون‌های چندگانه برای سنجش چند منظوره داده است و این ایده که برای سنجش هر حیطه یا حوزه محتوایی باید یک آزمون وجود داشته باشد، اکنون با شواهد و مطالعات موجود دیگر قابل دفاع نیست (رسائیان ۱۳۸۱). این نکته به ویژه برای ارزیابی مهارت پیچیده‌ای مانند استدلال بالینی که چندوجهی است، بسیار حیاتی است. چرا که اساساً جستجو برای یافتن یک آزمون که بتواند مهارت استدلال بالینی را بسنجد، اندیشه نادرستی است. به بیان دیگر، چند آزمون از منظرهای مختلف، مهارتی را ارزیابی می‌کنند و مجموع این آزمون‌ها، نمره نهایی فرد در آن مهارت است.

از این رو آزمون جامع[1] استدلال بالینی بر این ایده بنا شده است که ارزیابی استدلال بالینی با تنها یک آزمون، ارزیابی جامعی نیست چرا که هر کدام از آزمون‌های استدلال بالینی موجود، بخشی یا جنبه‌ای از مهارت استدلال بالینی را می‌سنجند و برای ارزیابی استدلال بالینی فراگیر نیاز به مجموعه یا آمیزه‌ای از آزمون‌های استدلال بالینی موجود داریم. برای کنار هم گذاشتن آزمون‌های استدلال بالینی نیاز به چارچوب نظری مناسبی داریم. به نظر می‌رسد، نظریه پردازش دوگانه[2] چارچوب مناسبی برای این کار در اختیارمان قرار می‌دهد (جهت کسب اطلاعات بیشتر در این خصوص به فصل یک همین بخش مراجعه نمایید).

بر پایه این فرض، مسلماً پرسش بعدی این خواهد بود که چه آزمون‌هایی را با چه منطقی باید کنار هم گذاشت. به همین دلیل باید یک چارچوب نظری طراحی کنیم تا بتوانیم این چندجانبه‌گی را در آن لحاظ کنیم. در این حالت آزمون جامعی طراحی می‌گردد که در آن برای ارزیابی مهارت استدلال بالینی، سه مهارت جمع‌آوری اطلاعات، ساختن فرضیه و ارزیابی فرضیه باید سنجیده شود تا بتوان تصویر کاملی از استدلال بالینی فرد به دست آورد.

همچنین باید به این مسأله دقت کنیم که هر کدام از آزمون‌های موجود، بیشتر کدام مهارت استدلال بالینی را مورد ارزیابی قرار می‌دهند. اگر از این منظر به آزمون‌های استدلال بالینی که تا اینجا به آنها اشاره شد، نگاه شود، می‌توان به این بررسی دست زد که کدام آزمون بیشتر به سمت تحلیلی تمایل دارد و کدام آزمون به سمت استدلال غیر تحلیلی. با توجه به ساختار و نحوهٔ دستورالعمل به نظر می‌رسد، آزمون‌های «ویژگی‌های کلیدی» و «استدلال بالینی» بیشتر به استدلال تحلیلی تمایل دارند و آزمون‌های سناریونویسی و پازل به استدلال غیرتحلیلی. آزمون «استدلال بالینی» از آنجا که بیش از یک تشخیص را از پاسخ‌دهنده طلب می‌کند و در ضمن دلایل تائیدکننده تشخیص را می‌پرسد، استدلال تحلیلی را فعال می‌کند. در مقابل، هر دو آزمون پازل و سناریونویسی بر پایهٔ یک تشخیص برای هر سناریو یا تابلوی بالینی استوار هستند. بنابراین بسان نوار قلب که یک لید به تنهایی نمی‌تواند وضعیت قلب را به درستی گزارش دهد و نیاز به چندین لید داریم، در ارزیابی استدلال بالینی هم نیاز به چند آزمون داریم. مطالعه بوشهری و همکاران نشان داده است که از میان چند آزمون فوق، آزمون استدلال بالینی بیشتر از بقیه تحلیلی است، آزمون

1. Comperehensive Exam
2. Dual-processing

«ویژگی‌های کلیدی» به میانهٔ طیف تمایل دارد و پازل بیشتر از سایرین غیرتحلیلی است (بوشهری ۱۳۹۴).

برای برطرف کردن مشکل ویژگی محتوا و مورد بهتر است سه آزمون به صورت جداگانه برگزار شوند، چرا که تعداد سناریوها در هر آزمون باید حداقل ۲۰ مورد باشد و در این شرایط اجرای یه آزمون کلی که مشتمل بر اشکال مختلف آزمون‌های استدلال به دلیل طولانی بودن، غیرعملی است. اما اگر به دلایلی می‌خواهیم هر سه مهارت را در یک امتحان بسنجیم، بهتر است سؤالات به صورت دسته‌بندی و از ساده به دشوار مرتب شوند. به علاوه باید توجه داشت که آزمون‌های جمع‌آوری اطلاعات، ساده‌ترین و آزمون‌های ارزیابی فرضیه، دشوارترین آزمون‌ها محسوب می‌شوند.

در مورد نحوه نمره‌دهی به آزمون جامع دو رویکرد را می‌توان پیش گرفت. یکی آنکه به تمام آزمون‌ها نمره مساوی تعلق گیرد و دیگری اینکه وزن هر کدام از آزمون‌ها متفاوت باشد اما نکته‌ای که توجه به آن ضروری است این است که مطالعات مختلف نشان داده‌اند که شیوه‌های متفاوت وزن‌دهی در نهایت باعث ایجاد تمایز و اختلاف چشمگیری در نمرات آزمون‌ها نخواهد شد (بلاند و همکاران[1] ۲۰۰۵).

ادیبی و همکاران ۱۳۸۹

اولین المپیاد علمی دانشجویان علوم پزشکی کشور به منظور شناسایی و توانمندسازی دانشجویان توسط وزارت بهداشت و با همکاری دانشگاه‌های وابسته در مرداد ۱۳۸۸ در شهر اصفهان برگزار شد. در این المپیاد ۳۶۴ دانشجو در قالب تیم‌هایی سه نفره از دانشگاه‌های علوم پزشکی سراسر کشور در رشته‌های مختلف شرکت کردند. کمیته‌های طرح سؤال با شرکت ۳۵ نفر از اعضای هیأت علمی از ۱۳ دانشگاه کشور به طرح سؤالاتی در فرمت‌هایی از قبیل آزمون‌های «ویژگی‌های کلیدی»، «همخوانی با شرحنامه»، سناریونویسی، جورکردنی و تشریحی برای دو روز آزمون فردی و دو روز آزمون گروهی در سه حیطه حل مسأله در مدیریت نظام سلامت، تفکر علمی در علوم پایه و استدلال بالینی پرداختند. به منظور ارزیابی کیفیت آزمون فرم‌های نظرسنجی بین دانشجویان توزیع شد. حدود ۶۰ درصد داوطلبان در مجموع از کیفیت برگزاری آزمون راضی بودند.

امینی و همکاران ۲۰۱۱

در این مطالعه از چهار ابزار ارزیابی استدلال بالینی شامل آزمون‌های «ویژگی‌های کلیدی»، «همخوانی با شرحنامه»، «استدلال بالینی» و «پازل ادغام‌یافته» در دومین المپیاد علمی دانشجویان علوم پزشکی استفاده شد. هدف از این مطالعه طراحی یک ابزار چند کارکردی برای ارزیابی مهارت‌های استدلال بالینی بود. ۱۳۵ دانشجوی پزشکی از ۴۵ دانشگاه علوم پزشکی در ایران در حیطه استدلال بالینی المپیاد علمی شرکت کردند. در این پژوهش، پایایی آزمون با استفاده از آلفای کرونباخ محاسبه شد. ضریب دشواری سؤالات و همبستگی بین نمره هر سؤال با نمره کل آزمون اندازه‌گیری شد. همچنین همبستگی بین نمره کل دانشجویان با هر کدام از آزمون‌های استدلال بالینی محاسبه گردید.
پایایی کل آزمون ۰/۹۱ گزارش شد. بالاترین میزان پایایی مربوط به آزمون پازل ادغام‌یافته بیماری‌ها بود (۰/۹۱). میزان پایایی آزمون «ویژگی‌های کلیدی» ۰/۸۳، آزمون «همخوانی با شرحنامه» ۰/۷۸ و آزمون «استدلال بالینی» ۰/۷۱ گزارش شد. ضریب دشواری اکثر سؤالات آزمون در محدوده‌ی بین ۰/۲ تا ۰/۸ قرار داشت. همبستگی بین نمره هر سؤال با کل آزمون برای هر چهار ابزار ارزیابی مثبت بود، به این ترتیب که بالاترین میزان همبستگی در آزمون «پازل ادغام‌یافته» و آزمون «ویژگی‌های کلیدی» مشاهده شد. نویسندگان بر اساس نتایج حاصل از مطالعه نتیجه گرفتند که آزمون جامع استدلال بالینی یک ابزار ارزیابی روا و پایا برای ارزیابی مهارت استدلال بالینی دانشجویان پزشکی استعداد درخشان است.

منجمی و همکاران ۲۰۱۲

در این مطالعه، آزمون جامعی بر اساس چارچوب نظری استدلال بالینی برای سنجش استدلال بالینی دانشجویان پزشکی در المپیاد علمی طراحی شد. در این مطالعه از آزمون‌های موجود در منابع، آزمون‌های ویژگی‌های کلیدی برای ارزیابی جمع‌آوری اطلاعات، آزمون استدلال بالینی برای ارزیابی مهارت ساختن فرضیه، آزمون همخوانی با شرحنامه و آزمون پازل بیماری‌ها برای سنجش مهارت ارزیابی فرضیه‌ها و حل مسأله مناسب تشخیص داده شد. در مورد مهارت جمع‌آوری اطلاعات با توجه به آنکه برخی از جنبه‌های این مهارت مانند مهارت جمع‌آوری اطلاعات قابل اعتماد و معتبر بر اساس منابع مختلف ارزیابی مورد سنجش قرار نمی‌گرفت، آزمون جمع‌آوری اطلاعات نیز طراحی گردید. در مورد ساختن فرضیه‌ها، آزمون دیگری به نام سناریونویسی طراحی شد که در آن توانایی فرد برای نوشتنِ یک سناریوی بالینی کامل بر اساس تعداد علامت و نشانه سنجیده می‌شود. هر کدام از این آزمون‌ها، به صورت آزمونی مستقل اجرا شد، مثلاً ۴۰ سؤال ویژگی‌های کلیدی به مدت دو ساعت، ۳۰ سؤال آزمون استدلال بالینی به مدت دو ساعت و نیم، ۲۰ سؤال پازل به مدت نود دقیقه و ۱۰ سؤال سناریو به مدت سه ساعت برگزار شد و نتایج چهار آزمون تجمیع و به عنوان یک نمره از ۱۰۰۰ گزارش گردید. سهم هر چهار آزمون در نمرهٔ نهایی یکسان و برابر بود. نتایج حاصل از اجرای آزمون نشان داد که عدم آشنایی دانشجویان با این آزمون‌ها، فراگیر نبودن استفاده از این آزمون‌ها در سایر رشته‌ها به جز پزشکی، ایجاد رقابت و تنش بیش از حد بین دانشجویان و نبود آموزش‌های ساختارمند جهت تقویت استدلال از مهمترین چالش‌های پیش رو برگزاری این آزمون بوده است

1. Bland et al.

نتایج تحقیقات روایی و پایایی قابل قبولی آزمون جامع را نشان می‌دهد (امینی و همکاران ۲۰۱۲، منجمی و همکاران ۲۰۱۳) از آنجا که آزمون جامع در المپیاد در سطح دانشجویان دوره پزشکی عمومی برگزار می‌شود و در مورد سطوح بالاتر هنوز مطالعه‌ای انجام نشده است، نیاز به مطالعات تکمیلی در این مورد احساس می‌شود.

منابع

1. Amini M, et al. An innovative method to assess clinical reasoning skills: Clinical reasoning tests in the second national medical science Olympiad in Iran. BMC Research Notes 2011; 4(418): 1-7
2. Ber R. The CIP (Comprehensive Integrative Puzzle) scoring system, paper presented at the AMEE annual conference, Prague, Czech Republic,. 1998.
3. Ber R, Brik R. Four years' experience with the CIP assessment method: formative vs summative, paper presented at the Ottawa in Africa Conference, Cape Town, South Africa. 2000.
4. Groves M, Scott I, Alexander H. Assessing clinical reasoning: a method to monitor its development in a PBL curriculum. Medical teacher 2002;24(5):507-515.
5. Monajemi A, et al. A comprehensive test of clinical reasoning for medical students: An olympiad experience in Iran. Journal of Education and Health Promotion 2012; 1: 1-5
6. Norman G, Bordage G, Curry L et al. Review of recent innovations in assessment. In: Wakeford R, ed. Directions in Clinical Assessment. Report of the Cambridge Conference on the Assessment of Clinical Competence.Cambridge: Office of the Regius Professor of Physic, Cambridge University School of Clinical Medicine, Addenbrooks Hospital 1985;8–27
7. Page G, Bordage G, Allen T. Developing key-feature problems and examinations to assess clinical deision-making skills. Acad Med 1995;70:194-201.
8. Page G, Bordage G. The medical council of Canada's key features project: A more valid written examination of clinical decision-making skills. Acad Med 1995;70:104-110
9. Van der Vlueten CP, Newble DI. How can we test clinical reasoning? Lancet 1995; 345:1032–1034.

۱۰. ادیبی پ و همکاران. برگزاری اولین المپیاد علمی دانشجویان علوم پزشکی کشور: گزارش یک تجربه. مجله ایرانی آموزش در علوم پزشکی، ۱۳۸۹؛ ۱۰(۵): ۱۰۰۶– ۱۰۱۷

۱۱. بوشهری ا. تبیین الگوی ذهنی استدلال بالینی پزشکان مبتدی و خبره در مواجهه با آزمون‌های استدلال بالینی، دانشگاه علوم پزشکی ایران. ۱۳۹۳.

۱۲. رساییان ن، نخعی س، صادقی قندهاری ن. مقایسه روش‌های آزمون‌های چندگزینه‌ای، صحیح غلط و کوته‌پاسخ در دانشجویان پزشکی، مجله ایرانی آموزش در علوم پزشکی ۱۳۸۱؛ ۵ (۴): ۲۷۱–۲۷۸.

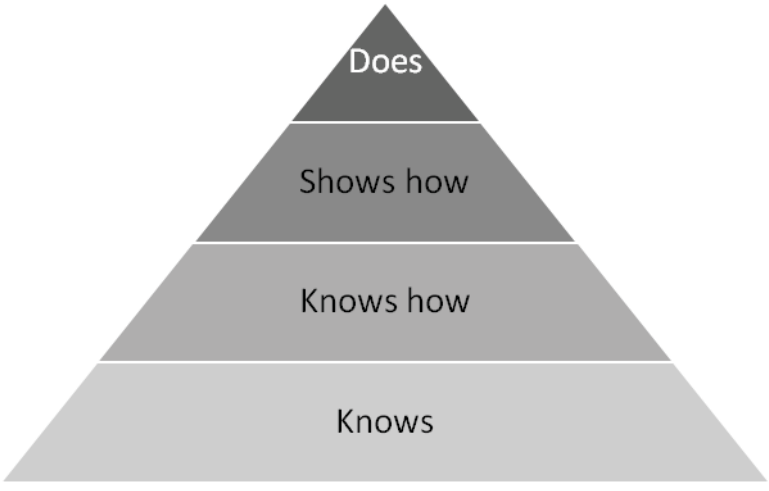
۵
آزمون‌های ساختارمند عینی

فصل ۱۷

خانواده آزمون‌های ساختارمند عینی

تاریخچه آزمون‌های ساختارمند عینی

ارزیابی صلاحیت بالینی فراگیران اهمیت زیادی دارد و موجب اطمینان خاطر جامعه در این خصوص می‌شود که پزشکانی که فارغ‌التحصیل می‌شوند، از صلاحیت لازم برای طبابت و مراقبت بیماران برخوردار هستند. شورای پزشکی عمومی[1] انگلیس بر اهمیت سنجش دقیق صلاحیت بالینی تاکید می‌کند. سنجش مهارت‌ها و صلاحیت‌های بالینی و عملی دانشجو در سالن امتحانات و در قالب آزمون‌های کاغذی معمول امکان‌پذیر نیست و لازم است شرایط و موقعیت‌های خاصی فراهم گردد تا این کار مقدور باشد. به عنوان مثال، بهترین روش برای سنجش مهارت معاینه فیزیکی دانشجو، در مواجهه با بیمار امکان‌پذیر است. در هرم میلر دو سطح فوقانی به عملکرد بالینی فراگیر اشاره دارند (شکل ۱-۱۷). با این تفاوت که دانشجو در سطح «انجام می‌دهد»[2]، در محیط کار واقعی و به صورت روزمره باید وظایف محوله را انجام دهد و در سطح «نمایش می‌دهد چگونه»[3] کافی است نشان دهد که قادر به انجام این موارد می‌باشد.



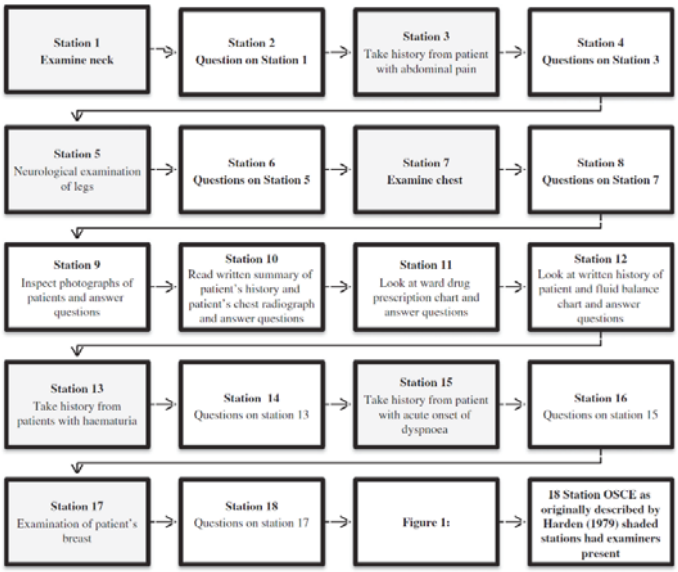
شکل ۱-۱۷: هرم میلر

1. General Medical Council (GMC)
2. Does
3. Shows

به صورت سنتی آزمون‌هایی که برای سنجش سطح «نمایش می‌دهد چگونه» استفاده می‌شد، «مورد بالینی کامل»[1] و «مورد بالینی کوتاه»[2] بود (خان و همکاران[3] ۲۰۱۳). این ابزارها در بخش بعدی کتاب مورد بحث قرار می‌گیرند اما به صورت کلی می‌توان گفت که همواره در کنار مزایا و کاربردهای این ابزارها انتقاداتی نیز به آنها وارد بوده است. از جمله اینکه این آزمون‌ها داری ساختار مشخص و استاندارد نبودند، در آنها تعداد موارد بالینی ارائه شده به هر دانشجو کم بود، موارد بالینی دانشجویان مختلف با یکدیگر متفاوت بود و دانشجویان توسط ارزیابی‌کنندگان متفاوت به نحوی مورد سنجش قرار می‌گرفتند که معمولاً زمینه قضاوت ذهنی و ارزیابی به دور از معیار و ملاک مشخص فراهم بود. همان‌طور که مشخص است منابع خطا در چنین آزمون‌هایی متعدد بودند و این مسائل روی هم رفته موجب روایی، پایایی و قابلیت تعمیم‌پذیری پایین آنها می‌شد (گرملی[4] ۲۰۱۱، خان و همکاران ۲۰۱۳).

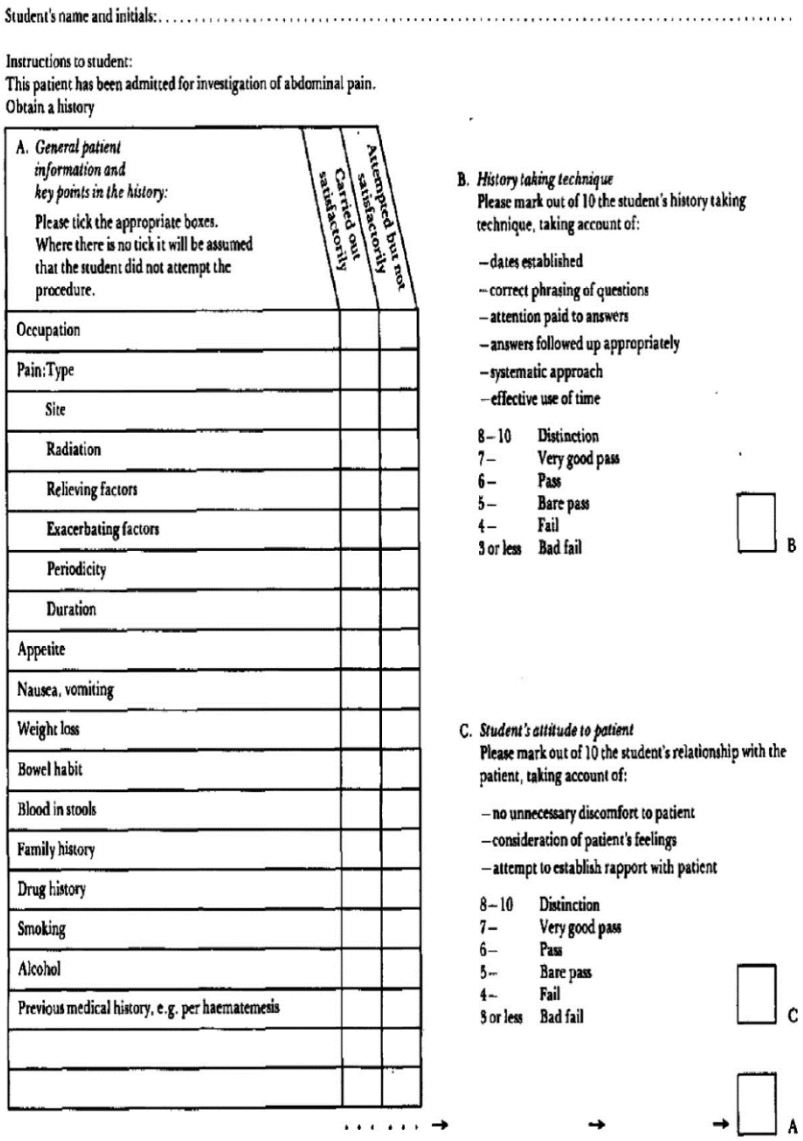
هر چند آزمون‌های فوق همچنان کم و بیش به منظور ارزیابی سطح «نمایش می‌دهد چگونه» مورد استفاده قرار می‌گیرند، در پاسخ به محدودیت آنها و برای کاهش منابع خطای اندازه‌گیری، آزمون عینی ساختارمند بالینی[5] (OSCE) معرفی شد (هاردن و همکاران[6] ۱۹۷۵). ویژگی بارز OSCE این است که یک آزمون واحد برای تمام دانشجویان در مدت زمان مشخص با نمره‌دهی یکسان برگزار می‌شود و از آنجا که دانشجو در یک موقعیت شبیه‌سازی شده مهارت‌های خود را نشان می‌دهد، شرایط نسبت به محیط واقعی، راحت‌تر و کنترل شده‌تر است. ایده‌ال این است که در OSCE منابع خطا به حداقل برسد و تنها متغیر موجود، سطح توانمندی شرکت‌کنندگان باشد. به عبارت دیگر، نمرات OSCE تنها منعکس‌کننده سطح توانمندی دانشجویان باشد و نه خطاهایی مانند تفاوت آزمونگران یا تفاوت سؤالات. البته این منابع خطا را نمی‌توان در OSCE حذف کرد و نهایتاً تاثیر خود را روی نتایج آزمون می‌گذراند ولی این تاثیر در مقایسه با سایر ابزارها بسیار کمتر و کنترل شده‌تر است.

آزمونی که اولین بار در سال ۱۹۷۲ در دانشگاه داندی اجرا شد به این صورت بود که هر دانشجو ۱۸ ایستگاه آزمون و ۲ ایستگاه استراحت را پشت سر می‌گذاشت (شکل ۲-۱۷ و شکل ۳-۱۷). محل آزمون در یک بخش بیمارستانی و زمان کل آزمون ۱۰۰ دقیقه بود. دانشجو در هر ایستگاه چهار دقیقه و نیم فرصت داشت و بین ایستگاه‌ها ۳۰ ثانیه فاصله در نظر گرفته شده بود (هاردن ۱۹۷۹)



شکل ۲-۱۷: نمای شماتیک آزمون OSCE که اولین بار توسط هاردن و همکاران در دانشگاه داندی برگزار شد (هاردن ۱۹۷۹)

1. Long case
2. Short case
3. Khan et al.
4. Gormley
5. Objective Structured Clinical Examination (OSCE)
6. Harden et al.



Student's name and initials:.....

Instructions to student:
This patient has been admitted for investigation of abdominal pain.
Obtain a history

A. *General patient information and key points in the history:* Please tick the appropriate boxes. Where there is no tick it will be assumed that the student did not attempt the procedure.	Carried out satisfactorily	Attempted but not satisfactorily
Occupation		
Pain: Type		
Site		
Radiation		
Relieving factors		
Exacerbating factors		
Periodicity		
Duration		
Appetite		
Nausea, vomiting		
Weight loss		
Bowel habit		
Blood in stools		
Family history		
Drug history		
Smoking		
Alcohol		
Previous medical history, e.g. per haematemesis		

B. *History taking technique*
Please mark out of 10 the student's history taking technique, taking account of:

- dates established
- correct phrasing of questions
- attention paid to answers
- answers followed up appropriately
- systematic approach
- effective use of time

8–10 Distinction
7– Very good pass
6– Pass
5– Bare pass
4– Fail
3 or less Bad fail

B

C. *Student's attitude to patient*
Please mark out of 10 the student's relationship with the patient, taking account of:

- no unnecessary discomfort to patient
- consideration of patient's feelings
- attempt to establish rapport with patient

8–10 Distinction
7– Very good pass
6– Pass
5– Bare pass
4– Fail
3 or less Bad fail

C

A

شکل ۳-۱۷: چک‌لیست نمره‌دهی یکی از ایستگاه‌های OSCE که اولین بار توسط هاردن و همکاران برگزار شد (هاردن ۱۹۷۹)

پس از آن OSCE به کرات برای سنجش توانمندی‌های بالینی دانشجویان علوم پزشکی مورد استفاده قرار گرفت. شواهد زیر میزان گستردگی استفاده از آزمون OSCE را نشان می‌دهند:

- OSCE از سال ۱۹۹۲ به عنوان قسمتی از آزمون MCCQE[1] (رزنیک و همکاران[2] ۱۹۹۲) و همچنین از سال ۲۰۰۵ به عنوان یکی از مراحل اجباری آزمون USMLE[3] در آمد (کیسی و همکاران[4] ۲۰۰۹). هرچند که ساختار آن با ساختار

1. Qualifying Examination of the Medical Council of Canada (MCCQE)
2. Reznick et al.
3. United States Medical Licensure Examination (USMLE)
4. Casey et al.

OSCE اولیه که توسط هاردن و همکاران پیشنهاد شده بود، دقیقاً یکسان نیست و حتی در مواردی از آن تحت عنوان شبهOSCE[1] نام برده می‌شود.

- نتایج یک پیمایش در امریکا نشان داد که در سال تحصیلی ۹۸-۱۹۹۷ از بین ۱۲۲ دانشکده معتبر[2]، ۴۹ دانشکده OSCE جامع[3] در پایان دوره برگزار می‌کردند و ۲۴ دانشکده در هیچ یک از مقاطع خود OSCE نداشتند (بارزانسکی و همکاران[4] ۱۹۹۸). تکرار مطالعه در سال تحصیلی ۱۰- ۲۰۰۹ نشان داد که از ۱۳۰ دانشکده، تنها یک دانشکده اصلاً OSCE برگزار نمی‌کرد و ۱۱۷ دانشکده OSCE جامع پایان دوره داشتند (بارزانسکی و اتزل[5] ۲۰۱۰).
- تعداد مطالعات صورت گرفته روی این آزمون روند رو به رشد داشته است. نتایج یک مطالعه مروری نظام‌مند نشان می‌دهد که در دهه اول معرفی OSCE (فاصله سال‌های ۱۹۷۵ تا ۱۹۸۴) متوسط تعداد مطالعات در مورد آن ۲/۱ در سال بوده و در دهه حاضر به ۶۱/۵ مطالعه در سال رسیده است (پاتریشیو و همکاران[6] ۲۰۱۳).
- این آزمون در پنج قاره و ۵۰ کشور جهان با فرهنگ‌ها و زمینه‌های مختلف مورد استفاده واقع شده است که تعداد قابل توجهی از آنها ناشی از همکاری بین چند دانشکده حتی از کشورهای مختلف بوده است (پاتریشیو و همکاران ۲۰۱۳).
- OSCE در رشته‌های مختلف مورد استفاده قرار گرفته است. بر اساس نتایج یک مطالعه مروری نظام‌مند، از مجموع ۱۰۶۵ مطالعه، ۸۷ درصد در حوزه پزشکی، پنج درصد پرستاری، سه درصد دندانپزشکی، دو درصد داروسازی و مابقی بین رشته‌ای بوده‌اند (پاتریشیو و همکاران ۲۰۱۳).
- اگرچه بر اساس مقاله مروری نظام‌مند پاتریشیو و همکاران، عمده موارد برگزاری OSCE در مقطع پزشکی عمومی[7] و دستیاری[8] (به ترتیب ۵۸ درصد و ۱۹ درصد مطالعات) بود اما به آنها محدود نمانده است و برای سایر مقاطع از جمله آموزش مدوام[9] و اخذ گواهینامه[10] برگزار شده است (پاتریشیو و همکاران ۲۰۱۳).
- در ایران نیز طی چند سال اخیر استفاده از آزمون OSCE برای ارزیابی مهارت‌های بالینی دانشجویان در رشته‌ها و مقاطع مختلف رواج یافته است. هرچند که در برخی از موارد آنچه واقعاً انجام می‌شود، شباهت چندانی به تعاریف پذیرفته شده OSCE ندارد از تجربه دانشگاه‌های کشور در زمینه برگزاری OSCE با کیفیت قابل قبول مقالاتی به چاپ رسیده است (تقوا و همکاران ۲۰۱۰، امینی و همکاران ۲۰۱۲، افتخار و همکاران ۲۰۱۲، میرزازاده و همکاران ۲۰۱۳، واثقی و همکاران ۲۰۱۳).

برای جمع‌بندی باید گفت که طی سه دهه اخیر از آزمون OSCE به کرات برای مقاصد گوناگون استفاده شده است. کاربردهای شایع OSCE در موارد زیر است:

- ارزیابی تکوینی عملکرد دانشجویان و ارائه بازخورد به آنها در طول یک دوره آموزشی
- ارزیابی عملکرد دانشجویان در پایان فازهای مختلف دوره برای اطمینان از دارا بودن حداقل استانداردهای صلاحیت بالینی برای ورود به فاز بعدی
- به عنوان قسمتی از آزمون جامع، سطح بالا[11] و مهم پایان دوره

1. OSCE-like
2. Accredited
3. Comprehensive
4. Barzansky et al.
5. Barzansky & Etzel
6. Patricio et al.
7. Undergraduate
8. Postgraduate
9. Continuous professional development
10. Licensure exam
11. High stake

- ارزیابی عملکرد فارغ‌التحصیلان برای دریافت مجوز و گواهینامه کار

علی‌رغم محبوبیت و استفاده فراوان OSCE نباید فراموش کرد که این آزمون به تنهایی نمی‌تواند وضعیت دانشجو را به طور کامل نشان دهد. در واقع، هیچ روش ارزیابی وجود ندارد که بتوان صرفاً با استفاده از آن به تصویری همه جانبه از عملکرد دانشجو دست یافت. برخی از جنبه‌های صلاحیت بالینی توسط OSCE قابل سنجش است. بنابراین توصیه می‌شود برای به دست آوردن نظری جامع در مورد سطح توانمندی یک دانشجو به جای استفاده از یک ابزار به صورت مقطعی، از ابزارهای متعدد و در طول زمان[1] استفاده شود (نیوبل[2] ۲۰۰۴، اپشتین[3] ۲۰۰۷).

انواع آزمون‌های ساختارمند عینی

شناخته‌شده‌ترین نوع آزمون در این خانواده همان OSCE است که در فصل بعدی به جزئیات آن خواهیم پرداخت. به صورت ساده می‌توان گفت OSCE از چند ایستگاه تشکیل شده است که دانشجویان به ترتیب آنها را پشت سر می‌گذارند و در هر یک از آنها معمولاً تحت نظارت یک آزمونگر کار خاصی که از آنها خواسته شده را انجام می‌دهند و مورد ارزیابی قرار می‌گیرند. با وجود اینکه اصول کلی و پایه‌ای OSCE همین است، باید گفت که ساختار استاندارد و کاملاً یکسانی برای برگزاری OSCE وجود ندارد. هر دانشکده و موسسه‌ای طبق ضوابط خود این کار را انجام می‌دهد و معمولاً تفاوت‌هایی از نظر تعداد ایستگاه‌ها، زمان ایستگاه‌ها، تعداد آزمونگران و ... بین OSCEهای مختلف وجود دارد.

با این حال، گاهی نحوه برگزاری آزمون یا عرصه‌ای که در آن برگزار می‌شود، آنقدر متفاوت است که دست‌اندرکاران آزمون ترجیح می‌دهند به جای لفظ OSCE، از نام‌های دیگری برای این امتحانات استفاده کنند؛ مانند:

OSCA[4]: این آزمون اولین بار توسط کالج سلطنتی جراحان استرالیا در سال ۱۹۹۰ مطرح شد و به تدریج تغییراتی در آن اعمال شد. اما به صورت کلی تعداد ایستگاه‌ها در این آزمون کمتر است (کمتر از ۵ ایستگاه) و زمان هر ایستگاه طولانی‌تر است (حدود ۱۵ تا ۲۰ دقیقه). آنچه دانشجو در ایستگاه با آن مواجه می‌شود، یک مهارت جزئی و اختصاصی نیست؛ بلکه در مواجهه کامل با یک بیمار قرار می‌گیرد و باید چندین توانمندی را به صورت همزمان نشان دهد. در واقع محتوای این ایستگاه‌ها ادغام‌یافته است و شامل مهارت ارتباطی، شرح‌حال، معاینه، ثبت و جمع‌بندی داده‌ها، استدلال بالینی و تهیه برنامه تشخیصی–درمانی می‌باشد (بوژاک[5] و همکاران ۱۹۹۱، سرپل[6] ۲۰۰۹). مشابه این آزمون در قسمت مهارت های بالینی آزمون USMLE (CS) استفاده می‌شود. هرچند به نظر می‌رسد این ابزار بسیار مشابه OSCE است، برخی ترجیح داده‌اند از نام دیگری برای متمایز کردن آن استفاده کنند. مقاله دیگری که به ذکر تجربه مشابه در دانشگاه تایوان در خصوص برگزاری آزمون ساختارمند عینی پرداخته است، از آن به عنوان iOSCE[7] نام برده است که به معنای OSCE ادغام‌یافته می‌باشد. این آزمون شامل دو مورد بالینی بود که هر یک در قالب پنج ایستگاه مرتبط مطرح شد و آزمونگران آن به ارزیابی عملکرد دانشجو در حیطه‌های مختلف پرداختند (لین و همکاران[8] ۲۰۱۳).

- OSATS[9]: این آزمون توسط دانشگاه تورونتو در سال ۱۹۹۰ برای ارزیابی مهارت‌های پروسیجرال و تکنیکی و ارائه بازخورد به دستیاران مخصوصاً در رشته‌های جراحی توسعه پیدا کرد. دستیاران باید چند ایستگاه را پشت سر می‌گذاشتند که در هر یک، از آنها خواسته شده بود طی زمان مشخصی (معمولاً ۱۵ تا ۲۰ دقیقه) پروسیجرهایی مانند

1. Longitudinal
2. Newble
3. Epstein
4. Objective Structured Clinical Assessment
5. Bujack et al.
6. Serpell
7. Integrated OSCE
8. Lin et al
9. Objective Structured Assessment of Technical Skill

بخیه، جراحی باز یا جراحی لاپارسکوپیک را انجام دهند. آزمونگران مستقر در ایستگاه عملکرد دستیاران را مشاهده می‌کردند و هم بر اساس چک‌لیست و هم با نمره‌دهی گلوبال به ارزیابی آنها می‌پرداختند (مارتین و همکاران[1] ۱۹۹۷، ون‌هوف و همکاران[2] ۲۰۱۰، چیپمن و اشمیتز[3] ۲۰۰۹). چک‌لیست برای ارزیابی تک‌تک مراحل تکنیک یا پروسیجر مورد نظر استفاده می‌شد و آیتم‌هایی داشت که به صورت یک و صفر نمره داده می‌شدند. مثلاً: هر چند که مبنای این آزمون استفاده از مدل‌های حیوانی و شبیه‌سازی شرایط واقعی بود، مقالاتی منتشر شده‌اند که به ارائه تجربه OSATS در اتاق عمل نیز پرداخته‌اند (نیتسو و همکاران ۲۰۱۲).

در حالی که نمره‌دهی گلوبال، برای موارد کلی‌تر که وابسته به یک مهارت خاص نبودند، به صورت لیکرت استفاده می‌شد. مثلاً:

نوع درست نخ بخیه را انتخاب کرد. بله☐ خیر☐
طی انجام پروسیجر، زمان‌بندی درستی داشت و از حرکات بی مورد اجتناب کرد. ☐ همیشه ☐ اکثر مواقع ☐ گاهی ☐ هرگز
یا:
طی انجام پروسیجر، مراقب بافت بود و با استفاده مناسب از ابزار به بافت آسیب نرساند. ☐ همیشه ☐ اکثر مواقع ☐ گاهی ☐ هرگز

- OSVE[4]: این آزمون برای ارزیابی مهارت‌های ارتباطی دانشجویان به شیوه ارزان و قابل اجرا طراحی شده است. از آنجا که مهارت‌های ارتباطی دانشجویان در مواجهه با بیماران متفاوت، متغیر است، برای ارزیابی درست مهارت ارتباطی لازم است تعداد مواجهات یعنی تعداد ایستگاه‌ها زیاد باشد که این امر موجب افزایش هزینه OSCE می‌شود. در OSVE تعدادی ویدئوی ضبط‌شده از مواجهه پزشک با بیماران به دانشجویان نشان داده می‌شود و از آنها خواسته می‌شود به تعدادی سؤال کتبی (بسته پاسخ با بازپاسخ) در مورد ارتباط صورت گرفته پاسخ دهند. این آزمون در محیط کلاس قابل اجرا است و نیاز به تمهیدات خاص و پرهزینه‌ای ندارد. برخی مطالعات با طراحی سؤالات مخصوص و به کار بردن روش‌های تصحیح خاص، تلاش کرده‌اند تا اسکریپت‌های نهان را از بین رفتار آشکار دانشجویان شناسایی و ارزیابی کنند (هامفریس و کنی[5] ۲۰۰۰، باریبو و همکاران[6] ۲۰۱۲). علی‌رغم شباهت اسمی شاید این روش را بتوان بیشتر جزء آزمون‌های کتبی در نظر گرفت تا آزمون‌های بالینی.
- TOSCE[7]: در این آزمون که با توجه به اهمیت و ضرورت کار تیمی و همکاری بین حرفه‌ای توسعه یافته است، سناریو به گونه‌ای طراحی و اجرا می‌شود که یک تیم متشکل از چهار تا پنج دانشجو باید با یکدیگر کار کنند و هر یک وظیفه مشخصی را برای مراقبت از بیمار انجام دهند. آزمونگران عملکرد آنها را مشاهده می‌کنند و در انتها در مورد نحوه تعامل آنها با یکدیگر و مهارت‌های بالینی و حرفه‌ای به آنها بازخورد می‌دهند (سینگلتن و همکاران[8] ۱۹۹۹، سیموندز و همکاران[9] ۲۰۰۳، امینی و همکاران ۲۰۱۲). برخی از مقالات به ارائه تجربه خود از برگزاری آزمونی مشابه اما با نام

1. Martin et al
2. van Hove et al
3. Chipman & Schmitz
4. Objective Structured Video Exam
5. Humphris & Kaney
6. Baribeau et al.
7. Team Objective Structured Clinical Examination
8. Singleton et al.
9. Symonds et al.

GOSCE[1] پرداخته‌اند (بیران[2] ۱۹۹۰، الیوت[3] ۱۹۹۴).

- OSPE[4] یا OSLE[5]: این آزمون برای عرصه‌هایی مانند آزمایشگاه یا کار عملی دروس علوم پایه طراحی شده است که دانشجو مهارت‌های عملی انجام می‌دهد اما لزوماً در مواجهه با بیمار و کار بالینی قرار نمی‌گیرد. به عنوان مثال، ایستگاه‌های این امتحان شامل این موارد هستند: مشاهده بافت زیر میکروسکوپ، آماده‌سازی لام خون محیطی، رنگ‌آمیزی لام باکتری و انجام آزمایش‌های بیوشیمی (هاردن و کرنکراس[6] ۱۹۸۰، دیسانایکه و همکاران[7] ۱۹۹۰، آبراهام و همکاران[8] ۲۰۰۹، کوندو و همکاران[9] ۲۰۱۳).
- OSTE:[10] این ابزار در واقع مرتبط با توانمندی‌های آموزشی استادان (و دستیاران) است که اگرچه برای ارزشیابی اعضای هیأت علمی می‌تواند به کار رود، با گنجاندن بازخورد و تمرین بیشتر در برنامه‌های توانمندسازی و به منظور تقویت مهارت‌های آموزشی اعضای هیأت علمی کاربرد دارد. مخصوصاً از آن جا که مشاهده تدریس استاد توسط سایر استادان و متخصصان آموزشی معمول نیست، با OSTE فرصت خوبی برای این کار فراهم می‌آید. این آزمون از تعدادی ایستگاه تشکیل شده است که برای هرکدام یک سناریوی آموزشی نوشته شده است و به جای بیمار، یک دانشجونما یا دانشجوی استاندارد[11] در ایستگاه به ایفای نقش می‌پردازد (استون و همکاران[12] ۲۰۰۳، جولیان و همکاران[13] ۲۰۱۲، بویلات و همکاران[14] ۲۰۱۲).

شایان ذکر است که پاتریشیو و همکاران در مطالعه مروری نظام‌مند خود به این مسأله اشاره دارند که این تعدد نامگذاری، احتمالاً به جز موارد معدودی مانند OSVE و TOSCE که واقعاً تفاوت‌های ساختاری مشخصی با OSCE اولیه دارند، ضروری نیست و برای بقیه موارد می‌توان از همان OSCE استفاده کرد (پاتریشیو و همکاران ۲۰۰۹).

از این بین در این کتاب، ما به تشریح آزمون OSCE خواهیم پرداخت و با توجه به شباهتهای ساختاری قابل توجه از پرداختن به جزئیات آنها صرف نظر شده است.

مزایا و محدودیت‌های آزمون‌های ساختارمند عینی

بی‌شک هیچ روش ارزیابی وجود ندارد که به طور کامل بی‌نقص باشد. هر کدام از ابزارهایی که برای سنجش دانش، مهارت و توانمندی دانشجویان به کار می‌روند، دارای مزایا و در عین حال محدودیت‌هایی هستند که موجب می‌شود بسته به شرایط قابل استفاده و بهره‌برداری باشند. یکی از رویکردهای خوب برای تصمیم‌گیری در مورد یک ابزار، مقایسه مزایا و محدودیت‌های آن با سایر ابزارها است.

1. Group Objective Structured Examination
2. Biran
3. Elliot et al.
4. Objective Structured Practical Examination
5. Objective Structured Laboratory Examination
6. Harden & Cairncross
7. Dissanayake et al.
8. Abraham et al
9. Kundu et al.
10. Objective Structured Teaching Exercise
11. Standardized student
12. Stone et al.
13. Julian et al
14. Boillat et al

مزایای آزمون‌های ساختارمند عینی

راشفورث[1] مزایای OSCE را به این شکل جمع‌بندی می‌کند (راشفورث ۲۰۰۷):

- قابلیت ارزیابی طیف متنوعی از مهارت‌ها در محیطی مشابه محیط واقعی
- عینیت بیشتر نسبت به اکثر ابزارهای ارزیابی
- تنوع آزمونگران که خود باعث کاهش سوگیری می‌شود
- یکسان بودن سؤالات و شرایط آزمون برای همه دانشجویان
- نگرش مثبت فراگیران و اساتید نسبت به آن
- ایجاد انگیزه یادگیری در دانشجویان
- پایایی و روایی بالا

محدودیت‌های آزمون‌های ساختارمند عینی

با توجه به مزایای ذکر شده، برخی OSCE را روش استاندارد طلایی برای ارزیابی فراگیران در حیطه سلامت و پزشکی می‌دانستند اما امروزه مشخص شده است که لزوماً تمام موارد فوق در تمام OSCEهای برگزار شده صدق نمی‌کند. به عنوان مثال، روایی و پایایی OSCE، عدد یا مقدار ثابتی نیست و به سؤالات آزمون، آزمونگران، شرایط آزمون و عوامل متعدد دیگر بستگی دارد و در هر بار آزمون باید مورد بررسی قرار گیرند (راشفورث ۲۰۰۷). از طرف دیگر، موارد زیر به عنوان مشکلات OSCE در متون و مقالات مطرح شده است:

- علی رغم اینکه دانشجویان نسبت به OSCE پذیرش خوبی دارند، استرس و اضطرابی که در OSCE بر آنها تحمیل می‌شود، قابل انکار نیست و می‌تواند بر عملکرد آنها تاثیر بگذارد. اگرچه برخی از نویسندگان معتقدند این استرس نشان دهنده روایی بالای آزمون است زیرا مواجهه با بیماران در محیط واقعی نیز استرس‌زا است (بوژاک و همکاران ۱۹۹۱، بارتفای و همکاران[2] ۲۰۰۴).
- آزمون OSCE پرهزینه است و به امکانات و تجهیزات زیادی احتیاج دارد. مسائلی مانند امنیت آزمون یا یکسان کردن ایستگاه‌های موازی از مواردی است که به پیچیدگی آن می‌افزاید. همچنین وقت زیادی از اعضای هیأت علمی باید صرف طراحی و اجرای آزمون شود (راشفورث ۲۰۰۷).
- مسأله دیگری که در کنار هزینه مطرح می‌شود ولی لزوماً با آن یکسان نیست، قابلیت اجرای آزمون است. به عنوان مثال دسترسی به بیماران واقعی یا بیماران استاندارد چندان آسان نیست و نیاز به برنامه‌ریزی دقیق دارد. مخصوصاً برای OSCE که در بخش خاصی مثل اطفال برگزار می‌شود، استفاده از کودکان در زمان طولانی می‌تواند خسته‌کننده باشد و روی همکاری آنها اثر بگذارد (کاراسیو و انگلندر[3] ۲۰۰۰).
- مسأله دیگر، محیط شبیه‌سازی شده در OSCE است که می‌تواند بر عملکرد دانشجویان اثر بگذارد و موجب شود برای یک مهارت معین، دقیقاً همان رفتاری که در محیط واقعی دارند، در OSCE از خود نشان ندهند (خان و همکاران ۲۰۱۳).
- نکته مشابه دیگر که شاید تا حدی به محدودیت قبلی برگردد این است که برخی از مهارت‌ها که اتفاقاً نقش مهمی در عملکرد در محیط واقعی دارند. به عنوان مثال همکاری در کار تیمی، مدیریت منابع، توانایی رهبری و ... به راحتی توسط OSCE قابل سنجش نیستند (خان و همکاران ۲۰۱۳).

1. Rushforth
2. Bartfay et al
3. Carraccio & Englander

- یکی دیگر از انتقاداتی که به OSCE وارد می‌شود، از بین رفتن دید جامع‌نگر و کل نگر[1] به توانمندی هنگام ارزیابی آن است. در واقع، این عقیده وجود دارد که شکستن توانمندی مراقبت بیمار به عنوان یک کل درهم‌تنیده به اجزای جداگانه کار چندان مطلوبی نیست و روایی ابزار را کاهش می‌دهد (نستل و همکاران[2] ۲۰۱۱). منفک کردن و شکستن توانمندی به اجزای آن، علاوه بر اینکه ناشی از ناتوانی OSCE در ارزیابی تمام مهارت‌ها است (محدودیت قبلی)، به این دلیل نیز رخ می‌دهد که تلاش می‌شود مهارت‌هایی در OSCE مورد ارزیابی قرار گیرند که توسط سایر ابزارها مانند آزمون کتبی یا کامپیوتری قابل سنجش نیستند. در همین راستا، سعی می‌شود سؤالات OSCE در مورد میزان دانش افراد یا میزان درک آنها از کاربرد دانش نباشد. بنابراین اگر به عنوان مثال طراح سؤال بخواهد توانمندی فراگیر را در تست پاپ اسمیر ارزیابی کند، ایستگاهی طراحی می‌کند که صرفاً مهارت عملی دانشجو را در اجرای تست بسنجد در حالی که اندیکاسیون تست و همچنین نحوه تفسیر آن از موارد مهم دیگری هستند که به صورت کلی این توانمندی را شکل می‌دهند ولی چون احتمالاً در آزمون‌های کتبی قابل ارزیابی هستند، در OSCE گنجانده نمی‌شوند.

در هر حال، مشخص است که OSCE نقاط قوت قابل توجهی دارد که قابل چشم‌پوشی نیستند و البته هنگام استفاده از آن باید مراقب محدودیت‌ها و مشکلات احتمالی بود. در این بین، روایی و پایایی آزمون اهمیت ویژه‌ای دارند، مخصوصاً اگر قرار است از نمرات و نتایج OSCE برای تصمیم‌های مهم و سرنوشت‌ساز استفاده شود.

سودمندی آزمون‌های ساختارمند عینی

همان‌طور که در فصول پیشین اشاره شد، سودمندی یک ابزار با پنج ویژگی شامل روایی، پایایی، مقبولیت، هزینه و تاثیر آموزشی بررسی می‌شود. از آنجا که اکثر مطالعات و مقالاتی که در مورد سودمندی آزمون‌های ساختارمند عینی منتشر شده‌اند، به سودمندی OSCE پرداخته‌اند، برای جلوگیری از تکرار مطالب، این مبحث در فصل بعدی ارائه خواهد شد.

سؤالات رایج آزمون‌های ساختارمند عینی

آیا OSCE آزمونی عینی و پایا است؟

این مسأله که آیا OSCE یک ابزار عینی و پایا هست، موضوع بسیاری از مقالات و پژوهش‌ها بوده است. نکته اول اینکه نمی‌توان به صورت کلی گفت یک نوع ابزار پایایی بالایی دارد یا پایین. زیرا OSCE به شکل‌های متنوع برگزار می‌شود. تعداد ایستگاه‌ها، زمان ایستگاه‌ها، تعداد آزمونگران و بسیاری از مسائل دیگر متفاوت است و به همین دلیل پایایی هر آزمونی که برگزار می‌شود، منحصر به فرد است و باید پس از برگزاری آزمون محاسبه شود. حتی OSCEهایی که یک دانشکده در سال‌های متوالی برگزار می‌کند، از نظر ویژگی‌های سایکومتریک از جمله پایایی یکسان نیست.

نکته بعدی که باید شفاف شود مفهوم پایایی در OSCE است. می‌دانیم که منظور از پایایی آزمون، کم بودن سهم خطای تصادفی و میزان تکرارپذیری نتایج آن است. باید این مسأله مهم را خاطرنشان کرد که منظور از نتایج OSCE، در بیشتر دانشکده‌ها تصمیم‌گیری‌های رد یا قبول است تا خود نمرات خام آزمون. به عبارت دیگر، وقتی از پایایی OSCE صحبت می‌شود به صورت اولیه مهم است که آیا دانشجویی که در آزمون قبول (یا رد) شده است، در صورت تکرار آزمون آیا باز هم قبول (یا رد) می‌شود؟ زیرا در آزمون‌های مبتنی بر توانمندی، هرچند که خوب است سطح توانمندی فراگیر دقیقاً مشخص

1. Holistic
2. Nestel et al.

شود، اما آنچه بیشتر حائز اهمیت است این موضوع است که آیا به سطح قابل قبولی از توانمندی مورد نظر رسیده است یا خیر (ون‌درولوتن و سوانسون ۱۹۹۰).

مسأله بعدی ارتباط مفهوم پایایی و عینیت آزمون است. تا قبل از OSCE انتقاداتی به آزمون‌های سنتی مانند آزمون شفاهی و «مورد بالینی کامل» وارد بود زیرا از ساختار مشخص و یکسان پیروی نمی‌کردند و ذهنیت آزمونگران در آنها نقش زیادی داشت. با معرفی OSCE به نظر می‌رسید عمده‌ترین مزیت آن، داشتن ساختار مشخص و چک‌لیست‌های اختصاصی است که منجر به آزمونی عینی و استاندارد می‌شود. تا مدت‌ها تصور بر این بود که هر چه ابزاری بیشتر واجد این دو خصوصیت باشد، پایایی بیشتری دارد. اما شواهد بعدی نشان دادند که پایایی آزمون بیشتر تحت تاثیر نمونه‌گیری دقیق و وسیع از محتوای بالینی مورد نظر است تا ساختارمند بودن و عینی بودن آن (ون‌درولوتن و سوانسون ۱۹۹۰، سوانسون و همکاران ۱۹۹۵، نورمن و همکاران ۲۰۰۶).

مسأله عینیت در OSCE جنبه دیگری نیز دارد که به ارزیابی بر مبنای چک‌لیست و نمره‌دهی گلوبال برمی‌گردد و در چالش دیگری بحث خواهد شد اما در همینجا از یک منظر دیگر آن را توضیح می‌دهیم: از جمله عواملی که به عنوان منبع خطا بر پایایی OSCE موثر هستند، خستگی دانشجویان، استرس زیاد، سوگیری شخصی آزمونگران، عملکرد متفاوت بیمارنماها، زبان، جنسیت و نژاد می‌باشد (شونیم–کلین و همکاران ۲۰۰۸، ترنر و دانکوسکی ۲۰۰۸). اما بر اساس نتایج مطالعات از جمله یک مطالعه مروری نظام‌مند، مشخص شد که منبع اصلی خطا در OSCE، ناشی از تفاوت عملکرد دانشجو در ایستگاه‌های مختلف است. به عبارت دیگر، هنگامی که تعداد ایستگاه‌ها کم است، تنوع موضوعاتی که می‌توان از آنها سؤال طرح کرد، کاهش پیدا می‌کند. به عبارت دیگر ویژگی مورد یا ویژگی محتوا وجود دارد که قابلیت تعمیم و پایایی را کاهش می‌دهد. یعنی اگر دانشجویی در یک ایستگاه پروسیجر مانند بخیه زدن به خوبی عمل کرد، نمی‌توان نتیجه گرفت که در ایستگاه گرفتن شرح حال یا حتی پروسیجر دیگری مانند تعبیه سونداژ نیز عملکرد خوبی دارد. به همین علت، باید به تعداد کافی ایستگاه در نظر گرفت تا بتوان حیطه‌ها و موضوعات مختلف را پوشش داد (ون‌درولوتن و سوانسون ۱۹۹۰، سوانسون و همکاران ۱۹۹۵، نورمن و همکاران ۲۰۰۶). به این ترتیب، پایایی OSCE بیشتر تحت تاثیر نمونه‌گیری دقیق و وسیع از محتوای بالینی مورد نظر است. به عبارت دیگر، هرچه تعداد ایستگاه‌ها بیشتر و زمان OSCE طولانی‌تر باشد، نتایج آن قابلیت تعمیم بیشتری دارند و پایاتر خواهند بود. البته این یافته محدود به OSCE نیست و در مورد تمام ابزارهای ارزیابی دانشجو از جمله آنها که بیشتر ذهنی‌گرا و کمتر استاندارد و ساختارمند هستند نیز صدق می‌کند (ون‌درولوتن و شوورث[1] ۲۰۰۵).

حال سؤال این است که چه تعداد ایستگاه برای OSCE کافی است؟ به این موضوع در قسمت بعدی پرداخته می‌شود.

چه تعداد ایستگاه برای OSCE کافی است؟

می توان تصور کرد که هرچه تعداد ایستگاه های OSCE بیشتر باشد، نمونه‌گیری بهتری از محتوا می‌توان داشت و بنابراین روایی محتوایی بهبود می‌یابد. همچنین با توجه به مسأله‌ای که در خصوص ویژگی محتوا و اثر آن روی پایایی گفته شد، باز هم افزایش تعداد ایستگاه‌ها مفید است. اما افزودن ایستگاه، عواقبی نیز به دنبال دارد؛ از جمله اینکه نیاز به بیمارنما و آزمونگر، زمان آزمون و هزینه برگزاری آزمون را زیاد می‌کند. بنابراین سؤال این است که حداقل تعداد ایستگاهی که می‌توانیم داشته باشیم تا روایی و پایایی مطلوبی تامین شود، چند عدد است؟

در دانشکده‌های مختلف تعداد ایستگاه‌هایی که برای یک OSCE در نظر گرفته می‌شود، متغیر است. این مسأله به هدف آزمون و شرکت کنندگان آزمون وابسته است. آزمون OSCE که به عنوان آزمون پایان یکی از بخش‌های کارآموزی طراحی می‌شود، خود به خود موضوعات کمتری برای ارزیابی دارد. در حالی که برای OSCEهای جامع که پایان یک

1. Van der Vleuten & Schuwirth

مقطع برگزار می‌شوند و از چندین درس و کورس تشکیل شده‌اند، می‌توان ایستگاه‌های بیشتری در نظر گرفت (کیسی و همکاران ۲۰۰۹).

به طور معمول یک OSCE جامع حداقل ۱۰ ایستگاه و حداکثر ۲۰ ایستگاه دارد (اسمی ۲۰۰۳، گرملی ۲۰۱۱). البته ایستگاه‌ها ممکن است کمتر یا بیشتر از این تعداد هم باشند؛ به طوری که یک مقاله به گزارش برگزاری OSCE با ۴۲ ایستگاه برای دوره دستیاری کودکان پرداخته است (جورابچی ۱۹۹۱). اما کمتر از ۱۰ ایستگاه برای پوشش دادن طیف مناسبی از مهارت‌ها ناکافی است و قابلیت اجرای بیشتر از ۲۰ ایستگاه نیز پایین است (سلبی و همکاران ۱۹۹۵). تعداد دانشجویان و تعداد هیأت علمی در دسترس نیز عواملی هستند که باید در نظر گرفته شوند.

همچنین، توجه به این نکته ضروری است که تعداد بهینه ایستگاه تحت تاثیر هدف آزمون قرار می گیرد: آیا با این آزمون قرار است دانشجویان رتبه‌بندی شوند؟ آیا صرفاً نتیجه رد و قبول برای هرکس اعلام می‌شود؟ یا اینکه نمره خام تک تک دانشجویان هم مهم است؟ به طور مثال، یک مطالعه که داده‌های چهار OSCE دندانپزشکی با ۱۶ تا ۱۸ ایستگاه را از طریق نظریه تعمیم‌پذیری بررسی کرده بود، به این نتیجه رسید که اگر هدف آزمون را رتبه‌بندی در نظر می‌گرفتند، ۱۲ ایستگاه کافی بوده اما برای دو هدف دیگر، نیاز به ۱۷ ایستگاه وجود داشته است (شونیم–کلین و همکاران ۲۰۰۸).

زمان OSCE چقدر باید باشد؟

تعداد ایستگاه‌ها و زمان آنها در مطالعات مختلف بسیار متفاوت است که منجر می‌شود زمان کلی آزمون نیز متغیر باشد. مدت زمانی که به هر ایستگاه OSCE اختصاص می‌یابد، در دانشگاه‌های مختلف بین ۴ تا ۲۰ دقیقه گزارش شده است اما رایج‌ترین زمان پنج دقیقه است (هاردن ۱۹۹۰، اسمی ۲۰۰۳). از آنجا که زمان آزمون بر پایایی آزمون موثر است، این سؤال وجود دارد که زمان ایستگاه و زمان کلی آزمون چقدر باید باشد تا ضریب پایایی خوبی (که از نظر مقدار بالای ۰/۸ است) برای OSCE داشته باشیم.

در یک مطالعه مروری نظام‌مند، پژوهشگران به آنالیز مجدد داده‌های مقالات از طریق نظریه تعمیم‌پذیری پرداختند تا پایایی نتایج آزمون‌ها را بررسی کنند. زمان ایستگاه در ۱۳ مطالعه به دست آمده، بین ۶ تا ۴۰ دقیقه متغیر بود. برای یکسان کردن تفاوت مطالعات مختلف همه آنالیزها، نه بر مبنای تعداد ایستگاه‌ها یا زمان هر یک از ایستگاه‌ها، بلکه در واحد زمان کلی آزمون انجام شد (یک، دو، سه، چهار، شش، هشت و دوازده ساعت). نتایج نشان داد که اگر آزمون یک ساعت باشد، ضریب تعمیم‌پذیری بین ۰/۱۹ تا ۰/۶۵ در مطالعات مختلف متغیر بوده است. با احتساب آزمون ۱۲ ساعته، ضریب پایایی بین ۰/۷۳ تا ۰/۹۶ به دست آمد اما در هر حال، محتوای آزمون‌ها با یکدیگر تفاوت داشت و همچنین منابع خطای آنها یکی نبود بنابراین نویسندگان ذکر کردند که نتایج تنها به صورت کلی و با اغماض تفاوت‌های دیگر، منطقی و مفهوم است (ون‌درولوتن و سوانسون ۱۹۹۰).

یافته دیگری که نویسندگان این مقاله مروری یافتند، این بود که در آزمون‌هایی که دو ایستگاه آن به هم مرتبط هستند، مثلاً در یک ایستگاه شرح حال و معاینه انجام می‌شود و در ایستگاه بعد دانشجو باید به یک سری سؤال کتبی جواب دهد یا پرونده بیمار را بنویسد، علی‌رغم افزایش زمان کلی آزمون، ضریب پایایی کمتر بوده است. یعنی هرچند استفاده از ایستگاه‌های مرتبط موجب می‌شود بتوانیم علاوه بر مهارت‌های عملی و تکنیکی، حیطه دانشی و شناختی را نیز ارزیابی کنیم و در نتیجه احتمالاً روایی آزمون افزایش پیدا می‌کند، اما با محدود شدن نمونه‌گیری، قابلیت تعمیم و پایایی آزمون کاهش می‌یابد (ون‌درولوتن و سوانسون ۱۹۹۰).

به جز مسأله مربوط به ایستگاه‌های مرتبط، نویسندگان ارتباط دیگری بین زمان یک ایستگاه و ضریب پایایی آزمون نیافتند. بسته به آزمونی که برگزار شده بود هم ایستگاه‌های کوتاه و هم ایستگاه‌های بلند گاهی منجر به پایایی بالا و گاهی منجر به پایایی پایین شده بودند. البته نویسندگان بیان کردند که آزمون‌ها با ایستگاه‌های خیلی طولانی (یک ساعته)

به علت کم بودن تعداد ایستگاه‌ها، آزمونگران و بیمارنمایان، منابع خطای دیگری دارند که در نهایت پایایی خوبی به دنبال ندارد. از طرفی، ایستگاه‌های خیلی کوتاه (یک تا دو دقیقه‌ای)، از دید آموزشی قادر نیستند مهارت‌ها و توانمندی‌های مناسبی را به خوبی ارزیابی کنند (ون‌درولوتن و سوانسون ۱۹۹۰).

به صورت کلی تصمیم‌گیری در مورد زمان یک ایستگاه بیشتر تابع روایی محتوایی و آن چیزی است که قرار است مورد سنجش قرار گیرد. زمان ایستگاه به مهارت و وظیفه مورد سنجش در ایستگاه بستگی دارد. اگر تردیدی وجود دارد که آیا دانشجو می‌تواند در زمان تعیین شده، مهارت را به اتمام برساند یا خیر، بهتر است تعدیل در زمان ایستگاه‌ها یا در وظیفه خواسته شده صورت گیرد.

در هر ایستگاه چه تعداد آزمونگر لازم است؟

یکی دیگر از باورهای رایج مربوط به OSCE این است که بهتر است بیش از یک نفر در ایستگاه به ارزیابی دانشجو بپردازد. ادعا می‌شود که در این حالت می‌توان با میانگین گرفتن از نظرات آزمونگران به ارزیابی دقیق‌تر و در نتیجه پایایی بیشتری دست یافت. اما نتایج یک مطالعه که در هر ایستگاه از دو آزمونگر استفاده کرده بود و پایایی آزمون را از طریق نظریه تعمیم‌پذیری به دست آورده بود، مشخص نمود که افزایش تعداد آزمونگر در هر ایستگاه، اثر بسیار کمی روی پایایی آزمون دارد (سوانسون و نورسینی ۱۹۸۹). همچنین نتایج یک مطالعه مروری نشان داد توافق بین آزمونگران در یک ایستگاه، عموماً بالا بوده است (ون‌درولوتن و سوانسون ۱۹۹۰).

به همین دلیل توصیه می‌شود که لزومی ندارد در یک ایستگاه بیش از یک آزمونگر داشته باشیم. در واقع حضور دو آزمونگر برای یک ایستگاه به نوعی ائتلاف منابع نیز محسوب می‌شود. اگر آزمونگر به تعداد زیاد در دسترس است، موثرتر است که تعداد ایستگاه‌های OSCE زیاد شود به جای اینکه در هر ایستگاه دو نفر به ارزیابی بپردازند. در واقع افزایش تعداد ایستگاه‌ها به علت کاهش همان مسأله ویژگی محتوا بیشتر روی بهبود پایایی موثر است تا تعداد آزمونگران هر ایستگاه.

نمره‌دهی توسط چک‌لیست بهتر است یا نمره‌دهی گلوبال؟

در اوایل معرفی OSCE، یکی از نقاط قوت آن که مخصوصاً در مقایسه با روش‌های قبلی بسیار مورد توجه قرار گرفت، وجود چک‌لیست بود. امروزه همچنان این فرض وجود دارد که در صورت نبود چک‌لیست، آزمونگران مجبور هستند به صورت کلی و بر اساس قضاوت ذهنی خود به فراگیر نمره دهند که احتمالاً موجب می‌شود ارزیابی‌ها خیلی شخصی شوند و هر کس به زعم خود به دانشجو نمره دهد و پایایی بین آزمونگران مطلوب نباشد. در حالی که اگر ارزیابی عملکرد دانشجو بر اساس آیتم‌های چک‌لیست باشد، آزمونگر را برای ارزیابی مهارت مورد نظر بهتر راهنمایی می‌کند و توافق بین آزمونگران و پایایی زیاد می‌شود.

آنچه استفاده از این رویکرد را تقویت می‌کند، یک باور عمومی شایع است که هر چقدر یک مهارت به ریزمهارت‌های آن شکسته شود، ارزیابی آن مهارت عینی‌تر خواهد بود و به افزایش عینیت و در نتیجه پایایی کمک می‌کند. به عنوان مثال، «دانشجو شکم بیمار را معاینه کرد»، خیلی کلی است و بهتر است به آیتم‌های جزئی‌تر مانند «چهار ربع شکم را لمس کرد»، «دق شکم را انجام داد»، و ... شکسته شود. در حالی که اکنون دیدگاه دیگری رواج یافته است که بر داشتن دید کلی و جامع هنگام ارزیابی تاکید می‌کند و باور فوق را به چالش می‌کشد. گفته می‌شود که به صرف ذهنی بودن ارزیابی نمی‌توان آن را بی‌ارزش دانست و عینی بودن قضاوت لزوماً ارزشمند نیست. در واقع، با فراگیر شدن OSCE به تدریج مشخص شد که چک‌لیست مخصوصاً اگر با دقت کافی طراحی نشود، مشکلاتی به همراه دارد:

اولاً تلاش برای دستیابی به عینیت بالا، موجب توجه زیاد به جزئیات می‌گردد که از آن تحت عنوان جزئی‌سازی[1] نام

1. Trivialization

برده می‌شود (نیوبل ۲۰۰۴، پل و همکاران[1] ۲۰۱۰). در واقع خیلی شایع است که طراح سؤال تلاش کند مهارت‌ها را به تعداد زیادی آیتم ریز کند. در این حالت، این نگرانی وجود دارد که واقعاً منعکس کننده مهارت نباشند و روایی را مخدوش کنند. این مسأله مخصوصاً هنگامی که وزن آیتم‌های مختلف با یکدیگر یکسان در نظر گرفته می‌شود، تشدید می‌شود. برای جلوگیری از این امر باید تلاش شود تا آیتم‌هایی که واقعاً اهمیت دارند و قادر هستند عملکرد دانشجوی ضعیف و قوی را از هم افتراق دهند، در چک‌لیست گنجانده شوند.

ثانیاً تمایل دانشجویان به حفظ کردن آیتم‌های چک‌لیست به جای درک عمیق آنها یکی از پیامدهای منفی آن است (بورسیکوت و همکاران[2] ۲۰۱۱). با استفاده از رویکرد قطعه‌قطعه‌سازی[3] که چند آیتم جزئی را در قالب یک آیتم کلی‌تر بیان می‌کند یا نمره‌دهی گلوبال، این مزیت وجود دارد که دانشجو نمی‌تواند صرفاً با حفظ کردن آیتم‌ها نمره خوبی دریافت کند.

بنابراین، تا اینجا مشخص شد که نمره‌دهی توسط چک‌لیست ریز آن اندازه که قبلاً تصور می‌شد، مطلوب نیست. اکنون باید دید که وضعیت نمره‌دهی گلوبال چگونه است. شواهد تجربی نشان می‌دهند اگر نمره‌دهی گلوبال در چارچوب یک مهارت ساختارمند و توسط فرد متخصص آموزش دیده و مطلع انجام شود و تاکید کافی بر مسأله آموزش و توجیه آزمونگران صورت گیرد، می‌تواند به همان اندازه یا بیشتر نتایج پایا تولید کند (کونینگتن و همکاران[4] ۱۹۹۷، رگر و همکاران[5] ۱۹۹۸، شوارتز و همکاران[6] ۱۹۹۹، پارک و همکاران[7] ۲۰۰۴، مالائو–عدولی و همکاران[8] ۲۰۱۲). و احتمالاً علاوه بر پایایی، واجد روایی بهتری در مقایسه با چک‌لیست باشد (هاجز و مک لیروی[9] ۲۰۰۳).

کونینگتن و همکاران ۱۹۹۷

در این مطالعه برای مقایسه چک‌لیست با نمره‌دهی گلوبال، ۹۶ دانشجوی پزشکی در یک OSCE با ۸ ایستگاه شرکت کردند. ۳۹ دانشجو از طریق چک‌لیست جزئی و ۵۷ دانشجو از طریق نمره‌دهی گلوبال مورد سنجش قرار گرفتند. یک گروه ۳۹ نفره از دانشجویان توسط دو آزمونگر مستقل ارزیابی شدند. پایایی بین آزمونگران و پایایی بین ایستگاه‌ها در ارزیابی چک‌لیست و گلوبال مشابه بود.

رگر و همکاران ۱۹۹۸

در این مطالعه، در یک OSCE با ۸ ایستگاه ۱۵ دقیقه‌ای برای دستیاران جراحی، ویژگی‌های سایکومتریک چک‌لیست و گلوبال مقایسه شد. در هر ایستگاه دو نفر جراح حاضر بودند که یکی از آنها ابتدا با چک‌لیست و سپس به صورت گلوبال به ارزیابی می‌پرداخت و دیگری صرفاً نمره‌دهی گلوبال انجام می‌داد. نتایج نشان داد در نمره‌دهی گلوبال توسط آزمونگران متخصص، پایایی بین ایستگاه، روایی سازه و روایی همزمان بهتر از ارزیابی توسط چک‌لیست است و همچنین وجود چک‌لیست موجب بهبود روایی یا پایایی نمره‌دهی گلوبال نمی‌شود.

البته به دلیل لزوم استفاده از آزمونگر متخصص و مجرب در نمره‌دهی گلوبال، نمی‌توان منکر این قضیه شد که گاهی محدودیت‌های اجرایی تعیین‌کننده هستند. به عنوان مثال، چنانچه نتوان برای ارزیابی تمام ایستگاه‌ها از هیأت علمی و آزمونگران متخصص دعوت کرد، نمی‌توان نمره‌دهی گلوبال را به راحتی انجام داد.

در هر حال، به صورت کلی نمی‌توان گفت که چک‌لیست بهتر از نمره‌دهی گلوبال است یا برعکس. هیچ کدام آنها واقعاً استاندارد طلایی نیستند (وس و همکاران[10] ۲۰۰۱). همان طور که چک‌لیست نمی‌تواند تمام جنبه‌های یک مهارت را پوشش دهد و روح کلی مهارت در بین آیتم‌های آن گم می‌شود، نمره‌دهی گلوبال نیز مستعد خطای شخصی و ذهنی افراد است.

1. Pell et al
2. Boursicot et al
3. Chunking
4. Cunnington et al
5. Regehr et al
6. Schwatz et al
7. Park et al.
8. Malau-Aduli et al.
9. Hodges & McIlroy
10. Wass et al.

یک توصیه، استفاده از رویکرد میانی و ترکیبی است. برای برخی از ایستگاه‌ها مخصوصاً آنها که مهارت تکنیکی و پروسیجر را می‌سنجند و مسیر مشخص و تعریف شده‌ای دارند، چک‌لیست مناسب‌تر است. در حالی که برای بعضی از ایستگاه‌ها که ماهیت آنها طوری است که تنها یک مسیر خط‌کشی شده در رویکرد به بیمار وجود ندارد؛ مثل مهارت ارتباطی و شاید برخی از ایستگاه‌های تشخیص، نمره‌دهی گلوبال بهتر است (نیوبل ۲۰۰۴).

همچنین راه حل دیگری برای ترکیب نمره‌دهی گلوبال و چک‌لیست وجود دارد که از آن تحت عنوان «ایستگاه بعد مواجهه»[1] نام برده می‌شود و در امتحان USMLE استفاده می‌شود: دانشجو ابتدا در مواجهه با بیمارنما وظایف محوله را انجام می‌دهد و بیمارنمای حاضر در ایستگاه او را توسط چک‌لیست ارزیابی می‌کند. سپس پنج تا هفت دقیقه وقت دارد تا به یک سری از پرسش‌ها به صورت کتبی پاسخ دهد. ممکن است لزوماً پرسشی در کار نباشد و بنا به موضوع ایستگاه از دانشجو خواسته شود خلاصه‌ای از شرح‌حالی که گرفته است، بنویسد یا با جمع‌بندی مشکلات بیمار، فهرستی از تشخیص‌های افتراقی تهیه کند یا برنامه تشخیصی–درمانی خود را ارائه دهد. این نوشته‌ها بعداً در اختیار اعضای هیأت علمی مجرب قرار می‌گیرد و تصحیح می‌شود. این رویکرد هم قابلیت اجرای خوبی دارد هم روایی و پایایی مطلوبی دارد. نتایج یک مطالعه که ارزیابی بیمارنما را با نوشته‌ها مقایسه کرده است، نشان می‌دهد نمرات و میزان قبولی ناشی از دو روش، هم در سطح ایستگاه و هم در سطح کل آزمون، مشابه بوده‌اند (ویلیامز و همکاران[2] ۱۹۹۹).

آیا استفاده از بیمار واقعی در ایستگاه نسبت به بیمارنما مزیت دارد؟

در ایستگاه‌های OSCE هنگامی که قرار است مواجهه پزشک و بیمار مورد ارزیابی قرار گیرد، از بیمار واقعی استفاده می‌شود یا فردی سالم که نقش بیمار را بازی می‌کند. هنگامی که صحبت از بیمار واقعی در برابر بیمار استانداردشده یا بیمار نما می‌شود، باید این مسأله را مدنظر قرار داد که مفاهیم فوق در قالب یک طیف تعریف می‌شوند که می‌تواند شامل موارد زیر باشد (کولینز و هاردن[3] ۱۹۹۸):

- بیمار واقعی که موافقت کرده در آزمون شرکت کند اما آموزش و تمرینی نداشته است.
- بیمار واقعی که آنچه باید در جلسه آزمون انجام دهد، به او آموزش داده شده است.
- بیمار واقعی که به او گفته شده قسمتی از شرح حال یا تاریخچه بیماری خود را در جلسه آزمون تغییر دهد.
- بیمارنمایی که به او گفته‌اند چه کاری باید در جلسه آزمون انجام دهد.
- بیمارنمایی که سناریویی به وی داده شده است تا با بیمار و سیر مشکلات او آشنا شود.
- بیمارنمایی که به وی آموزش داده شده است تا طبق سناریوی از قبل تعریف شده، نقش ایفا کند. نوع سؤال و جواب‌ها، رفتارهای کلامی و غیرکلامی طی جلسات تمرین به دقت مشخص شده است (بیمار استاندارد شده).

استفاده از بیماران واقعی، جلسه آزمون را به محیط کار طبیعی دانشجویان نزدیک می‌کند. عمده‌ترین مزیتی که با استفاده از بیماران واقعی ایجاد می‌شود، امکان طرح سؤال از بیماری‌هایی است که در OSCE قابلیت شبیه‌سازی ندارند و به همین دلیل معمولاً حذف می‌شوند. مثلاً سوفل قلبی در بیمار مبتلا به تنگی دریچه یا معاینه تیروئید در بیمار مبتلا به گواتر یا ندول در بیمار مبتلا به آرتریت روماتوئید مواردی نیستند که بتوان در یک فرد سالم ایجاد نمود. بنابراین، معمولاً طراحان OSCE از آنها می‌گذرند و به مواردی روی می‌آورند که در دسترس و قابل اجرا است.

اما در عین حال، استفاده از بیمار واقعی، مشکلاتی نیز ایجاد می‌کند به طوری که امروزه استفاده از بیمار واقعی در OSCE کم شده است (گرملی ۲۰۱۱). برخی از چالش‌های بیمار واقعی در OSCE شامل موارد زیر است:

- ممکن است بیمار واقعی در دسترس نباشد یا به تعداد مورد نیاز در دسترس نباشد. مثلاً اگر قرار است یک OSCE به

1. Post-encounter station
2. Williams and McLaughlin
3. Collins & Harden

صورت همزمان در دو لاین موازی برگزار شود، به این معناست که در آن واحد از یک مشکل، دو بیمار مورد نیاز است. از آنجا که دسترسی به بیمار کاملاً مشابه امکان‌پذیر نیست، این امر می‌تواند موجب مخدوش شدن پایایی آزمون گردد.

- استفاده از بیمار واقعی برای ارزیابی تمام توانمندی‌ها امکان‌پذیر نیست. ممکن است بیماری فوریت داشته باشد (به عنوان مثال، گاردینگ شکم بیمار با پریتونیت) یا ممکن است بیمار اذیت شود (به عنوان مثال، بیمار مبتلا به کانسر که برای مشاوره روانپزشکی مراجعه کرده است).
- آزمون OSCE معمولاً طولانی است. انجام معاینه مکرر توسط تعداد زیادی دانشجو می‌تواند برای بیمار واقعی از لحاظ جسمانی اذیت‌کننده باشد.
- احتمال دارد بیماران در برابر دانشجویان مختلفی که به ایستگاه می‌آیند، یکسان عمل نکنند. تفاوت در عملکرد بیماران، منبع خطایی است که پایایی آزمون را تحت تاثیر قرار می‌دهد زیرا می‌تواند موجب شود نمرات دانشجویان تغییراتی کند که ناشی از سطح توانمندی دانشجویان نیست. راه حلی که معمولاً در پیش گرفته می‌شود، آموزش دادن و تمرین است که در بیماران واقعی به علت سطح سواد یا سن آنها ممکن است به راحتی مقدور نباشد.
- استفاده از بیماران واقعی بر یادگیری دانشجویان تاثیر می‌گذارد که همیشه در جهت مثبت نیست. در یک مطالعه، دانشجویان پیش‌بینی کرده بودند که مواردی مانند فیبروز ریوی و سوفل قلبی به احتمال زیاد در OSCE مطرح می‌شوند زیرا استاندارد کردن آنها برای استفاده در لاین‌های موازی راحت‌تر از بیماری‌های دیگر است (گرملی و همکاران ۲۰۱۱). این مسأله موجب می‌شود دانشجویان علاقه‌ای نداشته باشند به یادگیری مواردی بپردازند که احتمال سؤال از آنها در OSCE کمتر است.

به این ترتیب با وجود اینکه احتمالاً رویارویی دانشجو با بیمار واقعی، روایی بهتری برای آزمون ایجاد می‌کند (اسمی[1] ۲۰۰۳)، استفاده از بیمارنمایی که آموزش دیده و فرایند استانداردسازی را پشت سر گذاشته است، علی‌رغم انرژی و هزینه‌ای که صرف آن میشود، مزایای خودش را دارد. در برخی از موارد، استفاده از بیمار واقعی برای برخی از ایستگاه‌ها و استفاده از بیمارنما برای ایستگاه‌های دیگر امکان ارزیابی مناسب‌تری را فراهم می‌کند

به طور کلی، به منظور تصمیم‌گیری برای استفاده از بیمار واقعی یا بیمارنما در OSCE باید به مسائل زیر توجه کرد (کولینز و هاردن ۱۹۹۸):

- چه چیزی قرار است مورد سنجش قرار گیرد؟ شرح حال، معاینه، مهارت ارتباطی؟ آیا لازم است که تعامل پزشک و بیمار زیاد باشد؟ آیا یافته‌های خاصی در معاینه مدنظر است؟ اگر میزان تعامل بیشتر باشد، استفاده از بیمارنما بهتر است. اگر یافته بالینی خاصی مدنظر است، استفاده از بیمار واقعی کمک‌کننده است.
- چه میزان یکسان سازی و استاندارد کردن لازم است؟ در آزمون‌های مهم، تراکمی و سرنوشت‌ساز که نمره و نتیجه دانشجویان ملاک تصمیم مهمی خواهد بود، مهم است که شرایط آزمون برای همه یکسان باشد. بنابراین ضرورت استفاده از بیمارنما بیشتر احساس می‌شود.
- قابلیت اجرا، دسترسی و هزینه استفاده از بیمار واقعی و بیمارنما چقدر است؟
- چقدر لازم است که آزمون واقعی به نظر برسد و مشابه شرایط روزمره طبابت باشد؟

آیا می‌توان برای نمره‌دهی به دانشجو از بیمارنما استفاده کرد؟

در برخی از موارد برای ارزیابی عملکرد دانشجو در ایستگاه، نمی‌توان به تعداد مناسب آزمونگری که پزشک و هیأت علمی باشد، پیدا کرد. سؤال این است که آیا در این موارد می‌توان از افراد غیرپزشک استفاده کرد و مثلاً از بیمارنمای حاضر در ایستگاه برای ارزیابی دانشجو استفاده نمود یا خیر.

1. Smee

حقیقت این است که استفاده از بیمارنما در OSCE تنها به ایفای نقش بیمار محدود نمی‌شود. در برخی از مواقع از بیمار حاضر در ایستگاه خواسته می‌شود تا علاوه بر ایفای نقش بیمار، عملکرد دانشجو را نیز ارزیابی کند (نستل و نیبون ۲۰۱۰). این ارزیابی بیشتر بر جنبه‌های انسانی و ارتباطی مواجهه بالینی متمرکز است و علاوه بر این که بر اهمیت بیمار–محوری در رابطه پزشک و بیمار تاکید می‌کند، درگیری بیمارنماها را در روند برگزاری OSCE افزایش می‌دهد. در عین حال، این نگرانی وجود دارد که آیا ارزیابی بیمارنما از دانشجو درست و دقیق است یا خیر؟

نتایج برخی از مطالعات نشان می‌دهند نمراتی که بیمارنماها به دانشجویان داده‌اند، رابطه خوبی با نمرات پزشکان داشته است (مَن و همکاران[1] ۱۹۹۰، کوهن و همکاران[2] ۱۹۹۰). یک مطالعه در آلمان روی ۲۱۴ دانشجوی پزشکی نشان داد که این کار، بدون آن که تاثیر نامطلوبی بر پایایی آزمون بگذارد کاهش هزینه‌ها را به دنبال دارد که پیامد مطلوبی است (شینوت و همکاران[3] ۲۰۰۷). همچنین دیده شده است که استفاده از نظرات بیمارنماها در کنار نمراتی که آزمونگران بر اساس چک‌لیست به دانشجو داده‌اند، پایایی ارزیابی را افزایش می‌دهد (هومر و پِل[4] ۲۰۰۹). همچنین بر اساس نتایج یک مطالعه در مورد نمره‌دهی گلوبال و چک‌لیست، دیده شد که توافق بین نمرات بیمارنمایان و آزمونگران در صوررتی خوب است که ارزیابی بر اساس چک‌لیست صورت گیرد و اگر ارزیابی به شیوه گلوبال انجام شود، توافق خوبی به دست نمی‌آید (هامفری–مورتو و همکاران[5] ۲۰۰۵). نتایج مطالعه دیگری نشان داد که پایایی هر دو روش سنجش (گلوبال و چک‌لیست) که توسط بیمارنمایان صورت گرفته است مطلوب بوده است (کیلمینستر و همکاران[6] ۲۰۰۷). همچنین، به نظر می‌رسد با آموزش آزمونگران–چه بیمارنما باشند و چه پزشک– پایایی خوبی به دست می‌آید. در یک مطالعه که آزمونگران مختلف شامل پزشک، دانشجوی پزشکی و افراد غیرپزشک عملکرد دانشجویان را در فیلم‌های ضبط شده ارزیابی کردند، تفاوتی که بین عملکرد آزمونگران گروه‌های مختلف قبل از آموزش وجود داشت، با آموزش از بین رفت (ون‌درولوتن و همکاران ۱۹۸۹).

در هر حال، توجه به چند نکته برای استفاده از بیمارنما به عنوان آزمونگر حائز اهمیت است:

- سنجشی که بیمارنما انجام می‌دهد، مراتب دارد. به طوری که می‌تواند صرفاً به صورت ارزیابی تکوینی و ارائه بازخورد به دانشجو باشد یا اینکه بیمارنما در قالب ارزیابی تراکمی نمره‌ای برای دانشجو در نظر بگیرد که در سرنوشت او موثر باشد.
- سنجشی که توسط بیمارنما انجام می‌شود، دقیقاً جایگزین همان ارزیابی نیست که توسط استاد صورت می‌گیرد. به عبارت دیگر، جنبه‌های مورد ارزیابی توسط این دو گروه یکسان نیست. در حالی که بیمارنما می‌تواند قضاوت خوبی در مورد مهارت برقراری ارتباط دانشجو داشته باشد، به منظور ارزیابی جنبه‌های بالینی و مهارت‌های تکنیکی و علمی باید از پزشکان استفاده کرد (ون‌درولوتن و سوانسون ۱۹۹۰). در USMLE معمولاً بیمارنمایی که در ایستگاه حضور دارد، به عنوان آزمونگر نیز محسوب می‌شود (بورسیکوت و همکاران ۲۰۱۱) اما حیطه‌ای که ارزیابی می‌کند، مهارت ارتباطی و مهارت زبان است که مستقل از نمره بالینی دانشجو و به عنوان معیارهای جداگانه مورد استفاده قرار می‌گیرند.
- برای ارزیابی توسط بیمارنما نمی‌توان از همان چک‌لیست‌هایی استفاده کرد که برای استادان طراحی شده است. بدین منظور باید چک‌لیست‌های مخصوص بیمارنما تدوین شود. به عنوان مثال، در دانشگاه کوییز بلفاست از بیمارنما خواسته می‌شود که بعد از اتمام کار دانشجو نظر خودش را در مورد این جمله اعلام کند: «من مایلم مجدداً به این

1. Mann et al.
2. Cohen et al.
3. Chenot et al.
4. Homer & Pell
5. Humphrey-Murto et al.
6. Kilminster et al.

دانشجو مراجعه کنم و مشکلاتم را با او در میان بگذارم» (گرملی ۲۰۱۱).

- برای اینکه بیمارنما کار ارزیابی را نیز انجام دهد، نیاز است تا علاوه بر جلسات و کارگاه‌های آموزشی ایفای نقش که به صورت معمول اجرا می‌شود، آموزش‌های اختصاصی ارزیابی نیز دریافت کند. در کارگاه‌هایی که برای آموزش بیمارنمایان ترتیب داده می‌شود، باید آموزش نحوه ثبت مشاهدات و نمره‌دهی بر اساس چک‌لیست نیز گنجانده شود. اگر قرار است بیمارنما به دانشجو بازخورد شفاهی دهد، آموزش نحوه دادن بازخورد نیز باید صورت گیرد. چنانچه نظر بیمارنما در تصمیم‌گیری سرنوشت‌ساز برای وضعیت رد و قبول دانشجو دخالت داده می‌شود، آموزش و تمرین فراوان بیمارنماها مورد نیاز می‌باشد (کللند و همکاران[1] ۲۰۰۹).

1. Cleland et al.

منابع

1. Abraham RR, Raghavendra R, Surekha K, Asha K. A trial of the objective structured practical examination in physiology at Melaka Manipal Medical College, India. Adv Physiol Educ 2009;33(1):21-3.
2. Amini M, Moghadami M, Kojuri J, et al. Using TOSCE (Team Objective Structured Clinical Examination) in the second national medical sciences Olympiad in Iran. J Res Med Sci 2012; 17(10): 975–978.
3. Baribeau DA, Mukovozov I, Sabljic T, Eva KW, deLottinville CB. Using an objective structured video exam to identify differential understanding of aspects of communication skills. Med Teach 2012;34(4):e242-50.
4. Barman A. Critiques on the Objective Structured Clinical Examination. Ann Acad Med Singapore 2005;34(8):478-82
5. Bartfay W, Rombough R, Howse E, Leblanc R. The OSCE approach in nursing education. Canadian Nurse 2004;100(3):18–23.
6. Barzansky B, Jonas HS, Etzel SI. Educational programs in US medical schools, 1997-1998. JAMA 1998;280(9):803-8, 827-35.
7. Barzansky B, Etzel SI. Medical Schools in the United States, 2009-2010. JAMA 2010;304(11):1247-1254.
8. Biran LA. Self-assessment and learning through GOSCE group objective structured examination. Medical Education 1991; 25(6):475–479.
9. Boillat M, Bethune C, Ohle E, Razack S, Steinert Y. Twelve tips for using the objective structured teaching exercise for faculty development. Med Teach 2012;34(4):269-73.
10. Boursicot K, Etheridge L, Setna Z, Sturrock A, Ker J, Smee S, et al. Performance in assessment: consensus statement and recommendations from the Ottawa conference. Med Teach 2011; 33(5): 370-83
11. Bujack L, McMillan M, Dwyer J, Hazelton M. Assessing comprehensive nursing performance: the Objective Structural Clinical Assessment (OSCA). Part 1--Development of the assessment strategy. Nurse Educ Today 1991;11(3):179-84.
12. Carraccio C, Englander R. The Objective Structured Clinical Examination: A Step in the Direction of Competency-Based Evaluation. Arch Pediatr Adolesc Med 2000;154(7):736-741.
13. Casey PM, Goepfert AR, Espey EL, et al. To the point: reviews in medical education--the Objective Structured Clinical Examination. Am J Obstet Gynecol 2009;200(1):25-34.
14. Chipman JG, Schmitz CC. Using objective structured assessment of technical skills to evaluate a basic skills simulation curriculum for first-year surgical residents. J Am Coll Surg. 2009;209(3):364-370.

15. Cleland JA, Abe K, Rethans J. The use of simulated patients in medical education: AMEE Guide No 42. Medical Teacher 2009,31:6,477-486
16. Collins JP, Harden RM. AMEE Medical Education guide No. 13: real patients, simulated patients and simulations in clinical examinations. Med Teach 1998; 20(6): 508-21.
17. Cox K. No Oscar for OSCA. Medical Education 1990; 24(6):540–545.
18. Cunnington JPW, Neville AJ, Norman GR. The Risks of Thoroughness: Reliability and Validity of Global Ratings and Checklists in an OSCE. In Advances in Medical Education. Springer Netherlands 1997;143-145
19. Dissanayake AS, Ali BA, Nayar U. The influence of the introduction of objective structured practical examinations in physiology on student performance at King Faisal University Medical School. Medical Teacher 1990 12(3–4):297–304.
20. Eftekhar H, Labaf A, Anvari P, Jamali A, Sheybaee-Moghaddam F. Association of the pre-internship objective structured clinical examination in final year medical students with comprehensive written examinations. Med Educ Online 2012;17
21. Elliot DL, Fields SA, Keenen TL, Jaffe AC, Toffler WL. Use of a group objective structured clinical examination with first-year medical students. Acad Med 1994;69(12):990-2.
22. Epstein RM. Assessment in medical education. N Engl J Med 2007;356(4):387-96.
23. GMC 2009. General Medical Council. Tomorrow's doctors: outcomes and standards for undergraduate medical education. 2nd ed. London: General Medical Council; 2009.
24. GMC 2009 supplementary. General Medical Council. Assessment in undergraduate medical education: advice supplementary to Tomorrow's Doctors. London: General Medical Council; 2009. Available online from: http://www.gmc-uk.org/Assessment_in_undergraduate_web.pdf_38514111.pdf
25. Gormley GJ, McCusker D, Booley MA, McNeice A. The use of real patients in OSCEs: a survey of medical students' predictions and opinions. Med Teach 2011;33(8):684.
26. Harden RM. How to Assess Clinical Competence – An overview. Med Teach 1979;1:289–296.
27. Harden RM, Cairncross RG. 1980. Assessment of Practical Skills: The Objective Structured Practical Examination (OSPE). Studies High Educ 5:187–196.
28. Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). Med Educ 1979;1)13): 41-54.
29. Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of Clinical Competence using Objective Structured Examination. BMJ 1975; 1:447–451.
30. Hodges B, McIlroy JH. Analytic global OSCE ratings are sensitive to level of training. Med Educ 2003;37(11):1012-6.
31. Humphris GM, Kaney S. The objective structured video exam for assessment of communication skills. Medical Education 2000; 34(11):939–945.

32. Julian K, Appelle N, O›Sullivan P, Morrison EH, Wamsley M. The impact of an objective structured teaching evaluation on faculty teaching skills. Teach Learn Med 2012;24(1):3-7.
33. Khan KZ, Ramachandran S, Gaunt K, Pushkar P. The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part I: An historical and theoretical perspective. Med Teach 2013;35(9):e1437-46. doi:10.3109/0142159X.2013.818634.
34. Khan KZ, Gaunt K, Ramachandran S, Pushkar P. The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part II: Organisation & Administration.Med Teach 2013;35(9):e1447-63.
35. Kilminster S, Roberts T, Morris P. Incorporating patients› assessments into objective structured clinical examinations. Educ Health 2007;20(1):6
36. Kundu D, Das HN, Sen G, Osta M, Mandal T, Gautam D. Objective structured practical examination in biochemistry: An experience in Medical College, Kolkata. J Nat Sci Biol Med 2013; 4(1): 103–107.
37. Lin CW, Clinciu DL, Swartz MH, et al. An integrative OSCE methodology for enhancing the traditional OSCE program at Taipei medical university ospital - a feasibility study. BMC Med Educ 2013;13(1):102.
38. Malau-Aduli BS, Mulcahy S, Warnecke E, et al. Inter-rater reliability: comparison of checklist and global scoring for OSCEs, Creative Education 2012;3(6A):937-942
39. Martin JA, Regehr G, Reznick R, MacRae H, Murnaghan J, Hutchison C, Brown M. Objective structured assessment of technical skill (OSATS) for surgical residents. Br J Surg 1997;84(2):273-8.
40. Nestel D, Kneebone R, Nolan C, Akhtar K, Darzi A. Formative Assessment of Procedural Skills: Students' Responses to the Objective Structured Clinical Examination and the Integrated Performance Procedural Instrument. Assess Eval Higher Educ 2011;36:171–183.
41. Newble D. Techniques for measuring clinical competence: objective structured clinical examinations. Med Educ 2004;38(2):199-203.
42. Niitsu H, Hirabayashi N, Yoshimitsu M, et al. Using the Objective Structured Assessment of Technical Skills (OSATS) global rating scale to evaluate the skills of surgical trainees in the operating room. Surg Today 2013;43(3):271-5.
43. Norman G, Bordage G, Page G, Keane D. How specific is case specifity? Med Educ 2006: 40: 618–623.
44. Patricio M, Juliao M, Fareleira F, Young M, Norman G, Vaz Carneiro A. A comprehensive checklist for reporting the use of OSCEs. Med Teach 2009;31(2):112-24.
45. Patrício MF, Julião M, Fareleira F, Carneiro AV. Is the OSCE a feasible tool to assess competencies in undergraduate medical education? Med Teach 2013;35(6):503-14
46. Pell G, Fuller R, Homer M, Roberts T. How to measure the quality of the OSCE: A review of metrics – AMEE guide no.49. Med Teach 2010; 32(10): 802-11

47. Regehr G, MacRae H, Reznick R, Szaky D. Comparing the psychometric properties of check-lists and global rating scale for assessing performance on an OSCEformat examination. Acad Med 1998;73:993–7.
48. Reznick R, Smee S, Rothman A, et al. An objective structured clinical examination for the licentiate: report of the pilot project of the Medical Council of Canada. Acad Med 1992;67(8):487-94
49. Rothman AI, Blackmore D, Dauphinee WD, Reznick R. The use of global ratings in OSCE station scores. Adv Health Sci Educ 1997;1:215
50. Schoonheim-Klein M, Muijtjens A, Habets L, et al. On the reliability of a dental OSCE, using SEM: effect of different days. Eur J Dent Educ 2008;12(3):131-7.
51. Schwatz MH, Colliver JA, Bardes CL, Charon R, Fried ED, Moroff S. Global ratings of videotaped performance versus global ratings of actions recorded on checklists: a criterion for performance assessment with standardized patients. Acad Med 1999; 74: 1028–32
52. Serpell JW. Evolution of the OSCA-OSCE-Clinical Examination of the Royal Australasian College of Surgeons. ANZ J Surg 2009;79(3):161-8.
53. Singleton A, Smith F, Harris T, Ross-Harper R, Hilton S. An evaluation of the Team Objective Structured Clinical Examination (TOSCE). Medical Education 1999 33(1):34–41.
54. Smee S. ABC of learning and teaching in medicine: Skill based assessment. BMJ 2003;326:703-6
55. Stone S, Mazor K, Devaney-O›Neil S, Starr S, Ferguson W, Wellman S, Jacobson E, Hatem DS, Quirk M Development and implementation of an objective structured teaching exercise (OSTE) to evaluate improvement in feedback skills following a faculty development workshop. Teach Learn Med 2003 Winter;15(1):7-13.
56. Swanson D, Norcini J. Factors influencing the reproducibility of tests using standardized patients. Teaching and Learning in Medicine 1989;l:158-66.
57. Swanson DB, Norman GR, Linn RL. Performance-based assessment: lessons from the health professions. Educ Res 1995: 24: 5–11.
58. Symonds I, Cullen L, Fraser D. Evaluation of a formative interprofessional team objective structured clinical examination (ITOSCE): a method of shared learning in maternity education. Med Teach 2003;25(1):38-41.
59. Taghva A, Leili Panaghi L, Rasoulian M, Bolhari J, Zarghami M, Nasr Esfahani M, Evaluation of Reliability and Validity of the Psychiatry OSCE in Iran. Academic Psychiatry 2010;34(2):154-157
60. Turner JL, Dankoski ME. Objective structured clinical exams: a critical review. Fam Med 2008;40(8):574-8.
61. Van der Vleuten CPM. Validity of final examinations in undergraduate medical training. BMJ 2000;321:1217
62. Van der Vleuten CP, Schuwirth LW. Assessing professional competence: from methods to programmes. Med Educ 2005;39(3):309-17

63. Van der Vleuten CPM, Swanson DB. Assessment of clinical skills with standardized patients: State of the art. Teaching and Learning in Medicine 1990;2(2):58-76
64. Van der Vleuten C, van Luyk S, Ballegooijen A, Swanson D. Training and experience of medical examiners. Medical Education 1989;23:290-6.
65. Van Hove PD, Tuijthof GJ, Verdaasdonk EG, Stassen LP, Dankelman J. Objective assessment of technical surgical skills. Br J Surg 2010;97(7):972-87
66. Vaseghi N, Alizadeh Naini M, Labaf Ghasemi R, Amiri S. Validity and Reliability of preinternship Objective Structured Clinical Examination (OSCE) in Shiraz Medical School. Adv Med & Prof 2013;1(3):85-88.
67. Wass V, Van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. Lancet 2001;357(9260):945-9.
68. Williams RG, McLaughlin MA, Eulenberg B, Hurm M, Nendaz MR. The patient findings questionnaire: one solution to an important standardized patient examination problem. Acad Med 1999; 74: 1118–24.

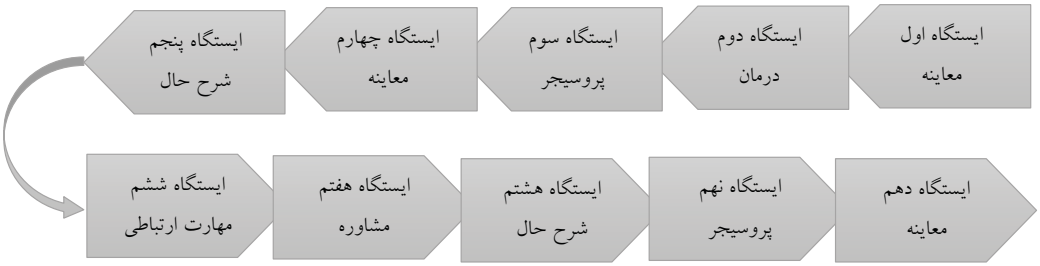
فصل ۱۸

آزمون بالینی ساختارمند عینی

ساختار OSCE

همان‌طور که در تاریخچه عنوان شد، از زمان معرفی OSCE تاکنون تغییرات زیادی در نحوه طراحی و اجرای این آزمون پیشنهاد شده است و نویسندگان مختلف، تعاریف گوناگونی از این آزمون ارائه داده‌اند. تعریفی که به عنوان جمع‌بندی در AMEE guide شماره ۸۱ در سال ۲۰۱۳ منتشر شده است، به این صورت است: «ابزار ارزیابی که بر پایه اصول عینیت و استاندارد کردن استوار است و در آن دانشجویان در ایستگاه‌هایی با زمان مشخص حرکت می‌کنند تا عملکرد حرفه‌ای آنها در یک محیط شبیه‌سازی شده توسط آزمونگران آموزش دیده و طبق روبریک‌های نمره‌دهی استاندارد مورد ارزیابی قرار گیرد» (خان و همکاران ۲۰۱۳).

بنابراین، ساختار کلی OSCE به این صورت است که هر دانشجو به صورت متوالی چندین ایستگاه را پشت سر می‌گذارد و در هر یک از آنها طی مدت زمان واحدی، یک وظیفه مشخص بالینی را انجام می‌دهد. با شنیدن صدای زنگ، دانشجو ایستگاه را ترک می‌کند و به ایستگاه بعدی می‌رود و دانشجوی دیگری وارد ایستگاه می‌شود. معمولاً OSCE به صورت چرخشی برگزار می‌شود مثلاً در یک OSCE ده ایستگاهی، ده دانشجو همزمان وارد ایستگاه‌ها می‌شوند و یک توالی مشخص را طی می‌کنند. به عنوان مثال، آخرین ایستگاه برای دانشجویی که از ایستگاه سوم وارد شده، ایستگاه دوم خواهد بود (شکل ۱–۱۸).



شکل ۱–۱۸: ساختار OSCE چرخشی با ده ایستگاه

ساختار و اجزای هر ایستگاه وابسته به مهارتی است که قرار است در آن ارزیابی شود. معمولاً این طور است که عملکرد دانشجو در مواجهه با بیمار مورد ارزیابی قرار می‌گیرد. بنابراین، یکی از اجزای ایستگاه «بیمار» است که ممکن است واقعی

باشد یا فردی معمولی که نقش بیمار را بازی می‌کند و به آن بیمارنما می‌گویند. به همین ترتیب، معمولاً یک «مشاهده‌گر»[1] یا «آزمونگر»[2] در هر ایستگاه وجود دارد که بر اساس «چک‌لیستی» که قبلاً تدوین شده است، عملکرد دانشجو را ارزیابی و نمره‌دهی می‌کند. همچنین «امکانات» مورد نیاز در ایستگاه نیز باید فراهم باشد.

اجزای مختلف OSCE و هدف از آنها در جدول ۱–۱۸ خلاصه شده‌اند:

جدول ۱–۱۸: اجزای مختلف OSCE و هدف آنها

اجزا	اهداف
سناریوی بالینی و تعامل با بیمارنما	روند تفکر و نحوه عملکرد دانشجو را در حین مواجهه با بیمار نشان می‌دهد.
چک‌لیست ساختارمند عینی	اساس ارزیابی در ایستگاه است و پایایی بین آزمونگران را بهبود می‌بخشد.
مشاهده مستقیم توسط آزمونگر یا ضبط جلسه	ارزیابی و بازخورد را تسهیل می‌کند.
ایستگاه‌های متعدد	تعداد و تنوع توانمندی‌های مورد سنجش را افزایش می‌دهد.
ارزیابی صلاحیت بالینی در حیطه‌های مختلف	علاوه بر ارزیابی نحوه گرفتن شرح حال، نحوه برقراری ارتباط، نحوه انجام معاینه و پروسیجر، از طریق مکتوب‌سازی شرح‌حال، یافته‌های بالینی و برنامه بیمار می‌توان توانمندی دانشجو را در جمع‌بندی مشکلات بیمار، استدلال و ارائه تشخیص‌های افتراقی ارزیابی نمود.

موارد زیر از جمله تغییرات ساختاری هستند که در همه OSCE ها دیده نمی‌شوند اما چون کم و بیش مورد استفاده قرار می‌گیرند، در این فصل به آنها اشاره خواهد شد:

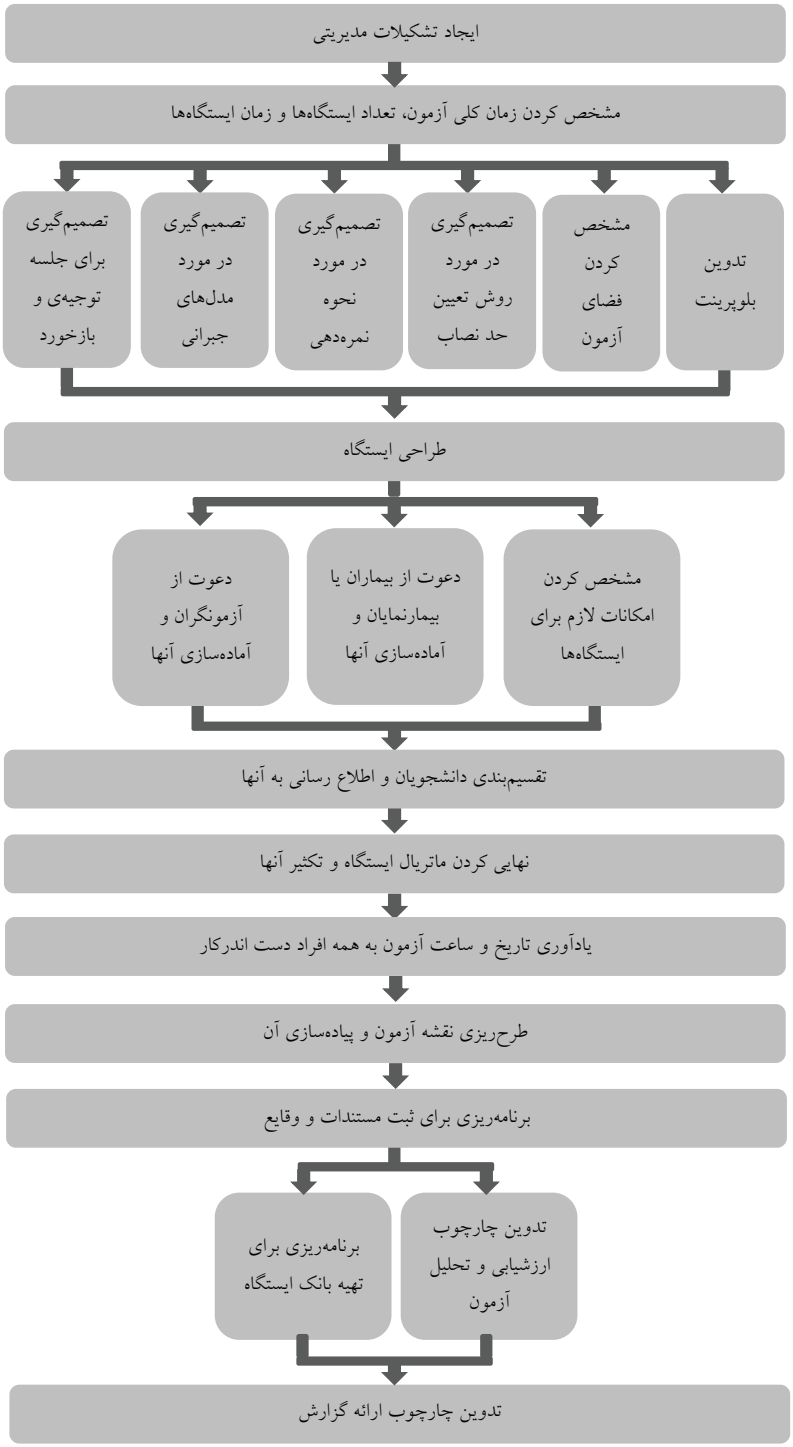
- ضبط عملکرد دانشجویان در ایستگاه‌ها
- برگزاری جلسه‌ای برای توجیه و بازخورد
- برگزاری OSCE در لاین‌های موازی
- برگزاری OSCE متوالی[3]

گام‌های طراحی OSCE

دستورالعمل‌ها و مقالات مختلف، مراحل گوناگونی برای طراحی OSCE ذکر کرده‌اند که بعضاً با یکدیگر همپوشانی‌هایی دارند و البته برخی هم به مطالب جدیدی اشاره می‌کنند (سلبی و همکاران[4] ۱۹۹۵، اسمی ۲۰۰۳، خان و همکاران ۲۰۱۳). در اینجا مراحل طراحی OSCE بر اساس این متون، جمع‌بندی و ارائه می‌شود.

باید دقت کرد که برخی از این مراحل حتماً باید به صورت متوالی انجام شوند ولی برخی دیگر به صورت موازی هم قابل انجام هستند. این موضوع از آن جهت حائز اهمیت است که با همزمان پیش بردن چند مرحله می‌توان زمان طراحی آزمون را به نحو بهتر و موثرتری مدیریت کرد. در عین حال، قبل از آغاز این راه دشوار، لازم است اطمینان حاصل شود که حمایت و پشتیبانی لازم برای برگزاری آزمون صورت می‌گیرد. برگزاری OSCE کاری پرهزینه است و نیاز به امکانات و منابع مالی در خور توجه‌ای دارد. شکل ۲–۱۸ خلاصه مراحل طراحی OSCE را نشان می‌دهد.

1. Observer
2. Rater, assessor, examiner, coder, marker
3. Sequential OSCE
4. Selby et al.



شکل ۲-۱۸: خلاصه مراحل طراحی OSCE

موضوع دیگر، استفاده از تکنولوژی در طراحی و برگزاری OSCE است. با توجه به حجم زیاد کاغذی که در OSCE مصرف می‌شود، نرم‌افزارهایی طراحی شده‌اند که به برگزارکنندکان OSCE در مراحل مختلف آن کمک می‌کنند. از جمله آنها می‌توان OSCErunner[1] و qpercom[2] را نام برد. به عنوان مثال، برگه‌ها و چک‌لیست‌های طراحی شده، به جای اینکه پرینت و تکثیر شوند، در نرم‌افزار قرار می‌گیرند. آزمونگر در ایستگاه، کامپیوتر یا تبلتی در اختیار دارد که ارزیابی دانشجو را در آن انجام می‌دهد. در نهایت به جای اینکه وقت زیادی صرف جمع آوری برگه‌ها، وارد کردن نمرات و تصحیح و تحلیل آنها شود، خود نرم افزار در مدت کوتاهی این کارها را انجام می‌دهد. با وجود اینکه استفاده از این نرم افزارها برگزاری آزمون را تسهیل می‌کند، در مراحل طراحی که ذکر خواهد شد، تغییر چندانی ایجاد نمی‌کند.

ایجاد تشکیلات مدیریتی

این مرحله مخصوصاً برای دانشکده‌ای اهمیت دارد که تصمیم گرفته است برای اولین بار OSCE برگزار کند. از آنجا که تعداد زیادی از افراد اعم از پرسنل و اعضای هیأت علمی درگیر برگزاری آزمون OSCE می‌شوند، مدیریت و هماهنگی امور مختلف اهمیت فراوانی دارد. در روند برگزاری OSCE موقعیت‌های بسیاری پیش می‌آید که نیازمند تصمیم‌گیری در مورد مسائل مختلف از جنبه‌های گوناگون است که احتمالاً به تنهایی از عهده یک نفر خارج است. بنابراین در اغلب موارد کمیته OSCE تشکیل می‌شود که در واقع جمعی از دست‌اندکاران، صاحب نظران و مسؤولان به تصمیم‌گیری امور می‌پردازند.

کمیته OSCE لزوماً یک نهاد مستقل نیست و حتی بهتر است در صورت وجود کمیته ارزیابی فراگیران، که مسؤولیت آزمون‌های دانشجویان را بر عهده دارد، زیر نظر آن فعالیت کند. تصمیماتی مانند اینکه آزمون برای دانشجویان سال چندم برگزار می‌شود، شرایط شرکت در آن چیست، یا عواقب عدم شرکت یا عدم قبولی در آن چه می‌تواند باشد، از جمله مسائلی است که در سطح کلان‌تر و در تعامل با کمیته ارزیابی فراگیران باید مورد بررسی قرار گیرند.

بهتر است یک نفر که واجد دانش و تجربه کافی در این زمینه است، برای مدیریت کمیته OSCE در نظر گرفته شود که مسؤولیت اصلی طراحی و اجرای آزمون را بر عهده بگیرد. چنانچه OSCE برای تعداد زیادی دانشجو و به صورت همزمان در چند مرکز برگزار می‌شود شاید لازم باشد رابطی برای هر یک از مراکز تعیین گردد که کارها با هماهنگی او پیش رود.

ضروری است که مسؤول OSCE خود را متعهد به مطالعه مقالات و دستورالعمل‌های منتشر شده بداند، در کارگاه‌های مرتبط شرکت کند و برای بازدید از تجربه سایر مراکز هماهنگی‌های لازم را به عمل آورد. همچنین مسؤول OSCE به منظور ایجاد هماهنگی بین قسمت‌های مختلف باید توانمندی رهبری قوی و همچنین مهارت‌های ارتباطی خوب داشته باشد و به امر آموزش دانشجویان علاقه‌مند باشد. مسؤول برگزاری OSCE باید جدول زمانی دقیقی داشته باشد که مراحل مختلف کار را در آن زمان‌بندی کند. مخصوصاً اگر آزمون برای بار اول برگزار می‌شود، مهم است که طراحی با فاصله زمانی معقولی نسبت به تاریخ آزمون و با در نظر گرفتن جمیع جوانب آغاز شود.

مشخص کردن زمان کلی آزمون، تعداد ایستگاه‌ها و زمان ایستگاه‌ها

هرچند که به نظر می‌رسد اولین کار کمیته باید تدوین بلوپرینت آزمون باشد، اما از لحاظ اجرایی معمولاً کار کردن روی بلوپرینت تا بعد از مشخص شدن تعداد و زمان ایستگاه‌ها به تعویق می‌افتد. بنابراین، یکی از اولین کارهای کمیته OSCE این است که در مورد تعداد و زمان ایستگاه‌ها تصمیم‌گیری کند. به طور ایده‌آل خوب است که هم تعداد و هم زمان ایستگاه‌ها زیاد باشند اما از لحاظ اجرایی این امر امکان‌پذیر نیست. در مورد تعداد ایستگاه‌های مورد نیاز و زمان آنها در قسمت سؤالات رایج در فصل قبلی صحبت شد.

1. www.oscerunner.com
2. www.qpercom.ie

در این مرحله باید مشخص شود که اجرای یک دور آزمون چقدر طول می‌کشد. یعنی از هنگامی که یک دانشجو وارد ایستگاه اول می‌شود چه زمانی طول می‌کشد تا از ایستگاه آخر خارج شود. بدیهی است که این موضوع تحت تاثیر «تعداد ایستگاه‌ها» و «زمان هر ایستگاه» است. به عنوان مثال، اگر OSCE از ۱۰ ایستگاه پنج دقیقه‌ای تشکیل شده باشد، زمان آزمون برای یک دانشجو ۵۰ دقیقه خواهد بود. سپس باید پیش‌بینی شود زمان لازم به منظور برگزاری آزمون برای کل دانشجویان چقدر خواهد بود. زمان کل آزمون بر اساس «تعداد دانشجویان شرکت‌کننده» افزایش می‌یابد زیرا تعداد دورها باید افزایش پیدا کند.

برای محاسبه زمان باید توجه شود که دو مدل برای حرکت دانشجویان می‌توان متصور شد. یک حالت اینکه هر دانشجو از ایستگاه یک شروع کند و تا ایستگاه آخر ادامه دهد و سپس دانشجوی دوم وارد شود. در این حالت توالی برای تمام دانشجویان یکسان خواهد بود. بنابراین اگر کلا ۲۰ دانشجو در امتحان شرکت کنند امتحان ۱۴۵ دقیقه طول می‌کشد. به عبارت دیگر، نفر بیستم، نوزده ضرب در ۵ دقیقه بعد از نفر اول امتحان را تمام می‌کند. حالت دیگر که معمول‌تر است، حرکت چرخشی است. یعنی لازم نیست دانشجوی دوم صبر کند تا دانشجوی اول تمام ایستگاه‌ها را پشت سر بگذارد و سپس وارد ایستگاه اول شود. بلکه به محض اینکه دانشجوی اول از ایستگاه اول خارج شود و به ایستگاه دوم برود، دانشجوی دوم وارد ایستگاه اول می‌شود. در این حالت ممکن است یک دانشجو با ایستگاه شماره ۱۰ شروع و با ایستگاه شماره ۱ امتحان را تمام کند. نحوه حرکت بر زمان آزمون اثر دارد و در حالت چرخشی به اندازه تعداد کل ایستگاه‌ها منهای یک، ضرب در زمان هر ایستگاه از کل مدت آزمون کاسته می‌شود. به این ترتیب در همان مثال قبلی، درست است که یک دور آزمون برای یک دانشجو ۵۰ دقیقه طول می‌کشد، ولی در همان ۵۰ دقیقه آزمون ۱۰ نفر تمام می‌شود. پس برگزاری OSCE برای ۲۰ نفر به معنای اجرای دو دور است که هر کدام ۵۰ دقیقه به طول می‌انجامد. پس برای کل OSCE برای ۲۰ دانشجو به صد دقیقه زمان نیاز است و ۴۵ دقیقه نسبت به حالت خطی صرفه جویی می‌شود.

گاهی علاوه بر ایستگاه‌های معمول که دانشجو باید در آنها کاری انجام دهد، یک ایستگاه استراحت در نظر گرفته می‌شود که در آن دانشجو کاری انجام نمی‌دهد و صرفاً منتظر اعلام زمان می‌ماند تا به ایستگاه بعد برود. ایستگاه استراحت، علاوه بر اینکه به دانشجویان فرصت استراحت در بین آزمون می‌دهد، فایده دیگری نیز دارد. تصور کنید در آزمون فوق، ۲۲ دانشجو شرکت کنند. از آنجا که برگزاری آزمون برای هر ده نفر ۵۰ دقیقه طول می‌کشد، سه دور آزمون باید برگزار شود که دو دور آن با شرکت ۱۰ دانشجو است. درست است که در دور آخر تنها ۲ نفر شرکت می‌کنند اما در هر حال باید آزمون به صورت کامل اجرا شود تا این دو دانشجو تمام ایستگاه‌ها را بگذرانند. پس به صورت کلی، سه راند ۵۰ دقیقه‌ای یعنی ۱۵۰ دقیقه وقت برای اجرای آزمون لازم است. در حالی که اگر علاوه بر ۱۰ ایستگاه، یک ایستگاه استراحت نیز در نظر گرفته شود و آزمون ۱۱ ایستگاهه شود، اگرچه هر راند ۵۵ دقیقه طول می‌کشد، اما هر بار ۱۱ دانشجو در آزمون شرکت می‌کنند و بنابراین کلاً دو راند یعنی ۱۱۰ دقیقه زمان لازم است.

محاسبه زمان کلی آزمون از آن نظر حائز اهمیت است که در اجرا به عنوان یک عامل محدودکننده مشکلاتی ایجاد می‌کند. در حالتی که تعداد دانشجویان بسیار زیاد و زمان کل محاسبه شده طولانی است، با توجه به اینکه نمی‌توان OSCE را به این شکل برگزار کرد، کمیته آزمون باید برای حل این معضل تصمیم‌گیری کند. راه‌کارهای موجود یکی از این چند حالت است:

- برگزاری آزمون در دو نوبت صبح و عصر
- برگزاری آزمون در دو (یا چند) روز متوالی
- برگزاری آزمون به صورت همزمان در مکان‌های مختلف؛ به عبارت دیگر اجرای چند لاین موازی

با توجه به اهمیت زمان کلی آزمون به عنوان یک فاکتور محدود کننده، باید به این موضوع نیز اشاره کرد که اگر زمان کلی آزمون را ثابت فرض کنیم، در حالتی که تعداد زیادی ایستگاه کوتاه داریم، احتمالاً نتایج پایایی به دست خواهد آمد که برای تصمیم‌گیری در مورد رد و قبول دانشجویان قابل اتکا هستند. اما در حالتی که تعداد کمتری ایستگاه با زمان طولانی‌تر داریم، اثر بیشتری بر یادگیری دانشجویان خواهد داشت؛ مخصوصاً اگر ارزیابی در هر ایستگاه با بازخورد همراه شود (اسمی ۲۰۰۳).

تدوین بلوپرینت

از آنجا که آزمون OSCE بری مقاطع و رشته‌های مختلف برگزار می‌شود، قبل از آغاز طراحی سؤال، باید در مورد توانمندهایی که قرار است در این آزمون ارزیابی شوند، به صورت شفاف تصمیم‌گیری شود و فهرستی از مهارت‌های موردنظر تهیه گردد. تهیه فهرست توانمندی‌های مورد انتظار که با همکاری و هماهنگی مسؤولان آموزشی و اجرایی دوره صورت می‌گیرد، شاید برای بار اول کمی دشوار باشد اما کار را برای دفعات بعدی آسان می‌کند. به عنوان مثال، در کانادا، چارچوب CanMEDS که هفت نقش را برای پزشک در نظر می‌گیرد، مبنای بسیاری از آزمون‌ها برای تدوین بلوپرینت است. در دانشگاه علوم پزشکی تهران، توانمندی‌های مورد انتظار از دانش‌آموخته پزشکی در هشت حیطه تعریف شده است که در بخش اول کتاب مورد بررسی قرار گرفت (میرزازاده و همکاران ۲۰۱۴).

پس از مشخص شدن فهرست توانمندی‌ها یا نقش‌های مورد انتظار، از آنجا که تمامی آنها در یک OSCE قابل ارزیابی نیستند، چند مهارت باید از بین فهرست انتخاب شوند. چون عملکرد دانشجو در یک ایستگاه، فاکتور ضعیفی برای پیش‌بینی عملکرد او در سایر حیطه‌ها می‌باشد، نمونه‌گیری درست و به تعداد مناسب از بین مهارت‌ها و توانمندی‌ها، روایی محتوایی و پایایی آزمون را تقویت می‌کند (ون‌درولوتن ۲۰۰۴). برای اینکه مهارت‌های انتخابی از انواع مختلف و در حیطه‌های گوناگون باشند، مانند هر آزمون دیگری استفاده از بلوپرینت یا جدول مشخصات آزمون توصیه می‌شود. نحوه تهیه بلوپرینت آزمون در بخش اول کتاب بیان شد. در اینجا چند نمونه بلوپرینت OSCE را مرور می‌کنیم.

جدول ۲-۱۸ تجربه جفریس و همکاران[1] (۲۰۰۷) در برگزاری یک OSCE با محتوای مراقبت‌های بارداری را نشان می‌دهد که در آن تلاش شده است هر ایستگاه به گونه‌ای طراحی شود که چند نقش از چارچوب CanMEDS را ارزیابی کند.

جدول ۲-۱۸: بلوپرینت یک آزمون OSCE بر اساس چارچوب CanMEDS

ایستگاه	متخصص پزشکی[1]	برقرارکننده ارتباط[2]	همکار[3]	مدیر[4]	ارتقاءدهنده سلامت[5]	محقق[6]	حرفه‌ای[7]
مشاوره بیماری قلبی	*	*					*
اداره بیماری قلبی	*	*	*			*	
برنامه ترخیص	*	*		*	*		
انتقال	*	*	*	*			*
مشاوره بارداری	*	*					*
پی‌گیری کاهش شنوایی	*	*	*	*	*		
آموزش تعبیه کاتتر	*	*				*	*
دادن خبر بد	*	*			*		*
وقایع مرتبط با شیردهی	*	*		*	*		*
احیاء	*	*	*				

1 Medical expert
2 Communicator
3 Collaborator
4 Manager
5 Health advocator
6 Health advocator
7 Professional

1. Jefferies et al

تدوین بلوپرینت فقط به یک مدل خاص محدود نمی‌شود. بسته به اهداف دوره و محتوای آن انواع گوناگونی از بلوپرینت قابل طراحی می‌باشند. برای تدوین بلوپرینت، همچنین می‌توان جدولی دوبعدی در نظر گرفت که از یک سمت اهداف و حیطه‌های مورد نظر در آن فهرست شده‌اند و از طرفی دیسیپلین‌هایی که قرار است در آزمون نقش داشته باشند. سپس باید سهم هر حیطه/دیسیپلین را از لحاظ تعداد سؤال در آزمون مشخص کرد. به جدول ۳–۱۸ توجه کنید.

جدول ۳–۱۸: نمونه بلوپرینت OSCE پایان دوره پزشکی با ۱۲ ایستگاه

دیسیپلین/حیطه	شرح حال	معاینه	پروسیجر	مهارت‌ارتباطی	تشخیص	درمان
اطفال		❊				❊
جراحی			❊	❊		
زنان	❊				❊	
اورژانس			❊		❊	
داخلی	❊			❊		
گوش، حلق و بینی						❊
چشم پزشکی		❊				

بلوپرینت آزمون باید متناسب با اهداف آموزشی دوره تدوین شود و امکان انتخاب نمونه خوبی از مهارت‌ها و توانمندی‌های دوره را تسهیل کند. به عنوان مثال، امتحانی که برای دانشجویان پزشکی در مقطع فیزیوپاتولوژی برگزار می‌شود، احتمالاً بیشتر بر اصول اولیه شرح حال و انجام معاینه طبیعی متمرکز است. در حالی که در OSCE پایان دوره پزشکی انتظار می‌رود دانشجو با سناریوی پیچیده‌تری مواجه شود یا در OSCE دستیاری مهارت‌هایی مانند حل مسأله، استدلال بالینی و پروسیجرهای مرتبط مورد ارزیابی قرار گیرد. به جدول ۴–۱۸ که برای OSCE در دوره دستیاری بیماری‌های داخلی تدوین شده است و جدول ۵–۱۸ که بلوپرینت یک OSCE دندانپزشکی را نشان می‌دهد (شونیم کلین و همکاران[1] ۲۰۰۵)، توجه کنید.

جدول ۶–۱۸ فهرست ایستگاه‌های آزمون OSCE پیش‌کارورزی دانشگاه علوم پزشکی تهران را نشان می‌دهد که در فاصله سال‌های ۱۳۸۸ تا ۱۳۹۱ برای دانشجویان پزشکی در بدو ورود به دوره پیش‌کارورزی برگزار شده است.

1. Schoonheim-Klein et al.

جدول ٤-۱۸: نمونه بلوپرینت OSCE دستیاری بیماری‌های داخلی

تخصص	فهرست بیماری‌ها	شرح حال	معاینه فیزیکی	تفسیر تست تشخیصی	تفسیر تصویربرداری	تفسیر نمای پاتولوژی	تشخیص بیماری	درمان و اداره بیماری
غدد	دیابت		❊					❊
	تیروئید			❊			❊	
	هیپوفیز				❊			
روماتولوژی	لوپوس	❊						
	روماتویید		❊					
	آرتریت							
گوارش	کیسه صفرا				❊			
	کانسر				❊			
ریه	بیماری‌های انسدادی ریه			❊				
قلب	دیس‌ریتمی			❊			❊	
	تامپوناد				❊			
نفرولوژی	گلومرلونفریت			❊				
	اختلالات اسید و باز			❊				❊
خون	لوکمی					❊		
	آنمی		❊					❊
نورولوژی	Stroke	❊	❊					
	بیماری‌های دمیلینیزان				❊			

جدول ۵-۱۸: نمونه بلوپرینت OSCE دندانپزشکی

گروه / حیطه	مهارت ارتباطی	مهارت بالینی	مهارت بالینی
پریودنتولوژی		پر کردن چارت وضعیت پاکت با استفاده از فانتوم	تشخیص رادیوگرافیک ارتفاع استخوان
ترمیمی		ترمیم موقت قرار دادن رابردم	تشخیص پوسیدگی در عکس
اندودانتیکس		درخواست ابزارهای اندو جهت پرکردن ریشه	
علوم اجتماعی	تامل در خصوص احساسات بیمار دادن خبر بد به بیمار		
رادیولوژی		گرفتن عکس از مانکن	تفسیر عکس
بهداشت دندان		کنترل عفونت جرم‌گیری	
ارتودانتیکس			اندازه گیری ارتودونتیک
پروتز	ارائه توضیح به بیمار دنچر	نصب دندان در دنچر تنظیم ارتیکولاتور انجام مراحل نصب روکش	ارزیابی تراش دندان برای روکش

جدول ۶-۱۸: ایستگاه‌های OSCE پیش کارورزی دوره پزشکی عمومی دانشگاه علوم پزشکی تهران طی سال‌های ۱۳۸۸ تا ۱۳۹۱

سال/حیطه	شرح حال و مهارت ارتباطی	معاینه	پروسیجر	نسخه	تفسیر
مهر ۸۸ (۱۰ ایستگاه)	بیماری تنفسی زن باردار	نورولوژی افتالموسکوپی	بخیه زدن نمونه خون شریانش	مراقبت پس از بخیه احیاء اطفال	گرافی ارتوپدی
اسفند ۸۸ (۱۳ ایستگاه)	سردرد کودک مبتلا به اسهال	پستان توش واژینال قلب و ریه رفلکس نوزادی کبد و طحال	بستن آتل سونداژ بخیه	درماتیت هایپوتیرووییدی مایع درمانی در سوختگی	
مهر ۸۹ (۱۲ ایستگاه)	سنکوپ اورژانس خودکشی	معاینه زنان ضعف اندام تحتانی معاینه پلاک مترشحه	نمونه گاز شریانی بخیه آتل	ترومای چشم احیاء بزرگسال	رسم منحنی تفسیر عکس قفسه صدری
اسفند ۸۹ (۱۳ ایستگاه)	فشار خون بیمار پانیک	افتالموسکوپی زن باردار معاینه حرکت دیافراگم	تعبیه راه وریدی شستشوی گوش جسم خارجی آتل گرفتن نوار قلب		تشخیص ضایعات پوستی رادیولوژی سوفلهای قلبی
اسفند ۹۰ (ایستگاه)	اوتیت اکسترن بیمار افسردگی بیمار دچار خونریزی واژینال	پاراپلژی فشار خون بخیه افتالموسکوپی پوست	انتوباسیون آتل گیری سونداژ	مشاوره ازدواج در تالاسمی واکسیناسیون	عکس ساده قفسه سینه و شکم، سی تی اسکن
مهر ۹۱ (۱۴ ایستگاه)	بیمار پانیک کودک با تشنج	کمردرد فشار خون افتالموسکوپی ضایعات پیگمانته	تست توبرکولین سونداژ شستشوی گوش احیا	مشاوره پیشگیری از بیماری‌های جنسی	نوار قلب در سکته قلبی CT در خونریزی مغزی عکس ساده گردن

سؤالی که در OSCE مطرح می‌شود، باید از جنس مهارت و اقدام عملی وظیفه باشد. دانشجو باید در هر ایستگاه وظیفه بالینی خاصی را انجام دهد که اغلب در مواجهه با یک بیمار واقعی یا بیمار شبیه‌سازی شده[1] (بیمارنما) تعریف می‌شود اما گاهی دانشجو باید در مقابل یک مانکن یا مولاژ یا طی یک شبیه‌سازی کامپیوتری کاری را انجام دهد. گاهی نیز نتایج آزمایش‌های مربوط به بیمار یا تصویر گرافی‌های به عمل آمده همراه با شرح‌حال بیمار در اختیار دانشجو قرار می‌گیرد تا به تفسیر آنها بپردازد یا نحوه استدلال و تصمیم‌گیری تشخیصی یا درمانی دانشجو مورد ارزیابی قرار گیرد که در این ایستگاه‌ها نیازی به آزمونگر نیست. موارد زیر از جمله حیطه‌های مورد ارزیابی در OSCE هستند:

- مهارت ارتباطی و حرفه‌ای (به عنوان مثال، دادن خبر بد به بیمار، توضیح عوارض جراحی)
- گرفتن شرح‌حال (به عنوان مثال، گرفتن شرح‌حال از بیماری که با درد حاد شکم مراجعه کرده است)
- مهارت معاینه فیزیکی (به عنوان مثال، معاینه قلبی در بیمار)
- انجام پروسیجر (به عنوان مثال، بخیه زدن)
- مهارت استدلال بالینی (به عنوان مثال، تفسیر داده‌های بالینی و آزمایشگاهی، تشخیص بیماری، تجویز دارو و نوشتن نسخه)

1. Simulated patient

- مهارت ثبت وقایع و مستندات (نوشتن شرح حال، برنامه تشخیصی، خلاصه پرونده)

برای انتخاب مهارت‌ها از بلوپرینت دو رویکرد قابل استفاده است: یکی اینکه هر ایستگاه روی وظایف یا مهارت‌های ذکر شده به صورت جداگانه تمرکز کند و رویکرد دیگر اینکه چندین مهارت به صورت ادغام‌یافته در یک ایستگاه ارزیابی شوند. مدل اول بیشتر شبیه چیزی است که هاردن به صورت اولیه توضیح داده بود. در حالی که در مدل دوم، زمان هر ایستگاه طولانی‌تر است و امکان این وجود دارد که مواجهه کامل دانشجو با یک بیمار از ابتدا تا انتها ارزیابی شود. همان طور که قبلاً ذکر شد، برخی از این مدل تحت عنوان OSCA یاد می‌کنند. بخش بالینی آزمون USMLE مثال خوبی از این رویکرد است که در آن چندین توانمندی از جمله مهارت ارتباطی، شرح حال، معاینه و ثبت اطلاعات مورد ارزیابی قرار می‌گیرند. در امتحانات پرستاری نیز نمونه ایستگاه ادغام‌یافته وجود دارد که در آن مهارت ارتباطی دانشجو، توانایی مشاهده، ثبت علایم حیاتی، پانسمان و مراقبت زخم، انتقال بیمار و ... در کنار یکدیگر مورد ارزیابی قرار می‌گیرد.

در هر حال، باید توجه شود که OSCE برای سنجش صلاحیت بالینی دانشجویان در محیطی مصنوعی ساخته شده است. مسائلی مانند رفتارهای حرفه‌ای، نگرش دانشجویان و مراقبت بلندمدت آنها از بیماران برای این آزمون مناسب نیستند و روایی آن را زیر سؤال می‌برند.

مشخص کردن فضای آزمون

در واقع OSCE را نمی‌توان در هر مکانی برگزار کرد. فضایی که به برگزاری OSCE اختصاص می‌یابد، باید ویژگی‌های خاصی داشته باشد. هرچند که برخی از دانشگاه‌ها مکانی مخصوص برگزاری OSCE دارند، در اغلب دانشکده‌ها، فضاهایی که به طور معمول برای کارهای دیگر به کار می‌رود، برای آزمون آماده می‌شود. بنابراین، گاهی لازم است که از مدت‌ها قبل رزرو محل صورت گیرد.

یکی از فضاهای مورد نیاز برای برگزاری OSCE فضای استقرار ایستگاه‌ها هستند. فضای مورد نظر باید چندین اتاق جداگانه داشته باشد تا از آنها به عنوان ایستگاه استفاده شود. گاهی یک سالن بزرگ را به صورت موقت توسط پاراوان به قسمت‌های کوچک تقسیم می‌کنند. در این حالت مهم است که صدا از خلال پرده عبور نکند. هر اتاق باید فضای کافی به اندازه یک بیمار، یک آزمونگر، یک دانشجو و تجهیزات داشته باشد. البته در مواردی که موضوع ایستگاه چیزی مانند تفسیر گرافی است، به فضای کوچک‌تری نیاز است. دقت به ریزه‌کاری‌ها به برگزاری بهتر آزمون کمک می‌کند؛ مثلاً بهتر است اتاق کم نور به افتالموسکوپی اختصاص داده شود یا ایستگاهی که در آن دانشجو باید از بیمار شرح‌حال بگیرد، کنار ایستگاه سمع ریه گذاشته نشود.

فضای دیگر قرنطینه[1] آزمون است. این فضا دانشجویانی را که امتحان داده‌اند از دانشجویانی که هنوز در امتحان شرکت نکرده‌اند، جدا می‌سازد. مخصوصاً هنگامی که OSCE در دو نوبت صبح و عصر برگزار می‌شود، این خطر وجود دارد که محتوای آزمون بین دانشجویان گروه‌های مختلف رد و بدل شود. به همین علت دانشجویان نوبت صبح و عصر را در دو قرنطینه جداگانه نگه می‌دارند. البته توافق کلی در متون وجود دارد که حتی در صورت لو رفتن سؤالات OSCE، اثر چشم‌گیری بر عملکرد دانشجویان در آزمون اتفاق نمی‌افتد (کولیور و همکاران[2] ۱۹۹۱(۱)، سوانسون و همکاران[3] ۱۹۹۹). زیرا در این آزمون بیشتر از اینکه دانش ارزیابی شود، مهارت مورد سنجش قرار می‌گیرد و در فاصله زمانی اندک بین دو آزمون یادگرفتن و تمرین مهارت‌ها اصلاً ساده نیست اما چنین مواردی، اعتبار آزمون را زیر سؤال می‌برند. به همین منظور برخی از دانشکده‌ها دانشجویان صبح و عصر را در دو سالن قرنطینه نگه می‌دارند. بسته به نحوه چیدمان ایستگاه‌ها، قرنطینه می‌تواند قبل از آزمون یا در پایان آزمون باشد. گاهی دانشجویانی که صبح امتحان داده‌اند، پس از امتحان، حق

1. Quarantining or Corralling
2. Colliver et al
3. Swanson et al

خارج شدن از فضای آزمون را ندارند و مستقیماً پس از ایستگاه آخر وارد قرنطینه می‌شوند. در زمان مقرر، تمام دانشجویانی که باید در نوبت عصر امتحان دهند، در قرنطینه دیگری جمع می‌شوند. به این ترتیب دو گروه با یکدیگر برخوردی ندارند. با شروع آزمون عصر، دانشجویان نوبت صبح می‌توانند از قرنطینه خارج شوند. طی مدتی که دانشجویان در قرنطینه هستند، نباید با بیرون ارتباط داشته باشند. از همین رو دسترسی ایشان به گوشی تلفن همراه و سایر وسایل ارتباطی الکترونیک محدود می‌شود.

فضای دیگر، محل ورود و ثبت‌نام دانشجویان است. دانشجویان در بدو ورود باید کارت شناسایی خود را نشان دهند و نام خود را ثبت کنند. همچنین فضایی باید به نگهداری گوشی همراه و وسایل شخصی دانشجویان اختصاص پیدا کند. ممکن است این فضا چندان مهم به نظر نرسد اما اگر از قبل به آن فکر نشود، در روز امتحان می‌تواند ازدحام ایجاد کند. همچنین لازم است اتاقی در نظر گرفته شود تا بیمارنماها و آزمونگران در آن جمع شوند، استراحت کنند یا پذیرایی شوند.

تصمیم‌گیری در مورد روش تعیین حد نصاب قبولی آزمون

روش‌های متعددی برای تعیین حد نصاب قبولی یا استاندارد آزمون وجود دارد که به دو دسته کلی هنجاری و معیاری تقسیم می‌شوند. از آنجا که معمولاً هدف OSCE تعیین صلاحیت بالینی دانشجویان است و نه مقایسه عملکرد دانشجویان با یکدیگر، منطقی‌تر است که از استانداردهای معیاری استفاده شود (نیوبل ۲۰۰۴، خان و همکاران ۲۰۱۳).

اهمیت تصمیم‌گیری در مورد روش تعیین حدنصاب از آن جهت است که برخی از روش‌ها قبل از آزمون اجرا می‌شوند و برخی دیگر در حین آزمون یا بعد از آزمون. همچنین هر یک از آنها تمهیدات خاصی را می‌طلبند که باید برنامه‌ریزی دقیقی برای آن صورت گیرد. به عنوان مثال، در روش انگوف[1]، بعد از طراحی سؤالات و قبل از برگزاری آزمون، پانلی از داوران متخصص تشکیل می‌شود تا با ارزیابی ایستگاه‌ها و چک‌لیست‌ها، میزان احتمال قبولی دانشجوی مرزی[2] را برآورد کنند. در حالی که روش گروه مرزی[3] حین OSCE اجرا می‌شود و مستلزم آن است که نمره‌دهی گلوبال نیز در کنار چک‌لیست گنجانده شود و همچنین برای ارزیابی عملکرد دانشجو حتماً از اعضای هیأت علمی استفاده گردد.

جزئیات مطالب مربوط به تعیین استاندارد در بخش هفتم کتاب ارائه خواهد شد.

تصمیم‌گیری در مورد نحوه نمره‌دهی

کمیته باید در مورد ابزار ارزیابی در ایستگاه تصمیم‌گیری کند. در ایستگاه مشاهده‌ای، دو رویکرد برای ارزیابی عملکرد دانشجو به کار می‌رود: یکی رویکرد جزئی‌نگر[4] یا همان استفاده از چک‌لیست که شناخته شده‌تر است و رویکرد کلی‌نگر[5] یا نمره‌دهی گلوبال[6]. در انتهای فصل اول این بخش به این پرداختیم که کدام یک برای استفاده در OSCE ارجح است و برای انتخاب آنها چه مسائلی باید مدنظر قرار گیرند. اما جدا از مسائل مربوط به پایایی که محور بحث در خصوص نمره دهی چک‌لیست و گلوبال است، کمیته باید مدنظر داشته باشد که برای تعیین استاندارد به شیوه گروه مرزی یا رگرسیون مرزی باید همزمان هر دو مقیاس را استفاده کند. همچنین، برای انتخاب از میان این دو مقیاس نمره‌دهی، نمی‌توان منکر این قضیه شد که گاهی محدودیت‌های اجرایی تعیین‌کننده هستند. به عنوان مثال، چنانچه نتوان برای ارزیابی تمام ایستگاه‌ها از هیأت علمی و آزمونگران متخصص دعوت کرد، نمی‌توان نمره‌دهی گلوبال را به راحتی انجام داد.

---

1. Angoff
2. Borderline
3. Borderline Group Method
4. Analytic
5. Holistic
6. Global rating

نکته قابل توجه در نمره‌دهی چک‌لیست است که به دو صورت دوحالته[1] و درجه‌ای[2] قابل تنظیم است:

- در نمره‌دهی دوحالته، که اولین بار توسط هاردن معرفی شد، در هر آیتم، اگر دانشجو کار مورد نظر را انجام دهد، یک نمره کسب می‌کند و اگر انجام ندهد، صفر می‌گیرد. در اینجا، کیفیت کار مطرح نیست و صرف انجام آن مهم است. سابقاً تصور می‌شد که این نحوه نمره دادن، عینیت ارزیابی را بسیار بالا می‌برد و نهایتاً پایایی آزمون را بهبود می‌بخشد اما این مسأله که عینیت به صورت مستقیم منجر به پایایی می‌شود، امروزه چندان مورد قبول نیست. شواهد حاکی از آن هستند که نمره‌دهی دوحالته به خوبی قادر به افتراق دانشجوی توانمند و غیرتوانمند نیست (خان و همکاران ۲۰۱۳).
- در حالت دیگر نمره‌دهی چک‌لیست که درجه‌ای نام دارد، برای هر آیتم، لیکرتی با پنج تا هفت درجه در نظر گرفته می‌شود. در این حالت، امکان قضاوت درباره کیفیت کار انجام شده نیز وجود دارد. توجه کنید که این نمره‌دهی با نمره‌دهی گلوبال متفاوت است.

در مورد جزئیات تدوین هر یک از موارد فوق در قسمت طراحی ایستگاه صحبت خواهد شد.

تصمیم‌گیری در مورد نحوه مدل‌های جبرانی، نیمه جبرانی و غیرجبرانی

مدل نمره‌دهی به این موضوع دلالت دارد که نمرات سؤالات/ ایستگاه‌های مختلف را با چه روشی با هم ترکیب کنیم تا به نمره کل برسیم. جزئیات این مبحث در بخش هفتم کتاب به تفصیل آمده است و در اینجا به طور مختصر به آن اشاره می‌شود:

- اگر نمرات حاصل از تمام سؤالات/ ایستگاه‌ها با هم جمع شود و در مجموع محاسبه شود که آیا دانشجو قبول شده یا نه، از مدل «کاملاً جبرانی»[3] استفاده شده است. یکی از انتقادات جدی که به این مدل وارد می‌شود این است که احتمال دارد دانشجو در یک مهارت اصلاً توانمندی نداشته باشد و در عین حال از کل آزمون نمره قبولی بگیرد زیرا عملکرد او در سایر ایستگاه‌ها، توانسته این نقص وی را جبران کند.
- گاهی لازم است دانشجو در تک‌تک سؤالات/ ایستگاه‌ها به حد نصاب برسد، در این صورت مدل «غیرجبرانی»[4] نامیده می‌شود. یعنی اگرچه در کل نمره دانشجو به حد نصاب رسیده است اما از آنجا که در یک یا چند ایستگاه کمتر از حد مورد انتظار ظاهر شده است، در نهایت در آزمون نمره قبولی نمی‌گیرد.
- نوع سومی هم وجود دارد که در واقع مخلوط این دو حالت است به این معنا که حدنصاب برای تعدادی از سؤالات/ ایستگاه‌ها و نه تمام آنها گذاشته می‌شود و در نهایت نمره کل نیز باید به حد نصاب برسد. به این مدل «نسبتاً جبرانی»[5] گفته می‌شود. مثلاً ممکن است کمیته آزمون تصمیم بگیرد از آنجا که حوزه مهارت‌های ارتباطی مهم است، دانشجو حتماً باید در ایستگاه مربوطه نمره قبولی را کسب کند و بدون آن در کل امتحان قبول محسوب نمی‌شود.

در اینجا باید به رابطه‌ای که ممکن است بین چند سؤال یا ایستگاه با یکدیگر وجود داشته باشد، توجه کرد. مثلاً اگر در یک ایستگاه، مهارت شرح‌حال مورد ارزیابی قرار می‌گیرد و در ایستگاه دیگر مهارت برقراری ارتباط، باید توجه داشت که چگونگی برقراری ارتباط دانشجو با بیمار، روی مهارت شرح‌حال‌گیری او هم اثر خواهد گذاشت. استفاده از روش غیرجبرانی در صورتی امکان‌پذیر است که بین محتوای مورد ارزیابی در ایستگاه های مختلف رابطه وجود نداشته باشد و یا این ارتباط قابل چشم‌پوشی باشد.

یکی دیگر از راه‌کارها برای رویارویی با مسأله جبران کردن عملکرد یک ایستگاه توسط ایستگاه دیگر، گزارش موارد نگران‌کننده و غیرقابل پذیرش است. از جمله مواردی که در طی OSCE نگرانی جدی ایجاد می‌کنند، رفتار غیرحرفه‌ای دانشجو با بیمار یا عملکرد مخاطره‌آمیز دانشجو است. این سیستم که در برخی از دانشگاه‌ها به «کارت زرد»[6] مشهور است،

1. Bianary
2. Rating scale
3. Total Compensatory
4. yellow card

به دانشجو در مورد رفتارش بازخورد می‌دهد. متعاقب این قضیه، دانشجو به یک عضو هیأت علمی معرفی می‌شود تا در مورد رخداد پیش آمده، بازاندیشی کند یا قبل از رفتن به مقطع بعدی باید دوره‌های اصلاحی[1] را پشت سر بگذارد. البته در مورد اثربخشی این سیستم، شواهد تجربی در دسترس نیست.

تصمیم‌گیری در مورد جلسه توجیهی[2] و بازخورد

کمیته آزمون باید با در نظر گرفتن هدف آزمون، امکانات و زمان مورد نیاز در مورد لزوم و نحوه برگزاری جلسه توجیهی و بازخورد تصمیم‌گیری کند.

جلسه توجیهی برای دانشجویان، بیمارنماها و آزمونگران برگزار می‌شود تا اطلاعات لازم را به آنها ارائه کند. بدیهی است که اطلاعات مورد نیاز این سه گروه با هم متفاوت است، بنابراین جلسات جداگانه‌ای باید برای آنها تنظیم شود. این جلسه قبل از برگزاری آزمون (مثلاً به فاصله یک تا دو هفته قبل) و همچنین در روز آزمون برگزار می‌شود که در مورد دوم خلاصه تر است و جزئیات آن در قسمت بعدی (اجرای OSCE) خواهد آمد.

جلسه بازخورد پس از برگزاری امتحان برگزار می‌شود و می‌تواند دو هدف داشته باشد؛ بازخورد به دانشجویان در مورد نحوه عملکرد و سطح مهارت آنها؛ و بازخوردی که دانشجویان و آزمونگران در مورد کیفیت امتحان ارائه می‌دهند.

در مواردی که OSCE تکوینی است و در طی دوره برگزار می‌شود یا قرار است دانشجویان را برای آزمونی مهم آماده کند، ضرورت جلسات بازخورد بیشتر می‌شود. از طرفی اگر تعداد دانشجویان بسیار زیاد است، همکاری هیأت علمی مطلوب نیست یا زمان آزمون بیش از حد طولانی است، شاید بهتر باشد که از آن صرف‌نظر گردد یا به شکل دیگری تعدیل شود؛ مثلاً:

- به جای اینکه جلسه بازخورد پس از هر ایستگاه برای هر دانشجو برگزار شود، به صوت کلی پس از آزمون در نظرگرفته شود؛
- به جای اینکه جلسه بازخورد برای تک‌تک دانشجویان به صورت جداگانه تشکیل شود، یک جلسه کلی با حضور همه دانشجویان برگزار شود یا تنها برای آن دسته از شرکت‌کنندگان که عملکرد ضعیفی از خود نشان داده‌اند؛
- به جای استفاده از هیأت علمی، جلسات توسط افراد آموزش دیده‌ای غیر از هیأت علمی اداره شود.

خودارزیابی را نیز می‌توان به عنوان قسمتی از جلسه بازخورد در نظر گرفت. می‌توان به گونه‌ای برنامه‌ریزی نمود که بازخورد دانشجویان نسبت به آزمون نیز در همان زمان اخذ شود. همچنین اگر جلسات ضبط شوند، بازخورد دادن روی عملکرد شرکت‌کنندگان تسهیل می‌شود.

طراحی ایستگاه

تا اینجا مشخص شده است که آزمون چند ایستگاه دارد و زمان هر ایستگاه چقدر است. همچنین بر اساس بلوپرینت دوره، موضوع ایستگاه‌های مختلف و نحوه نمره‌دهی در آنها نیز تعیین شده است. در این مرحله باید طراحی ایستگاه‌ها صورت گیرد. حتی اگر دانشکده سابقه برگزاری OSCE دارد و از بانک ایستگاه استفاده می‌کند، مواقعی پیش می‌آید که نیاز به طراحی ایستگاه‌های جدید وجود دارد. برای دانشکده‌ای که قبلاً OSCE برگزار نکرده است، بدیهی است که طراحی ایستگاه از ابتدا ضرورت پیدا می‌کند. نظر به اهمیت این موضوع و نکات و ریزه‌کاری‌های مهم، پس از پایان این قسمت یعنی طراحی آزمون به صورت مفصل به طراحی ایستگاه پرداخته خواهد شد. در اینجا چندین نکته مهم که مسؤول OSCE باید قبل از آغاز طراحی ایستگاه‌ها به آن توجه کند، بیان می‌شوند:

- قبل از مشخص کردن اسامی هیأت علمی و قبل از واگذاری کار به آنها اول باید مشخص شود که نوع ایستگاه چیست. لزوماً تمام ایستگاه‌ها به صورت برخورد پزشک با یک بیمار نیستند و از این نظر به چهار دسته تقسیم می‌شوند

1. Remedial
2. Debriefing

که در قسمت طراحی ایستگاه به آن پرداخته می‌شود. سپس مسؤول OSCE، هر ایستگاه را به یک نفر از اعضای هیأت علمی که از نظر محتوایی متخصص موضوع است، می‌سپارد.

- طراحان ایستگاه‌ها لزوماً دست اندرکار برگزاری آزمون و درگیر مسائل اجرایی آن نیستند. پس لازم است مسؤول برگزاری OSCE به اندازه کافی اطلاعاتی در مورد هدف آزمون، ساختار آزمون، تعداد ایستگاه‌ها، زمان هر ایستگاه، نحوه نمره‌دهی، سطح شرکت‌کنندگان و ... در اختیار آنها بگذارد.
- طراحان ایستگاه ممکن است تجربه قبلی برای این کار نداشته باشند و آموزش ساختارمندی دریافت نکرده باشند. پس ضروری است در ابتدای کار در یک جلسه یا کارگاه توجیهی، با اصول و مفاهیم پایه OSCE و نحوه طراحی پایه سؤال و سناریو، تدوین چک‌لیست، تدوین راهنمای بیمارنما و ... آشنا شوند. می‌توان در قالب یک تمرین از طراحان خواست که یک ایستگاه را به عنوان نمونه تدوین کنند و سپس به آنها بازخورد داد.
- از آنجا که طراحان متعدد هستند، احتمالا در طراحی ایستگاه‌ها مطابق با سلیقه خودشان عمل می‌کنند. بنابراین، دور از ذهن نیست که از فونت، رنگ، کادربندی و ساختارهای متفاوتی استفاده کنند. بهتر است مسؤول OSCE یک چارچوب و قالب یکسان برای سناریوی ایستگاه، راهنمای دانشجو، راهنمای آزمونگر، چک‌لیست، راهنمای بیمارنما و ... تدوین کند و همان را در اختیار طراحان قرار دهد. اگر این امر از ابتدا صورت نگیرد، مسؤول OSCE نهایتا خودش باید تمام متون و مستندات را یکسان کند که وقت‌گیر است. اگر هم بخواهد ایستگاه‌ها را به همان شکل متنوع و ناهمگون در اختیار دانشجویان قرار دهد، کار مناسبی نیست و باعث سردرگمی فراگیران می‌شود.
- طراحان ایستگاه لزوماً همان افرادی نیستند که در روز امتحان به عنوان آزمونگر در ایستگاه حضور دارند. پس لازم است که از آنها خواسته شود مواد لازم را طوری آماده کنند که بدون حضور خودشان ایستگاه قابل درک و قابل اجرا باشد.
- حتی اگر طراحان سؤال، سابقه طراحی ایستگاه را داشته باشند، ضروری است که پس از دریافت ایستگاه‌ها جلسه مرور برای ارزیابی ایستگاه‌ها برگزار شود. این کار بهتر است در یک پانل با حضور اعضای هیأت علمی برگزار شود. به این صورت که تک تک ایستگاه‌ها خوانده شوند و بر اساس اصول طراحی مورد نقد و ارزیابی قرار گیرند. اما گاهی به دلیل محدودیت‌های اجرایی، مسؤول OSCE به تنهایی این کار را انجام می‌دهد. حالت بهتر این است که برای حداکثر استفاده از نظر هیأت علمی و در عین حال صرفه جویی در وقت، طراحان به یک جلسه مشترک دعوت شوند تا در ابتدای آن به طراحی ایستگاه‌های خود بپردازند، در ادامه، ایستگاه‌های یکدیگر را ارزیابی کنند و در نهایت، بر اساس بازخوردهای داده شده، اصلاحات و تغییرات را اعمال نمایند. سنجش روایی نیز در این مرحله با اخذ نظر هیأت علمی قابل انجام است. توضیحات بیشتر در این خصوص در قسمت ارزشیابی آزمون خواهد آمد.

در قسمت بعدی (طراحی ایستگاه) به صورت مفصل گام‌های طراحی یک ایستگاه بیان خواهد شد.

مشخص کردن امکانات لازم برای ایستگاه‌ها

زمانی که ایستگاه طراحی شود، مشخص می‌شود که چه امکاناتی برای چیدن ایستگاه لازم است. در صورتی که آزمون چند لاین داشته باشد، وسایل و تجهیزات باید به تعداد لاین‌ها تهیه شود. هیچ وسیله‌ای را نباید کم اهمیت شمرد. نبود کوچکترین و پیش پاافتاده‌ترین وسیله در روز آزمون می‌تواند منجر به تاخیر در برگزاری آزمون، به هم خوردن نظم ایستگاه‌ها و مخدوش شدن کل امتحان شود. به عنوان مثال، یکی از موارد مهم، داشتن ذخیره کافی برای موارد مصرفی ایستگاه است که از این بین می‌توان باتری را نام برد.

بسته به میزان منابع مالی که برای آزمون در نظر گرفته شده است، می‌توان از تکنولوژی‌های پیشرفته از جمله استفاده از شبیه‌سازی‌ها، نمره‌دهی کامپیوتری، فرم‌های ارزشیابی مبتنی بر شبکه و ... در مراحل مختلف آزمون استفاده کرد. اما جایی که امکانات محدود است، وسایل جایگزین قابلیت اجرای OSCE را افزایش می‌دهند.

دعوت از بیماران یا بیمارنمایان و آماده‌سازی آنها

زمانی که سناریوی ایستگاه نوشته شد، مشخص می‌شود که آیا دانشجو در ایستگاه قرار است در مواجهه با یک بیمار/بیمارنما قرار بگیرد یا خیر. معمولاً دانشگاه‌ها یا مراکز مهارت بالینی، بانکی از بیماران واقعی و بیمارنمایان دارند که افراد در آن عضو هستند و تحت نظارت مسؤول برنامه بیمارنما، برای شرکت در OSCE آموزش دیده‌اند. به فراخور امتحانی که قرار است برگزار شود، به تعداد لازم بیمار یا بیمارنما از بانک فراخوانده می‌شود. در صورتی که بانکی از بیمارنمایان وجود ندارد، افرادی باید برای این کار مشخص و دعوت شوند.

از آنجا که یکی از منابع شناخته شده خطا در OSCE نحوه عملکرد بیماران است، باید تلاش شود تا عملکرد بیمارنمایان استاندارد و یکسان شود. تفاوت در عملکرد بیماران، منبع خطایی است که پایایی آزمون را تحت تاثیر قرار می‌دهد زیرا می‌تواند موجب شود نمرات دانشجویان تغییراتی کند که ناشی از سطح توانمندی دانشجویان نیست. برای استاندارد کردن عملکرد بیمارنمایان، توجه به موارد زیر لازم است:

- انتخاب بیمار یا بیمارنما: حین انتخاب افراد باید به مشخصات و ویژگی‌های آنها دقت کرد که برخی از آنها ذاتی و هنگام انتخاب افراد مشهود است اما برخی در دوره آموزشی مشخص می‌شود (جدول ۷–۱۸)
- تدوین برگه‌های راهنمای بیمارنما: راهنمای بیمارنما توسط طراح ایستگاه آماده می‌شود و در اختیار بیمارنمایان قرار می‌گیرد. در این باره در قسمت طراحی ایستگاه بیشتر صحبت خواهد شد. فقط اینکه باید دقت شود که در راستای آموزش بیمارنمایان، سؤالات و موضوعات ایستگاه‌ها نزد دانشجویان فاش نشود. تمام هماهنگی‌ها و آموزش‌ها باید با حفظ مسائل مربوط به امنیت آزمون صورت گیرد.
- آموزش بیمار یا بیمارنما: گذاشتن جلسات توجیهی حضوری قبل از آزمون برای بیمارنمایان به صورت کارگاه، نمایش فیلم و جلسات بحث حائز اهمیت است. همچنین ایجاد فرصتی که طی آن بیماران سناریوی ایستگاه را عملاً تمرین کنند و بازخورد دریافت نمایند، برای اجرای بهتر و با آمادگی بیشتر در روز آزمون کمک‌کننده است. میزان و سطح آموزشی که ارائه می‌شود، بستگی به سطح افراد و زمینه قبلی آنها دارد. هرچند که توافق در این خصوص وجود ندارد که زمان کافی برای آموزش یک بیمارنما چقدر باید باشد، یک مطالعه برآورد کرده است که احتمالاً این فرایند بسته به نقش مورد نظر و تجربه و وخصوصیات بیمارنما تا ۱۵ ساعت هم طول خواهد کشید (شاموی و هاردن[1] ۲۰۰۳).

از آنجا که حضور به موقع تمام افراد برای شروع OSCE ضروری است، علاوه بر اینکه باید به بیمارنماها در مورد اهمیت حضور به موقع تاکید شود، باید افرادی را به عنوان رزرو در نظر گرفت تا در صورت بروز اتفاقات غیرمنتظره جایگزین شوند.

برای توضیحات بیشتر در مورد تفاوت بیمار واقعی و بیمارنما به سؤالات رایج در فصل اول مراجعه کنید.

جدول ۷–۱۸: خصوصیات ذاتی و اکتسابی بیمارنمایان

(خصوصیات ذاتی بیمارنمایان (هنگام انتخاب	(خصوصیات اکتسابی بیمارنمایان (طی دوره آموزشی
سن، زبان، جنس، نژاد، عادات و استایل فیزیکی	سرعت و آمادگی در یادگیری
یافته‌های بالینی مانند اسکار، ضایعات پوستی	توانایی تطابق با انواع مختلف مصاحبه پزشکی
سطح سواد و تحصیلات	توانایی ایفای نقش موثر مطابق سناریو
تجربه در بیان بیماری	توانایی گوش دادن فعال و تغییر مطابق بازخوردهای ارائه شده
در دسترس بودن برای کارگاه یا جلسه آزمون	توانایی پاسخ دادن به سؤال غیرمترقبه با توجه به تجربیات حین تمرین
توانایی جسمانی	توانایی ثبت دقیق موارد و ارائه بازخورد سازنده از دیدگاه یک بیمار

1. Shumway & Harden

دعوت از آزمونگران و آماده‌سازی آنها

زمانی که سناریوی ایستگاه نوشته شد، مشخص می‌شود که آیا آزمونگر در ایستگاه مورد احتیاج هست یا خیر. در صورتی که نیاز باشد عملکرد دانشجو در ایستگاه توسط آزمونگر مورد مشاهده و ارزیابی قرار گیرد، باید آزمونگران مشخص شوند و از ایشان برای شرکت در روز آزمون دعوت به عمل آید. در اینجا هم برای مقابله با اتفاقات غیرمنتظره لازم است افرادی به عنوان آزمونگر رزرو در نظر گرفته شوند تا در صورت لزوم جایگزین شوند.

هرچند که معمول است آزمونگر در ایستگاه حاضر شود و به صورت مستقیم عملکرد دانشجو را مشاهده کند، اگر تجهیزات صوتی–تصویری وجود داشته باشد و ایستگاه ضبط شود، ممکن است آزمونگران در اتاق دیگری به مشاهده عملکرد دانشجو بپردازند.

آزمونگران نقش مهمی در برگزاری OSCE منصفانه و سالم ایفا می‌کنند. همان‌طور که قبلاً اشاره شد، به صورت ایده‌ال، تفاوتی که در نمرات امتحان OSCE دانشجویان دیده می‌شود، صرفاً باید به دلیل تفاوت بین سطح توانمندی خود دانشجویان باشد و علت دیگری نداشته باشد. هر عاملی که به عنوان منبع خطا موجب تفاوت بین نمرات شود، روی پایایی آزمون اثر می‌گذارد. در همین راستا، اگر نحوه نمره‌دهی دو آزمونگر یا میزان سختگیری آنها متفاوت باشد، مطلوب نیست و پایایی را مخدوش می‌کند. برای اینکه نحوه عملکرد آزمونگران یکسان و با ثبات باشد، دو مسأله باید مورد توجه قرار گیرد:

- **انتخاب آزمونگران:** کمیته باید در مورد این مسأله تصمیم‌گیری کند که ارزیابی در چه سطحی صورت می‌گیرد. در حالی که در انگلیس، معمول است که پزشکان متخصص به مشاهده و ارزیابی عملکرد دانشجویان بپردازند (گرملی ۲۰۱۱)، در برخی از موارد به دلیل محدودیت‌های اجرایی به جای عضو هیأت علمی، از دستیار، دانشجوی MD-PhD، دانشجوی سال بالاتر، دانش‌آموخته، پرستار و ... استفاده می‌شود تا در هزینه‌ها صرفه‌جویی شود. همچنین در برخی از دانشکده‌ها از بیمار حاضر در ایستگاه خواسته می‌شود تا عملکرد دانشجو را ارزیابی کند که در قسمت مسائل چالشی فصل قبل مورد بحث قرار گرفت. هنگام انتخاب آزمونگران ویژگی‌های آنان از جمله رشته تخصصی، سابقه آموزشی، سابقه کار بالینی و سابقه کار در OSCE باید در نظر گرفته شود. همچنین آزمونگران باید مشاهده‌گران خوب و دقیقی باشند.
- **آماده‌سازی و آموزش آزمونگران:** پس از انتخاب آزمونگران چه استاد باشند، چه دانشجو و چه بیمارنما، باید به طور جدی برای آموزش، تمرین و بازخورد دادن به آنها برنامه‌ریزی صورت گیرد. هرچند که تا مدتها تصور بر این بود که برخی از آزمونگران ذاتاً سختگیر و برخی دیگر سهل‌گیر هستند و این ویژگی قابل تغییر نیست و مداخلات آموزشی در این زمینه بی‌تاثیر هستند، مطالعات نشان داده‌اند آموزش، موجب یکسان شدن رفتار آزمونگران و کاهش واریانس بین آنها می‌شود و پایایی آزمون را بهبود می‌بخشد (خان و همکاران ۲۰۱۳). به طوری که در یک مطالعه نویسندگان تاثیر آموزش را بر آزمونگران مختلف شامل پزشک، دانشجوی پزشکی و افراد غیر پزشک بررسی کردند. گروه‌های آموزش دیده و آموزش ندیده، عملکرد دانشجویان را در فیلم‌های ضبط شده ارزیابی کردند. نتایج نشان داد که ضرورت آموزش و همچنین تاثیر آن برای گروه‌های مختلف متفاوت است. به طوری که بیشترین نیاز و تاثیر در نحوه ارزیابی افراد غیرپزشک مشاهده شده و کمترین نیاز و تاثیر در نحوه عملکرد پزشکان به دست آمد. با این حال، تفاوتی که بین عملکرد آزمونگران گروه‌های مختلف (بیمارنما، پزشک و دانشجوی پزشکی) قبل از آموزش وجود داشت، با آموزش از بین رفت (ون‌درولوتن و همکاران ۱۹۸۹).

دستورالعمل‌ها و توصیه‌های متعددی توسط موسسات مختلف برای مشخص کردن نقش، وظایف و مسؤولیت‌های آزمونگران دوین شده‌اند (GMC ۲۰۰۹، AoME ۲۰۱۲). همچنین، موسسات مختلف برنامه آموزشی ویژه آزمونگران خود ترتیب داده‌اند. بدیهی است که آموزش باید منطبق با سطح و زمینه قبلی آزمونگران باشد اما اهدافی که باید به صورت کلی مدنظر باشند، شامل این موارد هستند:

- اهداف و اصول OSCE را درک کنند.

- عملکرد حرفه ای خود را به طور یکنواخت در سرتاسر آزمون حفظ کنند.
- اصول نمره‌دهی (چک‌لیست و گلوبال) را بفهمند و به درستی به کار برند.
- در صورت لزوم در آزمون تراکمی، بازخورد کتبی به عملکرد دانشجویان دهند.
- در صورت لزوم در آزمون تکوینی، بازخورد شفاهی به عملکرد دانشجویان دهند.
- اصول رازداری و محرمانه بودن اطلاعات و نمرات دانشجویان را رعایت کنند.
- عملکرد نامناسب یا رفتار خطرناک دانشجویان را درک کنند.

از بین موارد فوق، اصول نمره‌دهی اهمیت بسیار زیادی دارد. نمره‌دهی بر اساس چک‌لیست برای آزمونگران مبتدی آسان‌تر است. زیرا آیتم‌ها نوشته شده و تنها کاری که آزمونگر باید انجام دهد، مشاهده دقیق است اما نمره‌دهی گلوبال نیاز به تجربه و آموزش بیشتری دارد. در ابتدا باید معنای هر کدام از سطوح در مقیاس لیکرت (مثلاً مرزی یا قابل قبول) برای آزمونگران تبیین شود. سپس باید فرصت‌هایی مهیا کرد تا آزمونگران تمرین کنند و بازخورد دریافت کنند. دانشگاه کوییز بلفاست یک دوره آموزشی انلاین[1] به منظور توانمندسازی آزمونگران در نمره‌دهی گلوبال طراحی کرده است که در آن آزمونگران پس از مشاهده فیلمی از عملکرد یک دانشجو به او نمره می‌دهند و سپس در مورد نحوه ارزیابی خود بازخورد دریافت می‌کنند.

آزمونگران باید موارد زیر را هنگام نمره‌دهی گلوبال به یاد داشته باشند:

- از تجربه و تخصص خود برای قضاوت در مورد عملکرد دانشجو استفاده کنند.
- نمره‌دهی گلوبال را مستقل از نمره چک‌لیست انجام دهند.
- مقطع دانشجویان شرکت‌کننده در آزمون را در نظر بگیرند.
- نمره‌دهی گلوبال را به صورت کلی نگر و جامع انجام دهند.
- از تمام شاخص‌ها استفاده کنند و خود را به قسمت میانی مقیاس، محدود نکنند.
- نگران اینکه دانشجویی را رد کنند، نباشند.

تقسیم‌بندی دانشجویان و اطلاع رسانی به آنها

فهرست اسامی دانشجویان شرکت کننده در آزمون باید از قبل نهایی شود و تقسیم بندی آنها در گروه‌ها صورت گیرد. باید مشخص شود که نوبت هر دانشجو دقیقاً چه زمانی است. یعنی در چه لاینی است، راند چندم است و با ایستگاه چندم شروع می‌کند. تقسیم‌بندی دانشجویان باید به نحو مناسب به اطلاع آنها برسد. بهتر است که برای سهولت کار، تقسیم‌بندی با کد شناسایی صورت گیرد. به عنوان مثال تصور کنید OSCE با ۱۵ ایستگاه و شرکت ۳۰۰ دانشجو در دو نوبت صبح و عصر انجام می‌شود که هریک دو لاین دارند (A و B و C و D). به عبارت دیگر، در هر لاین ۷۵ دانشجو حضور خواهد داشت که در ۵ دور امتحان می‌دهند. در هر دور ۱۵ دانشجو وارد آزمون می‌شوند. بنابراین کد C-3-12 دانشجویی است که از ایستگاه دوازدهم دور سوم نوبت عصر آزمون خود را شروع می‌کند. در برخی از آزمون‌ها برای صرفه‌جویی در وقت در ایستگاه آزمونگر لازم نیست نام و نام خانوادگی دانشجو را بپرسد و یادداشت کند. بلکه به هر دانشجو به تعداد ایستگاه برچسبی داده می‌شود که کد شناسایی مخصوص روی آن قید شده است. دانشجو با ورود به هر ایستگاه، یک برچسب به آزمونگر می‌دهد تا روی چک‌لیستی که تکمیل خواهد کرد، بچسباند.

همچنین در مورد اطلاع رسانی، باید اطمینان حاصل شود که دانشجویان تمام اطلاعات مورد نیاز را برای شرکت در این امتحان دارند. مخصوصاً اگر OSCE تجربه جدیدی است، باید توجه شود که اطلاع‌رسانی در خصوص اینکه شیوه برگزاری امتحان چگونه است، از ابتدای دوره صورت گیرد، نه نزدیک امتحان. ماهیت این آزمون فشار و استرس زیادی روی دانشجویان اعمال می‌کند. آشنا کردن دانشجویان با اتفاقاتی که در ایستگاه‌ها می‌افتد و نحوه چرخش آنها در کاهش

1. www.med.qub.ac.uk/OSCE

اضطراب آنها موثر است. به این منظور می‌توان از نمونه ایستگاه‌ها و یا فیلم‌ها و عکس‌های موجود استفاده کرد. به همین ترتیب، دادن نقشه آزمون به دانشجویان به آنها کمک می‌کند دیدی کلی نسبت به فضای آزمون پیدا کنند.

نهایی کردن ماتریال ایستگاه و تکثیر آنها

همه ماتریال لازم از جمله شماره لاین‌ها، شماره ایستگاه‌ها، فهرست اسامی دانشجویان و تقسیم‌بندی آنها، راهنمای آزمونگر، راهنمای بیمارنما، راهنمای دانشجو و چک‌لیست باید چاپ و تکثیر شوند اما خیلی مهم است که قبل از این کار، برای بار آخر مرور نهایی صوت گیرد. از آنجا که معمولاً ایستگاه‌ها توسط افراد مختلف طراحی می‌شود، مهم است که ساختار و شکل تمام آنها به خصوص در قسمت نمره‌دهی یکسان شود.

در مواردی که OSCE به صورت مبتنی بر شبکه برگزار می‌شود، ماتریالی پرینت گرفته نخواهد شد اما اطمنیان از اینکه همه موارد به درستی در نرم‌افزار وارد شده است، اهمیت دارد.

یادآوری تاریخ و ساعت آزمون به همه افراد دست اندرکار

از آنجا که حضور به موقع تک‌تک دست‌اندرکاران برای برگزاری موفقیت‌آمیز OSCE ضروری است و از طرفی می‌توان تصور کرد که این افراد تا چه میزان درگیر سایر امور کاری هستند، حتماً لازم است که در هفته منجر به آزمون و حتی مجدداً روز قبل آزمون به تمام افراد یادآوری شود که رأس ساعت در محل برگزاری حاضر باشند. به منظور رعایت جانب احتیاط اعلام زمان زودتر از ساعت مقرر توصیه می‌شود. برای کنترل موارد پیش‌بینی‌نشده، در نظر گرفتن چند آزمونگر و بیمارنمای جایگزین و ذخیره باعث می‌شود استرس برگزارکنندگان در روز آزمون کاهش باید.

طرح‌ریزی نقشه آزمون و پیاده‌سازی آن

قبل از اینکه عملاً ایستگاه‌ها چیده شوند، باید ایستگاه‌ها شماره‌گذاری شوند، نقشه محل آزمون ترسیم شود و یک بار تمام جزئیات از جمله محل ایستگاه‌ها، میز ثبت نام دانشجویان، محل قرنطینه قبل و بعد آزمون، مسیر حرکت دانشجویان، محل استقرار بیمارنماها و آزمونگران و ... روی کاغذ مشخص شوند. پس از اینکه همه چیز روی کاغذ تثبیت شد، می‌توان نقشه را اجرا کرد، ایستگاه‌ها را سازمان‌دهی نمود و وسایل و تجهیزات مورد نظر را در آنها قرار داد.

شماره‌گذاری ایستگاه‌ها باید به صورت منطقی انجام شود و کاملاً در معرض دید باشد تا حتی دانشجویان بسیار مضطرب بتواند به راحتی و درستی راه خود را پیدا کنند. نشان دادن مسیر با فلش و راهنما روی کف سالن نیز کمک‌کننده است. دانشجویی که در مسیر خود به اشتباه وارد ایستگاه دیگری می‌شود، علاوه بر اینکه تمرکز خود را از دست می‌دهد، موجب به هم ریختگی در وضعیت سایر دانشجویان نیز می‌گردد.

برای اعلام پایان زمان ایستگاه‌ها معمولاً از زنگ استفاده می‌شود. یک نفر از پرسنل باید مسؤولیت این کار را بر عهده بگیرد. وسیله‌ای برای نگه داشتن زمان باید در اختیار وی قرار داده شود. صدای زنگ باید در تمام ایستگاه‌ها با در بسته به خوبی شنیده شود. برای اطمینان از این امر، صدای زنگ باید در سر و صدای یک روز عادی و معمولی از قبل چک شود. گاهی از تکنولوژی‌های پیشرفته‌تر استفاده می‌شود که طبق زمان‌بندی مشخص به صورت خودکار پایان وقت را اعلام می‌کند. در این حالت، می‌توان به گونه‌ای به کامپیوتر برنامه داد که یک دقیقه مانده به اتمام زمان ایستگاه یا که زمانی آزمونگر باید بازخورد را شروع کند (در آزمون تکوینی)، نیز وقت را اعلام کند.

در اغلب امتحانات برای حفظ امنیت آزمون، مسائلی از قبیل کور کردن آنتن موبایل مدنظر قرار می‌گیرد.

بعد از اینکه همه چیز آماده شد، چیدمان و آرایش فضای آزمون باید مرور شود. پایلوت کردن یا اجرای امتحانی قبل از اجرای اصلی برای روشن کردن و برطرف کردن نقائص کار بسیار کمک‌کننده است. به این منظور می‌توان از افرادی که

در روند چیدمان فضای آزمون دخیل نبوده‌اند و قرار هم نیست در امتحان شرکت کنند، مانند عضو هیأت علمی یک بخش دیگر، یا دانشجوی سال بالاتر یا یک دانش‌آموخته درخواست کرد دقیقاً مانند روز آزمون با صدای زنگ از یک ایستگاه به ایستگاه دیگر حرکت کند. اگر این افراد که آزمودنی‌های تمرینی[1] نامیده می‌شوند، در پیدا کردن مسیر به مشکل برخورد نکردند، احتمالاً دانشجویان نیز مشکل جدی نخواهند داشت. این کار علاوه بر کنترل کردن مسیر حرکت دانشجو، به شناسایی مشکلات دیگر از جمله میزان رسا بودن صدای زنگ نیز کمک می‌کند. گاهی برای کنترل مناسب بودن زمان ایستگاه و کافی بودن امکانات موجود، از این افراد خواسته می‌شود تا مانند آزمون واقعی واقعاً وظایف محوله در هر ایستگاه را انجام دهند. به این ترتیب می‌توان متوجه شد که آیا مثلاً وظیفه خواسته شده در طی زمان در نظرگرفته شده قابل انجام است یا اینکه مثلاً چک‌لیست تدوین شده نقص دارد یا اینکه همه چیز مرتب است. به این کار که همانند پایلوت کردن است، اجرای نمایشی[2] می‌گویند (سلبی و همکاران ۱۹۹۵، کیسی و همکاران ۲۰۰۹).

برنامه‌ریزی برای ثبت مستندات و وقایع

تجربه برگزاری OSCE ارزشمند است و نباید به آسانی از دست برود. مکتوب‌سازی و نگهداری تمام برنامه‌ریزی‌های صورت گرفته از جمله جدول زمان‌بندی مراحل، ساختار OSCE و چیدمان ایستگاه‌ها، سناریوها و چک‌لیست‌ها، اسامی و مشخصات آزمونگران و بیمارنمایان، تعداد و مشخصات دانشجویان و ... علاوه بر اینکه در اجراهای بعدی آزمون کمک‌کننده است، اطلاعات مورد نیاز برای دو مرحله بعد یعنی تحلیل آزمون و تهیه بانک سؤال را نیز فراهم می‌کند. همچنین ثبت رخدادهای ناخواسته و تصمیماتی که متعاقب آنها اخذ شده است، تجربه برگزارکنندگان را به مسؤولان آتی منتقل می‌سازد.

تدوین چارچوب ارزشیابی و تحلیل آزمون

مانند هر آزمون دیگری لازم است که پس از اجرا، نتایج آزمون برای بررسی کیفیت سؤالات مورد تحلیل قرار گیرند. جزئیات مربوط به نحوه محاسبه و تفسیر شاخص‌ها در بخش تحلیل آزمون (بخش هشتم) مفصلاً مورد بحث قرار می‌گیرند اما صرفاً برای نام بردن باید گفت کمیته OSCE باید به فکر استخراج داده‌ها در موارد زیر باشد:

- روایی آزمون: سنجش روایی آزمون می‌تواند از طریق نظرخواهی از اعضای هیأت علمی در مورد اهمیت ایستگاه و ارتباط آن با اهداف آموزشی دوره صورت گیرد. همچنین نظر شرکت‌کنندگان در آزمون نیز می‌تواند در تعیین روایی آزمون کمک‌کننده باشد.

چند سؤال مرتبط با روایی آزمون در نظرخواهی از اعضای هیأت علمی
آیا سطح ایستگاه با سطح دستیاران سال دوم داخلی مطابقت داشت؟
آیا این ایستگاه قائل به افتراق بین دستیاران خوب و بد بود؟
آیا این ایستگاه شبیه شرایط واقعی کار دستیاران بود؟
اگر دستیاری نمره حدنصاب ایستگاه را کسب کند، آیا در مواجه با بیمار نیز توانمندی مورد نظر را داراست؟
آیا این ایستگاه همان چیزی را می‌سنجید که در بلوپرینت دوره مقرر بود بسنجد؟

چند سؤال مرتبط با روایی آزمون در نظرخواهی از دانشجویان
آیا وظیفه خواسته شده در ۵ دقیقه قابل اجرا بود؟
آیا موضوع ایستگاه در اهداف دوره آموزشی وجود داشت؟
آیا موضوع ایستگاه به شما تدریس شده بود؟
آیا موضوع ایستگاه با کار روزمره شما ارتباط داشت؟
آیا عملکرد بیمارنما قابل باور بود؟

1. Mock Examinee
2. Dummy Run

- پایایی آزمون و ایستگاه‌ها: سنجش پایایی بر اساس نظریه کلاسیک یا نظریه تعمیم‌پذیری قابل انجام است که می‌تواند شامل این موارد باشد: آلفای کرونباخ در سطح آیتم‌ها، آلفای کرونباخ در سطح ایستگاه‌ها، آلفا در صورت حذف آیتم یا ایستگاه، میزان توافق آزمونگران، ضریب تعمیم‌پذیری مطلق یا نسبی
- میزان قبولی و ردی در آزمون و در هر ایستگاه
- شاخص دشواری و تمیز ایستگاه طبق نظریه کلاسیک یا نظریه سؤال پاسخ
- مجذور رابطه بین نمره چک‌لیست با نمره‌دهی گلوبال هر ایستگاه
- رضایت آزمونگرن، بیمارنمایان و دانشجویان
- هزینه صرف شده

تهیه بانک ایستگاه

ایستگاه‌ها با کمی تغییرات و اصلاحات که مبتنی بر نتایج ارزشیابی و بازخوردهای امتحان قبلی بوده است، مجدداً قابل استفاده هستند. با توجه به زمانی که صرف طراحی ایستگاه‌ها شده، معقول نیست که کل فرایند طراحی برای هر آزمون مجدداً از ابتدا انجام شود. ایستگاه‌هایی که طراحی می‌شوند، باید همراه با چک‌لیست و راهنماهای مربوطه در بانک OSCE ثبت و نگهداری شوند تا برای آزمون‌های بعدی مورد استفاده قرار گیرند. نکته حائز اهمیت برای قرار دادن یک ایستگاه در بانک این است که حتماً فرایند مرور توسط متخصصان صورت گیرد و خطاهای طراحی ایستگاه حذف و ویرایش شوند. همچنین محاسبه خصوصیات سایکومتریک ایستگاه از جمله ضریب دشواری و تمیز، اطلاعات مفیدی در مورد کیفیت ایستگاه در اختیار می‌گذارد (خان و همکاران ۲۰۱۳).

برخی از دانشکده‌ها ایستگاه‌ها را فقط از بانک انتخاب می‌کنند زیرا معتقد هستند که هر ایستگاهی قبل از اجرا باید یک بار پایلوت شود تا بتوان از کیفیت آن مطمئن بود. گاهی این پایلوت کردن در خلال یک آزمون تکوینی یا آزمون نه چندان سطح بالا صورت می‌گیرد. یعنی بعد از اجرای آزمون تکوینی، اطلاعات مربوط به کیفیت ایستگاه مورد نظر استخراج می‌شود تا در بانک قرار گیرد و برای آزمون‌های اصلی استفاده گردد. گاهی نیز پایلوت ایستگاه جدید در جریان یک آزمون سطح بالا و مهم صورت می‌گیرد اما چون از کیفیت آن مطمئن نیستند، نمره آن را برای دانشجویان منظور نمی‌کنند و فقط از اطلاعات آن برای آزمون‌های بعدی استفاده می‌نمایند.

تدوین چارچوب ارائه گزارش

اگر قرار است تجربه برگزاری OSCE گزارش شود، رعایت یک سری اصول در تدوین آن ضروری است. نویسندگان یک مطالعه مروری در روند بررسی مقالات و گزارش‌های منتشر شده درباره OSCE به این نتیجه رسیدند که فقدان ساختار منسجم، عدم ارائه اطلاعات کافی و ناهمگونی در استفاده از لغات و عبارات از جمله ضعف‌هایی است که کار را برای خوانندگان دشوار می‌کنند (پاتریشیو و همکاران ۲۰۰۹). آنها یک چک‌لیست جامع به منظور ارائه گزارش OSCE ارائه دادند (جدول ۸–۱۸) که در سه بخش تنظیم شده است:

- اطلاعات مربوط به دوره آموزشی
- اطلاعات مربوط به طراحی OSCE
- اطلاعات مربوط به نتایج OSCE

گزارش تمام موارد ذکر شده در این چک‌لیست برای تمام OSCEها امکان‌پذیر نیست و برگزارکنندگان باید بسته به هدف و ماهیت آزمون از بین آنها انتخاب کنند. توجه بیشتر مخصوصاً هنگامی بسیار مهم است که هدف از ارائه اطلاعات، ارزشیابی و تحلیل خود OSCE است و نه صرفاً ارزیابی دانشجویان و یا برنامه آموزشی.

جدول ۸-۱۸: چک‌لیست جامع به منظور گزارش OSCE

۱. اطلاعات مربوط به دوره آموزشی	
دانشگاه/دانشکده	نام دانشگاه و دانشکده‌ای که مسؤول طراحی و برگزاری آزمون است.
رشته	پزشکی، پرستاری، دندانپزشکی، ...
تعداد سال دوره	هفت سال، شش سال، ...
مقطع	ارشد، دکترای عمومی، دستیاری، ...
سال	سوم، چهارم، ...
تعداد دانشجویان	تعداد دانشجویانی که در دوره و در آزمون شرکت کرده‌اند
اهمیت آزمون	سطح بالا، متوسط، پایین
۲. اطلاعات مربوط به طراحی OSCE	
هدف امتحان	برگزاری OSCE در فرمت جدید، ارزشیابی برنامه آموزشی و روش‌های تدریس، ارزیابی دانشجوی ان، مقایسه ارزیابی چک‌لیست با گلوبال، سنجش روایی و پایایی و قابلیت اجرا، مقایسه OSCE با سایر امتحانات، رابطه OSCE با متغیرهایی مانند سن و جنس، ...
نقش امتحان	تکوینی، تراکمی، هردو، پایلوت، فقط برای بازخورد به مدرسان، فقط برای آموزش مدرسان، ...
تعداد دانشجویان	تعداد دانشجویانی که در آزمون شرکت کرده‌اند.
تعداد دانشجویان مورد مطالعه	تعداد دانشجویانی که برای مطالعه انتخاب شده‌اند و دلایل ورود آنها به مطالعه
تعداد چرخش‌ها	در مواردی که تعداد دانشجویان زیاد است، یک امتحان در چندین چرخش باید برگزار شود
جلسه توجیهی دانشجویان	نحوه آشناسازی دانشجویان با ساختار و روند آزمون
تعداد محل‌های برگزاری	اگر یک آزمون در چند مکان برگزار می‌شود.
تعداد و زمان امتحانات	اگر گزارش چند سال ارائه می‌شود.
تعداد لاین‌های موازی یا مشابه	در صورتی که به دلیل تعداد زیاد دانشجویان آزمون در لاین‌های موازی یا مشابه برگزار شده است.
تعداد دانشجویان هر لاین	در صورتی که به دلیل تعداد زیاد دانشجویان آزمون موازی برگزار شده است، تعداد دانشجویانی که در هر لاین در آزمون شرکت کرده‌اند.
آزمون پایلوت	مشخصات ایستگاه‌هایی که برای پایلوت قبل از آزمون اصلی برگزار شده‌اند.
موضوع و محتوای امتحان	طب داخلی، زنان، بیهوشی، جراحی، پاتولوژی، اطفال، پزشکی اجتماعی، ...
نحوه ارزیابی و ثبت آن	چک‌لیست یا گلوبال، کاغذی یا الکترونیک
نحوه تصمیم‌گیری برای رد و قبول	روش تعیین استاندارد و مدل نمره‌دهی
تعداد ایستگاه‌ها	
تعداد ایستگاه‌های مورد مطالعه	اگر فقط برخی از ایستگاه‌ها برای مطالعه انتخاب شده‌اند، تعداد و دلیل آن.
نوع ایستگاه‌ها	پروسیجر، کتبی، ترکیبی، استراحت

ادامه جدول ۸-۱۸: چک‌لیست جامع به منظور گزارش OSCE

۲. اطلاعات مربوط به طراحی OSCE	
هدف مورد ارزیابی در ایستگاه	شرح حال، معاینه، تشخیص، مهارت ارتباطی، پروسیجر، تجویز دارو، نسخه‌نویسی، آموزش بیمار، اخلاق، فناوری اطلاعات، تفسیر داده‌ها، حل مسأله
زمان ایستگاه	اگر زمان ایستگاه‌ها نامساوی است، حتما ذکر شود
ضبط	تعداد و جزئیات ایستگاه‌هایی که ضبط شده‌اند
تعداد و مشخصات طراحان ایستگاه‌ها	کسانی که ایستگاه، ستاریو و چک‌لیست را تدوین کرده‌اند.
تعداد و مشخصات مشاهده‌گران	کسانی که روز آزمون در جلسه امتحان یا در ایستگاه‌های unmanned حضور داشته‌اند (نه آزمونگران).
تعداد و مشخصات آزمونگران	کسانی که عملکرد دانشجویان را در ایستگاه ارزیابی کرده‌اند
تعداد و مشخصات بیمارنماها	
تعداد و مشخصات بیماران واقعی	
نحوه آموزش آزمونگران و بیمارنماها	مدت زمان، مطالب آموزش داده شده، نحوه ارائه
شواهد روایی OSCE	بلوپرینت آزمون، تعداد و مشخصات کسانی که درگیر طراحی و برنامه‌ریزی آزمون بوده‌اند، نحوه مشارکت کسانی که به طراحی ایستگاه‌ها پداخته‌اند (جلسه غیررسمی، جلسه متمرکز[1]، کارگاه، پرسشنامه، تکنیک دلفی ...)، شواهد پشتیبان در متون مربوط به OSCE روایی صوری، روایی محتوایی، روایی سازه، روایی پیش‌بینی کننده، روایی همزمان
شواهد پایایی OSCE	پایایی بین آزمونگران: مانند ضریب کاپا ثبات درونی: مانند ضریب آلفا ضریب تعمیم‌پذیری
شواهد قابلیت اجرای OSCE	هزینه و زمان صرف شده برای برنامه‌ریزی، طراحی و اجرای OSCE
بازخورد به دانشجویان و بیمارنماها	توسط چه کسانی، چه وقت، چگونه، چه مطالبی
۳. اطلاعات مربوط به نتایج OSCE	
نتایج OSCE	باید بر اساس هدف آزمون گزارش شوند
بازخورد در مورد OSCE	اطلاعات حاصل از بازخورد دانشجویان، استادان، آزمونگران، بیمارنماها و ... در مورد آزمون.
ارتباط موضوعی[2]	آیا آزمون از دید دانشجویان، استادان، بیماران و ... مرتبط بوده است؟
عدالت[3]	آیا آزمون از دید دانشجویان با توجه به مطالبی که در دوره به آنها آموزش داده شده است، عادلانه بوده است؟
تاثیر روی یادگیری	از دید دانشجویان و استادان OSCE چگونه باعث یادگیری دانشجویان شده است؟
تاثیر روی آموزش	از دید استادان آیا OSCE نقاط قوت و ضعف دانشجویان را آشکار کرده است؟ آیا بر اساس این اطلاعات، مباحث آموزشی یا نحوه ارائه آنها تغییر کرده‌اند؟
مشکلات و دشواری‌های OSCE	اطلاعات در مورد مشکلات مربوط به امتحان و در صورت امکان، راه‌حل‌های آنها

1. Focus group
2. Relevance
3. Fairness

طراحی ایستگاه

همان‌طور که در قسمت قبل گفته شد، یکی از مراحل مهم طراحی OSCE، طراحی ایستگاه‌های آن است. یعنی پس از اینکه بلوپرینت دوره تدوین شد، تعداد ایستگاه‌ها و زمان آنها مشخص شد و بر سر موضوع ایستگاه‌ها توافق حاصل شد، زمان آن است که محتوای هر ایستگاه طراحی شود.

معمول است که مسؤول برگزاری OSCE، طراحی هر ایستگاه را به یکی از اعضای هیأت علمی که تخصص محتوایی دارد، واگذار کند. اما همان‌طور که قبلاً اشاره شد مسؤول برگزاری OSCE نمی‌تواند مسؤولیت طراحی ایستگاه‌ها را به صورت کامل به اعضای هیأت علمی بسپارد بلکه باید اصول کار را برای آنها تشریح کند، با آنها ارتباط مستمر داشته باشد، بر روند انجام کار نظارت کند و پس از دریافت ایستگاه‌ها به مرور و ارزیابی آنها بپردازد و به طراحان بازخورد دهد.

طراح سؤال باید مطابق گام‌هایی که اشاره خواهد شد، به آماده کردن ایستگاه بپردازد اما قبل از آغاز طراحی، باید به این نکته توجه شود که محتوای ایستگاه بر اساس مهارت مورد ارزیابی، احتمالاً یکی از سه حالت زیر را داراست. اهمیت این تقسیم‌بندی از آن جهت است که گام‌های طراحی آنها کمی با یکدیگر متفاوت هستند. در زیر انواع ایستگاه توضیح داده می‌شود. خلاصه ویژگی آنها و مزایا و محدودیت‌های هر یک از انواع در جدول ۹–۱۸ آمده است.

جدول ۹–۱۸: انواع ایستگاه های مورد استفاده در OSCE و ویژگی های آنها

نوع ایستگاه	توضیح	مثال	مزایا	محدودیت‌ها
ایستگاه مشاهده‌ای	آزمونگر طی آزمون حاضر است و عملکرد دانشجو را مشاهده و ارزیابی می‌کند.	مهارت‌های تباطی، پروسیجر، مهارت‌بالینی	مشاهده مستقیم، امکان ارزیابی سطوح بالاتر یادگیری، امکان بازخورد فوری	زمان‌بر برای آزمونگر
ایستگاه غیرمشاهده‌ای	آزمونگر در ایستگاه حضور ندارد. پاسخ روی برگه نوشته و تحویل داده می‌شود.	نسخه‌نویسی، مشاهده نمونه‌پاتولوژی، تفسیر رادیولوژی، تکنولوژی اطلاعات	عدم نیاز به آزمونگر در ایستگاه	نبود مشاهده مستقیم، عدم ضرورت برگزاری OSCE و امکان ارزیابی این مهارت‌ها توسط سایر ابزارها
ایستگاه ترکیبی	دو ایستگاه متوالی که بر اساس یک سناریو طراحی شده‌اند. می‌تواند مشاهده‌ای یا غیرمشاهده‌ای باشد.	مشاهده معاینه ریه در ایستگاه اول و جمع‌بندی یافته‌ها و برنامه‌ریزی درمانی در ایستگاه دوم	امکان ارزیابی مهارت‌های بیشتر با تعداد سناریوهای محدود استفاده بهینه از آزمونگران	
ایستگاه مبتنی بر فناوری پیشرفته	از مولاژ و مانکن‌های تکنولوژیک و شبیه‌سازها استفاده می‌شود	معاینه فیزیکی خاص مانند معاینه رکتوم، تصمیم‌گیری بالینی در شرایط پیچیده و بیمار بدحال و ناپایدار	افزایش حیطه‌های مورد ارزیابی در OSCE و تنوع ایستگاه‌ها	هزینه خریداری و نگهداری، آموزش پرسنل برای استفاده

ایستگاه مشاهده‌ای: در ایستگاه «مشاهده‌ای»، تعامل دانشجو با بیمار مورد مشاهده قرار می‌گیرد و عملکرد دانشجو در مواجهه با بیمار ارزیابی می‌شود. طراح سؤال باید بیماری را در نظر بگیرد که به پزشک مراجعه کرده است و سناریویی بنویسد که بعداً بیمارنما همان را اجرا کند:

- در برخی از موارد، هدف ایستگاه گرفتن «شرح‌حال» بیمار است. در این صورت از بیمارنمایی استفاده می‌شود که شرح‌حال یک بیمار فرضی را بازگو می‌کند. به عنوان مثال، اگر موضوع ایستگاه، گرفتن شرح حال از بیمار مبتلا به

سردرد باشد، بیمارنما تظاهر می‌کند که در چند روز گذشته سردرد داشته است و بر اساس آنچه در متن سناریو نوشته شده است، پاسخ سؤالات دانشجو را در مورد شدت و مدت و کیفیت آن می‌دهد.

- گاهی ارزیابی «مهارت ارتباطی» پزشک با بیمار در ایستگاه مد نظر است. در این حالت نیز سناریویی نوشته می‌شود و بر اساس آن بیمار به پزشک مراجعه می‌کند و نحوه برقراری رابطه پزشک با وی مورد ارزیابی قرار می‌گیرد.
- هدف برخی از ایستگاه‌ها، ارزیابی مهارت دانشجو در «معاینه فیزیکی» است که به این منظور گاهی از مانکن و مولاژ استفاده می‌شود (مانند معاینه واژینال در خانم باردار) و گاهی از بیمارنما (مانند معاینه شکم). در حالت دوم، اگر لازم است بیمارنما واکنش خاصی انجام دهد، مثلاً در پاسخ به لمس عمقی شکم، به نشانه درد چهره خود را در هم کشد، ضروری است که طراح سؤال آن را در راهنمای مربوطه ذکر کند و آموزش لازم نیز به بیمارنما داده شود. در صورتی که یافتن یک «علامت بالینی»[1] مدنظر باشد، از بیمار واقعی (مانند سوفل قلبی) یا بیمارنماها می‌توان استفاده کرد. در حالت دوم، بیمارنما باید علامت بالینی را تقلید نماید. این موارد با اینکه معدود هستند اما امکان‌پذیر می‌باشند. به عنوان مثال، در معاینه نورولوژیک، بیمارنما وانمود می‌کند که نمی‌تواند با چشم بسته درست بایستد یا راه برود.
- گاهی در یک ایستگاه، آزمون اجرای یک «پروسیجر» مدنظر است که ممکن است روی مانکن و مولاژ صورت گیرد (مانند بخیه زدن) و یا روی بیمارنما (مانند بستن آتل). امروزه این دو در کنار یکدیگر امکان شبیه‌سازی هیبرید را نیز فراهم کرده‌اند. به عنوان مثال، می‌توان به مانکنی برای گرفتن نمونه خون وریدی اشاره کرد که به شکل بازو است و در کنار بدن بیمارنما قرار می‌گیرد یا به بدن او وصل می‌شود (شکل ۳-۱۸). مزیت این حال این است که نه تنها جنبه‌های تکنیکی پروسیجر ارزیابی می‌شود، بلکه ابعاد انسانی قضیه مانند نحوه برقراری رابطه با بیمار و گرفتن اجازه برای شروع کار نیز مورد توجه قرار می‌گیرد. مثال دیگر شبیه‌سازی هیبرید، استفاده از تاتوهای موقت مانند تاتوی ملانوم بدخیم است که به عنوان ضایعه پوستی روی بدن بیمارنما نصب می‌شود.

شکل ۳-۱۸: شبیه‌سازی هیبرید

1. Relevant

ایستگاه غیرمشاهده‌ای: نوع دیگر ایستگاه‌ها «ایستگاه‌های غیر مشاهده‌ای» هستند. هدف این قبیل ایستگاه‌ها، ارزیابی یک مهارت عملی نیست. بلکه سؤالی مطرح می‌شود که دانشجو باید به آن پاسخ دهد یا داده‌های آزمایشگاهی و تصاویر رادیولوژیک در اختیار او قرار می‌گیرد تا آنها را تفسیر کند. در این حالت، نیازی به حضور بیمار و آزمونگر در ایستگاه نیست؛ دانشجو پاسخ خود را در پاسخنامه‌ای می‌نویسد که بعداً تصحیح می‌شود. اصطلاحاً به این ایستگاه‌ها unmanned هم گفته می‌شود.

ایستگاه ترکیبی: برخی از ایستگاه‌ها ترکیبی از دو حالت فوق (مشاهده‌ای و غیر مشاهده‌ای) هستند. یعنی دانشجو باید ابتدا در مواجهه با بیمارنما وظایف محوله را انجام دهد و بیمارنما یا آزمونگر حاضر در ایستگاه او را توسط چک‌لیست ارزیابی می‌کند. سپس وقت دارد (پنج تا هفت دقیقه) تا به یک سری از پرسش‌ها به صورت کتبی پاسخ دهد. ممکن است لزوماً پرسشی در کار نباشد و بنا به موضوع ایستگاه از دانشجو خواسته شود خلاصه‌ای از شرح‌حالی که گرفته است بنویسد یا با جمع‌بندی مشکلات بیمار، فهرستی از تشخیص‌های افتراقی تهیه کند یا برنامه تشخیصی–درمانی خود را ارائه دهد. این نوشته‌ها بعداً در اختیار اعضای هیأت علمی مجرب قرار می‌گیرد و تصحیح می‌شود. در برخی از موارد این دو مرحله از یکدیگر تفکیک می‌شوند و در دو ایستگاه پشت هم سنجیده می‌شوند. یعنی دانشجو در ایستگاه اول در مواجهه با بیمار کاری را انجام می‌دهد و سپس وارد ایستگاه بعدی می‌شود که غیرمشاهده‌ای است و دانشجو باید نوت‌های مربوط به بیمار را تدوین کند. به این رویکرد، «ایستگاه بعد از مواجهه» می‌گویند. در این حالت، نحوه حرکت دانشجویان در مسیر ایستگاه‌ها باید به دقت برنامه‌ریزی شود. زیرا نقطه شروع هیچ دانشجویی نمی‌تواند از ایستگاه دوم باشد.

ایستگاه‌های مبتنی بر فناوری پیشرفته[1]: در این ایستگاه‌ها از مانکن، مولاژ و شبیه‌سازهای تکنولوژیک استفاده می‌شود تا مواردی که در حالت معمول نمی‌توان ارزیابی کرد، در OSCE راحت‌تر مورد سنجش قرار گیرند.

با در نظر داشتن مسائل فوق، اکنون به ارائه اصول طراحی یک ایستگاه و مراحلی که برای اینکار باید مورد توجه قرار گیرد، می‌پردازیم. چارچوبی که طراحی ایستگاه را نظم می‌بخشد، به صورت خلاصه در جدول ۱۰–۱۸ آمده است.

جدول ۱۰–۱۸: اجزا و چارچوب طراحی ایستگاه

شناسنامه ایستگاه	
موضوع ایستگاه	از قبل توسط کمیته OSCE مشخص شده است.
تاریخ طراحی	
نام طراح	
سطح دانشجویان	از قبل توسط کمیته OSCE مشخص شده است.
توانمندی(های) مورد ارزیابی	از قبل توسط کمیته OSCE مشخص شده است.
زمان ایستگاه	از قبل توسط کمیته OSCE مشخص شده است.
نوع مشکل	مزمن، حاد یا تحت حاد
خلاصه سناریو	

1. High fidelity

ادامه جدول ۱۰–۱۸: اجزا و چارچوب طراحی ایستگاه

اطلاعات مورد نیاز برای کمیته اجرایی	
مشخصات بیمارنما	مواردی مانند سن و جنس و ویژگی‌های مربوط به بیماری که توسط طراح ایستگاه مشخص می‌شود تا کمیته اجرایی بتواند فرد مناسبی برای ایستگاه پیدا کند.
امکانات مورد نیاز	وسایل مورد نظر توسط طراح ایستگاه مشخص می‌شود تا توسط کمیته اجرایی مهیا شود. همچنین اگر امتحان در چند لاین یا چند مکان برگزار می‌شود، برای استانداردسازی آزمون لازم است.
چیدمان ایستگاه	نحوه چیدمان وسایل در ایستگاه (به عنوان مثال، محل قرار گرفتن و میز صندلی دانشجو و بیمارنما و آزمونگر) توسط طراح ایستگاه مشخص می‌شود تا کمیته اجرایی بتواند ایستگاه را به درستی پیاده‌سازی کند.
راهنمای دانشجو	
خلاصه سناریو	اطلاعات اصلی و کلیدی در مورد ایستگاه باید همان ابتدا در اختیار دانشجو قرار گیرد.
مکان سناریو	اینکه دانشجو در کجا (اورژانس، درمانگاه، بخش، ...) قرار گرفته است، بر رویکرد او در مواجهه با بیمار می‌تواند تأثیرگذار باشد.
گزارش اقدامات قبلی	اگر قبل از مواجهه دانشجو با بیمار، اقداماتی روی بیمار صورت گرفته است، مثلاً فشار خون او توسط پرستار چک شده است، باید ذکر شود. در این صورت، شاید طراح سؤال، انتظار داشته باشد دانشجو با توجه به مورد ذکر شده، درباره اندازه فشار خون بیمار از پرستار سؤال کند.
کاری که انتظار می‌رود دانشجو انجام دهد.	مثلاً: «روی بیمارنما معاینه کامل ریه انجام دهید.»
کاری که انتظار نمی‌رود دانشجو انجام دهد.	بهتر است با توجه به محدودیت وقت، کاری که انتظار نمی‌رود دانشجو انجام دهد، شفاف بیان شود تا دانشجو از وقت خود به درستی استفاده کند. مثلاً: «روی بیمارنما معاینه ریه انجام دهید. دق ریه لازم نیست.»
امکان تعامل یا درخواست اطلاعات بیشتر	معمولاً در راهنمای دانشجو نوشته می‌شود که مجاز به تعامل با آزمونگر نیست. اگر سناریو ایجاب می‌کند که دانشجو با آزمونگر تعامل داشته باشد، باید ذکر شود. مثلاً اینکه: «اگر به جواب آزمایش نیاز دارید، از آزمونگر سؤال کنید.»
راهنمای آزمونگر	
خلاصه سناریو	اطلاعات اصلی و کلیدی در مورد ایستگاه باید در اختیار آزمونگر قرار گیرد.
هدف ایستگاه و آنچه از دانشجو انتظار می‌رود انجام دهد.	مثلاً: «در این ایستگاه از دانشجو خواسته شده است معاینه کامل ریه انجام دهد.»
نقش آزمونگر	مهم است شفاف شود که آزمونگر چه کارهایی را باید یا نباید که انجام دهد. مثلاً اینکه می‌تواند با دانشجو صحبت کند یا خیر.
اطلاعاتی که باید در اختیار دانشجو قرار بگیرد	ممکن است طراح در نظر داشته باشد که آزمونگر جواب آزمایش خون بیمار را در اختیار دانشجو قرار دهد.
اطلاعاتی که نباید در اختیار دانشجو قرار بگیرد.	آزمونگر باید بداند محدوده اطلاعاتی که می‌تواند در اختیار دانشجو بگذارد، به چه میزانی است.
اطلاعات مرتبط بالینی	آزمونگر برای ارزیابی درست به یک سری اطلاعات مرتبط با سؤال نیاز دارد؛ مخصوصاً اگر از دیسیپلین مربوطه نباشد.

ادامه جدول ۱۰-۱۸: اجزا و چارچوب طراحی ایستگاه

راهنمای بیمارنما	
راهنمای بیمارنما	ویژگی‌ها و خصوصیات بیمارنما مانند سن، جنس، ویژگی فیزیکی و ... باید توضیح داده شود.
زمینه اجتماعی-اقتصادی آنها	
شکایت اصلی، شرح بیماری حاضر و سابقه بیماری قبلی	شرح بیماری تا حد امکان کامل ولی موجز و مرتبط با موضوع ایستگاه باشد.
جزئیات مربوط به احساسات و نگرانی آنها در مورد بیماری	شرح دقیق نگرانی‌ها و احساسات بیمار در استاندارد کردن عملکرد بیمارنماهای ایستگاه‌های مشابه اهمیت دارد.
آنچه باید انجام دهند (و ندهند) و آنچه باید بگویند (و نگویند).	شرح دقیق رفتار و گفتار بیمارنماها نقش مهمی در استانداردسازی عملکرد آنها دارد.
سؤالاتی که باید از دانشجو بپرسند.	مشابه قبلی

تدوین شناسنامه ایستگاه و طراحی سناریو

موضوع ایستگاه از قبل توسط کمیته OSCE مشخص شده است. اما بهتر است مسؤول OSCE یک بار دیگر آن را با جزئیات بیشتر با طراح ایستگاه چک کند. به این منظور داشتن چارچوبی که در آن مشخصات مربوط به ایستگاه ذکر شده است، کمک‌کننده است. بهتر است که برای تمام ایستگاه‌ها شناسنامه استانداردی در نظر گرفته شود و از ابتدا همان در اختیار طراحان سؤال قرار گیرد تا آن را تکمیل کنند.

برای طراحی سؤال، طراح باید مطابق با مهارت مورد ارزیابی در ایستگاه، یک سناریوی بالینی آماده کند. این سناریو لزوماً در اختیار دانشجو قرار نمی‌گیرد و بیشتر برای مهیا کردن سایر متریال ایستگاه مانند چک‌لیست و راهنماها به کار می‌آید. اما مکتوب کردن آن در شکل دادن و عینیت بخشیدن به افکار طراح سؤال کمک می‌کند. مواردی مانند نام، سن، جنس، محل مراجعه و شکایت اصلی بیمار معمولاً باید در سناریو مشخص شوند.

سناریو یک ایستگاه	
موضوع ایستگاه: شرح حال سردرد نام طراح: ..... زمان ایستگاه: ۸ دقیقه نوع مشکل: ☐مزمن ■ حاد ☐ تحت حاد	تاریخ طراحی: ..... سطح دانشجویان: سال چهارم پزشکی توانمندی مورد ارزیابی: برقراری ارتباط و گرفتن شرح حال
سناریو: خانم ۳۰ ساله با سردرد و تهوع شدید به اورژانس مراجعه کرده است و سابقه سردردهای مشابه را از سه ماه پیش می‌دهد. دانشجو با رعایت اصول ارتباطی باید شرح حال متمرکز و مرتبط با شکایت بیمار بگیرد.	
سناریو یک ایستگاه	
موضوع ایستگاه: مشاوره جلوگیری از بارداری نام طراح: ..... زمان ایستگاه: ۵ دقیقه نوع مشکل: ☐ مزمن ☐ حاد ■ تحت حاد	تاریخ طراحی: ..... سطح دانشجویان: پیش کارورزی توانمندی مورد ارزیابی: مشاوره با حفظ اصول رازداری[1]
سناریو: خانم جوان ۱۸ ساله برای دریافت نسخه قرص ضدبارداری به درمانگاه زنان مراجعه کرده است. دانشجو باید اطلاعاتی را جمع‌آوری کند تا اطمینان حاصل کند تجویز قرص مناسب است، در مورد رابطه جنسی ایمن مشاوره دهد و در خصوص ضرورت معاینه لگنی بحث کند.	

1. Confidentiality

باید توجه شود که فرایند طراحی ایستگاه خطی نیست و طراح می‌تواند در مراحل بعدی برای ویرایش به مرحله قبلی بازگردد. مثلاً پس از نوشتن نسخه اولیه‌ای از چک‌لیست یعنی مرحله سوم، می‌توان مجدداً برگشت و سناریو را کامل کرد.

تهیه فهرست امکانات مورد نیاز برای تیم اجرایی

طراح سؤال باید فهرست کاملی از آنچه در ایستگاه لازم است یا مورد استفاده قرار می‌گیرد، تهیه کند و آن را در اختیار برگزارکنندگان بگذارد تا آنها وسایل مورد نظر را تهیه کنند و در ایستگاه قرار دهند.

مهم است که نحوه چیدمان وسایل نیز مشخص شود. به عنوان مثال، نحوه قرارگیری بیمارنما و دانشجو باید طوری باشد که دید آزمونگر را مختل نسازد. یا مانکن باید به گونه‌ای روی تخت قرار گیرد که علاوه بر اینکه دانشجو برای انجام وظیفه محوله راحت باشد، آزمونگر نیز به خوبی قادر به مشاهده و ارزیابی نحوه عملکرد وی باشد.

همچنین طراح باید مشخصات بیمارنمای مورد نظر را بیان کند تا برگزارکنندگان فرد مناسبی برای ایستگاه پیدا کنند که نقش بیمار را بازی کند.

تدوین راهنمای دانشجو

هنگامی که دانشجو وارد ایستگاه می‌شود، یک برگه راهنما در اختیار او قرار می‌گیرد. گاهی این برگه بر روی در اتاق نصب می‌شود تا دانشجو هنگامی که بیرون از ایستگاه منتظر صدای زنگ و ورود به اتاق است، آن را مطالعه کند.

در راهنمای دانشجو، خلاصه سناریو به صورت مختصر اما کامل و شفاف آمده است و همچنین آنچه از دانشجو انتظار می‌رود در ایستگاه انجام دهد، ذکر شده است. برای استانداردسازی، مکان سناریو نیز باید به اطلاع دانشجو برسد. زیرا در بسیاری از شرایط، بسته به اینکه دانشجو در بخش یا درمانگاه یا اورژانس با بیمار مواجه شود، باید رویکرد متفاوتی را در پی بگیرد. از آنجا که معمولاً دانشجو در ایستگاه با آزمونگر تعاملی ندارد ولی دانشجویان ممکن است حین آزمون خواستار توضیحات بیشتر از آزمونگر شوند، این موضوع نیز در برگه راهنما باید شفاف شود. به مثال‌های زیر توجه کنید:

راهنمای دانشجو

بیمار خانم ۵۳ ساله مبتلا به سردرد است که به مطب شما مراجعه کرده است. از شما انتظار می‌رود در این ایستگاه طی ۵ دقیقه، یک شرح حال متمرکز و مرتبط از بیمار بگیرید. حین آزمون تعاملی با آزمونگر نخواهید داشت.

یا:

راهنمای دانشجو

بیمار کودک ۱۰ ساله با شکایت کبودی مکرر است که به درمانگاه آورده شده است. از شما انتظار می‌رود در این ایستگاه طی هفت دقیقه، معاینه متمرکز و مرتبط برای مشخص شدن علت کبودی بیمار انجام دهید. حین آزمون تعاملی با آزمونگر نخواهید داشت.

راهنمای دانشجو

بیمار، آقای ۶۵ ساله است که به دلیل به اختلال راه رفتن در بخش نورولوژی بستری است. از شما انتظار می‌رود در این ایستگاه طی پنج دقیقه، معاینه اعصاب کرانیال را روی بیمار انجام دهید. حین آزمون تعاملی با آزمونگر نخواهید داشت.

گاهی ایستگاه بیش از یک هدف آموزشی را ارزیابی می‌کند. در این صورت، باید برای دانشجو به طور شفاف بیان شود که چه کارهایی را باید انجام دهد:

راهنمای دانشجو

بیمار، خانم ۲۲ ساله مبتلا به آکنه است که به درمانگاه پوست مراجعه کرده است. از شما انتظار می‌رود در این ایستگاه طی هفت دقیقه، شرح حال متمرکزی از بیمار بگیرید و برای او نسخه بنویسید. حین آزمون تعاملی با آزمونگر نخواهید داشت.

با توجه به محدودیت زمان هر ایستگاه و برای اینکه دانشجو بهتر بتواند وقت خود را تنظیم کند، اگر نیاز نیست اقدام خاصی را انجام دهد، بهتر است که این موضوع نیز ذکر شود:

راهنمای دانشجو

بیمار خانم ۵۰ ساله است که با شکایت سرفه به مطب شما مراجعه کرده است. از شما انتظار می‌رود در این ایستگاه طی چهار دقیقه، ریه بیمارنما را معاینه کنید. دق ریه مورد نظر نیست. حین آزمون تعاملی با آزمونگر نخواهید داشت.

راهنمای دانشجو در ایستگاه‌های غیرمشاهده‌ای نیز تقریباً به همین شکل است:

راهنمای دانشجو

بیمار آقای ۷۰ ساله است که با درد قفسه سینه ناگهانی به اورژانس مراجعه کرده است. از شما انتظار می‌رود در این ایستگاه طی چهار دقیقه، نوار قلب بیمار را تفسیر کنید و به سؤالات به صورت کتبی جواب دهید.

گاهی سناریو به گونه‌ای نوشته می‌شود که قبل از ویزیت بیمار توسط دانشجو، یک سری اقدامات برای بیمار انجام شده است. دراین صورت، اقدامات باید ذکر شوند و اگر قرار است دانشجو نتیجه اقدامات را سؤال کند، این نکته باید در راهنمای وی ذکر شود:

راهنمای دانشجو

بیمار آقای ۲۵ ساله مبتلا به درد شکم است که به اورژانس مراجعه کرده است. پرستار قبل از حضور شما، فشار خون بیمار را اندازه‌گیری کرده است و نمونه خون بیمار را برای آزمایشگاه ارسال کرده است. از شما انتظار می‌رود در این ایستگاه طی ۱۰ دقیقه، یک شرح حال متمرکز و مرتبط از بیمار بگیرید، شکم بیمار را معاینه کنید و برنامه تشخیصی-درمانی خود را برای آزمونگر شرح دهید. می‌توانید فشار خون و نتایج آزمایش را از آزمونگر بپرسید.

تدوین ابزار ارزیابی (گلوبال، چک‌لیست، پاسخنامه)

هنگامی که پایه و سناریوی سؤال تقریباً آماده شد، طراح باید ابزاری را که اساس کار آزمونگر برای ارزیابی و نمره‌دهی است، تدوین کند. اگر ایستگاه غیر مشاهده‌ای باشد، برای آن باید پاسخنامه تدوین شود. در ایستگاه مشاهده‌ای همان طور که قبلاً ذکر شد، ارزیابی به دو صورت گلوبال و مبتنی بر چک‌لیست قابل انجام است. اینکه کدام یک استفاده شود، قبلاً در کمیته OSCE تعیین شده است.

تدوین پاسخنامه: اگر ایستگاه غیرمشاهده‌ای است، به جای چک‌لیست، پاسخنامه‌ای باید تهیه شود که در همان ایستگاه توسط دانشجو تکمیل می‌شود و بعدا تصحیح می‌شود. ساختار کلی آن شبیه پاسخنامه سؤال تشریحی یا کوته‌پاسخ است.

تدوین مقیاس گلوبال: در نمره‌دهی گلوبال از آزمونگر خواسته می‌شود تا با توجه به مهارت مورد سنجش در ایستگاه و همچنین سطح و مقطع دانشجویان شرکت‌کننده در آزمون، نظر خود را به صورت کلی در مورد میزان توانمندی دانشجو

در حیطه مورد نظر اظهار کند. در اینجا نه تنها این مسأله که مهارت انجام شده است، مورد ارزیابی قرار می‌گیرد، بلکه اینکه به خوبی انجام شده است یا خیر نیز، مورد بررسی قرار می‌گیرد. بنابراین در مواردی که کیفیت کار مهم است، بهتر از چک‌لیست عمل می‌کند.نمره‌دهی گلوبال مخصوصاً برای ارزیابی همدردی، مهارت ارتباطی، قضاوت و نحوه سازمان‌دهی اطلاعات دانشجو مفید است (خان و همکاران ۲۰۱۳). به عنوان مثال:

در مجموع نظر شما در مورد عملکرد این دانشجو در این ایستگاه چیست؟ (فقط یک مورد را انتخاب کنید)
☐ عالی ☐ خوب ☐ قابل‌قبول ☐ مرزی ☐ رد

تدوین چک‌لیست: هرچند نمره‌دهی گلوبال اخیراً بیشتر مورد توجه قرار گرفته است، معمول‌تر است که برای ایستگاه چک‌لیست تدوین شود. در واقع، OSCE را به چک‌لیست‌های آن می‌شناسند. چک‌لیست فهرستی از کارهایی است که انتظار می‌رود دانشجو در ایستگاه انجام دهد. معمولاً به منظور تهیه چک‌لیست، در ابتدا کارهایی که دانشجو باید به ترتیب انجام دهد تا مهارت بالینی مورد نظر را نشان دهد، فهرست شود. سپس باید نکات مهمی که افتراق‌دهنده عملکرد دانشجوی ضعیف از قوی هستند، شناسایی شوند و به عنوان آیتم چک‌لیست نوشته شوند. تمرکز چک‌لیست باید بر همین موارد افتراق‌دهنده باشد. این نکته از آن جهت حائز اهمیت است که امتیاز اضافه به دانشجویی که همه سؤالات و اقدامات را امتحان می‌کند، داده نشود. به عنوان مثال اگر قرار باشد دانشجو برای بیمار مبتلا به آنمی آزمایش مشخصی درخواست کند اما چندین آزمایش بنویسد که آزمایش موردنظر هم یکی از آنها باشد، در واقع نمی‌توان به این دانشجو نمره کامل را داد.

طول چک‌لیست بستگی دارد به مهارت مورد نظر، زمان ایستگاه و کسی که عملکرد دانشجو را ارزیابی می‌کند. یک چک‌لیست برای ایستگاه پنج دقیقه‌ای شرح حال در صورتی که آزمونگر، هیأت علمی است، می‌تواند ۲۵ آیتم داشته باشد اما برای ارزیابی توسط بیمارنما تعداد آیتم‌ها باید کمتر باشد (اسمی ۲۰۰۳). در دستورالعمل سازمان پزشکی کانادا برای طراحی OSCE با ایستگاه پنج دقیقه‌ای ۸ تا ۲۵ آیتم توصیه شده است. در حالی که بر اساس نتایج یک مطالعه، در نظر گرفتن بیش از ۱۰ تا ۱۲ آیتم برای ایستگاه در هر حیطه‌ای اعم از شرح حال و معاینه فیزیکی و ...، روی روایی و پایایی اثر معکوس داشت (ویلکینسون و همکاران[1] ۲۰۰۳).

از آنجا که لزوماً آزمونگر حاضر در ایستگاه، همان طراح سؤال نیست، مهم است که چک‌لیست طوری تدوین شود که نیازمند تفسیر نباشد و برای آزمونگران شفاف و قابل درک باشد. توجه به نکات زیر هنگام طراحی چک‌لیست اهمیت دارد:

- آیتم‌های چک‌لیست باید مبتنی بر همان مهارتی باشند که در پایه سؤال از دانشجو خواسته شده است.
- هر آیتم باید بر اساس رفتار قابل مشاهده باشد (اسمی ۲۰۰۳). به عنوان مثال، در ایستگاهی که از دانشجو خواسته شده است گوش مانکن را با اتوسکوپ مشاهده کند، آیتم «پرده صماخ را مشاهده کرد»، قابل ارزیابی توسط آزمونگر نمی‌باشد و نباید در چک‌لیست منظور شود.
- هر فعالیت باید در قالب یک آیتم جدا نوشته شود و از به کار بردن «و» «یا» خودداری شود.
- هنگام تدوین چک‌لیست، طراح باید مراقب سطح دانشجویان شرکت کننده در امتحان باشد. ممکن است یک موضوع مشابه هم برای دانشجوی پزشکی و هم برای دستیار تخصصی در نظر گرفته شود. اما آنچه از دستیار انتظار می‌رود بتواند انجام دهد، حتماً در سطح دشواری بالاتری است. بنابراین، این موضوع باید هنگام تدوین چک‌لیست لحاظ شود.
- برای تدوین چک‌لیست باید به زمان ایستگاه دقت کرد و فعالیت‌ها را به زمان تعیین شده محدود کرد.

در AMEE guide شماره ۸۱ آمده است نمره‌دهی چک‌لیست به دو صورت دوحالته[2] و درجه‌ای[3] قابل تنظیم است:

1. Wilkinson et al.
2. Bianary
3. Rating scale

در OSCE که اولین بار توسط هاردن و همکاران اجرا شد، نمره‌دهی دوحالته به کار رفت. در این مدل، اگر دانشجو آیتم مورد نظر را انجام دهد، یک نمره کسب می‌کند و اگر انجام ندهد، صفر می‌گیرد. سپس نمرات کل آیتم‌ها یا یکدیگر جمع می‌شود.

باید دقت شود که گاهی آیتم‌های مختلف اهمیت یکسانی ندارند. به این منظور می‌توان وزن‌دهی کرد و بر اساس اهمیت نمرات بیشتری برای برخی از آیتم‌ها در نظر گرفت. در حالت معمول، بسته به این که دانشجو کار را به درستی انجام دهد، یا اصلاً انجام ندهد یا به صورت ناقص انجام دهد، نمره یک، صفر یا نیم به او تعلق می‌گیرد. ممکن است طراح تصمیم بگیرد که چون برخی از آیتم‌ها مهم‌تر هستند، وزن و نمره بیشتری به آنها اختصاص دهد. وزن‌دهی ممکن است میزان قبولی کلی آزمون را تغییر ندهد اما روایی چک‌لیست را بهبود می‌بخشد و ممکن است بر وضعیت رد یا قبول هر دانشجو اثر بگذارد (اسمی ۲۰۰۳).

برخی از منتقدان معتقد هستند که بعضی از مهارت‌های بالینی مراحلی دارند که انجام آنها بسیار ضروری است. به طوری که حتی وزن‌دهی نیز منعکس کننده اهمیت این آیتم‌ها نیست و تصمیم جدی‌تری برای افزایش میزان تمایز آنها مورد نیاز است که آن را اقدامات بحرانی[1] می‌نامند. پین و همکاران[2] در مطالعه خود چونگی استفاده از این روش و نتایج آن را تشریح می‌کنند.

پین و همکاران ۲۰۰۸

این مطالعه برای بررسی نمره‌دهی آیتم‌های چک‌لیست به شیوه اقدامات بحرانی روی OSCE هایی که در فاصله سال‌های ۲۰۰۳ تا ۲۰۰۶ برای ۳۹۸ دانشجوی دوره کارآموزی پزشکی دانشگاه ویرجینیا برگزار شده بود، انجام شد. پس از مرور مجدد آیتم‌ها و ایستگاه‌ها مجموعاً تشخیص داده شد که ۱۰ ایستگاه از ۲۵ ایستگاه حاوی آیتم‌های بحرانی هستند. دانشجویان بر اساس نمره کل OSCE به دو دسته قوی (با نمرات بالای میانه) و ضعیف (با نمرات کمتر از میانه) تقسیم شده بودند.
نتایج رگرسیون لاجیستیک نشان داد آیتم های بحرانی را بیشتر از دانشجویان اجرا کرده بودند. با این وجود ۶ تا ۴۶ درصد آنها نتوانسته بودند ۹ مورد از ۱۰ مورد را به درستی انجام دهند.
نویسندگان نتیجه‌گیری کردند که صرف گرفتن نمره بالا در کل آزمون نشان دهنده این نیست که دانشجویان تمام موارد ضروری را انجام داده‌اند و اعمال نمره‌دهی به شیوه بحرانی می‌تواند قدرت OSCE را در تمایز دانشجویان قوی از ضعیف افزایش دهد.

یکی از انتقاد دیگری که به نمره‌دهی دوحالته وارد می‌کنند، این است که در این مدل، کیفیت کار ارزیابی نمی‌شود و صرف انجام آن مهم است. سابقاً تصور می‌شد که این نحوه نمره دادن، عینیت ارزیابی را بسیار بالا می‌برد و نهایتاً پایایی آزمون را بهبود می‌بخشد. اما این مسأله که عینیت به صورت مستقیم منجر به پایایی می‌شود، امروزه چندان مورد قبول نیست. شواهد حاکی از آن هستند که نمره‌دهی دوحالته به خوبی قادر به افتراق دانشجوی توانمند و غیرتوانمند نیست (خان و همکاران ۲۰۱۳).

انتقاد دیگر این است که جزئی کردن بیش از حد یک توانمندی به آیتم‌ها یعنی جزئی‌سازی بیش از حد موجب می‌شود نتوان ارزیابی درستی از کل مهارت ارائه داد. برای جلوگیری از این موضوع، توصیه می‌شود از رویکردی استفاده شود که در آن به جای آن که آیتم‌های چک‌لیست خیلی ریز و جزئی شوند، چندین آیتم با یکدیگر تحت یک عنوان کلی در نظر گرفته می‌شوند. گفته می‌شود که استفاده از چنین مقیاسی باعث بهبود پایایی می‌شود (پل و همکاران ۲۰۱۰). به مثال زیر توجه کنید:

دست خود را شست.	☐ بله	☐ خیر
خود را به بیمار معرفی کرد.	☐ بله	☐ خیر
نام بیمار را پرسید.	☐ بله	☐ خیر
هدف مصاحبه را توضیح داد.	☐ بله	☐ خیر
شکل بهتر		
آغاز مصاحبه با بیمار را چگونه انجام داد؟	☐ عالی	☐ متوسط ☐ ضعیف

1. Critical action
2. Payne et al

این تغییر رویکرد نمره‌دهی در واقع منجر به نمره‌دهی درجه‌ای می‌شود. در این مدل، آیتم‌ها به صورت مراحل کلیدی نوشته می‌شوند که در طول وظیفه محوله باید رعایت شوند و برای هر یک، لیکرتی با، پنج یا هفت درجه در نظر گرفته می‌شود. در این حالت، امکان قضاوت درباره کیفیت کار انجام شده نیز وجود دارد.

از آنجا که در این نمره‌دهی از مقیاس لیکرت استفاده می‌شود، به نمره‌دهی گلوبال شباهت دارد اما از آنجا که قضاوت در مورد عملکرد کلی فراگیر در ایستگاه صورت نمی‌گیرد و در مورد هر آیتم باید قضاوت جداگانه صورت گیرد، در واقع با آن متفاوت است. به همین دلیل برای جلوگیری از اشتباه از عبارت درجه‌ای برای آن استفاده می‌شود.

چک‌لیست ایستگاه معاینه قفسه سینه با نمره‌دهی دوحالته

۱. معرفی خود	☐ بله	☐ خیر
۲. اخذ اجازه	☐ بله	☐ خیر
۳. وضعیت مناسب	☐ بله	☐ خیر
۴. رویکرد حرفه‌ای	☐ بله	☐ خیر
۵. معاینه فیزیکی عمومی	☐ بله	☐ خیر
۶. مشاهده	☐ بله	☐ خیر
۷. لمس	☐ بله	☐ خیر
۸. دق	☐ بله	☐ خیر
۹. سمع	☐ بله	☐ خیر
۱۰. تشکر بابت همکاری	☐ بله	☐ خیر
جمع نمره از ۱۰		

چک‌لیست ایستگاه معاینه قفسه سینه با نمره‌دهی درجه‌ای

☐ وظایف کلیدی									
☐ ۱. قبل و بعد معاینه از الکل استفاده کرد و هنگام ضرورت، دستکش پوشید	۱		۲		۳		۴		۵
☐ ۲. از بیمار اجازه گرفت و توضیحات لازم را به او ارائه داد.	۱		۲		۳		۴		۵
☐ ۳. در صورت لزوم از پرستار خواست که بر بالین بیمار حاضر شود.	۱		۲		۳		۴		۵
☐ ۴. از بیمار سؤال کرد که طی لمس یا معاینه قسمتی از بدن دچار درد می‌شود؟	۱		۲		۳		۴		۵
☐ ۵. بی جهت بیمار را خجالت زده نگرد یا وی را آزار نداد.	۱		۲		۳		۴		۵
☐ ۶. مناطق مرتبط را معاینه کرد یا پیشنهاد معاینه آنها را داد.	۱		۲		۳		۴		۵
☐ ۷. بعد از اتمام کار، قسمت های برهنه شده بیمار را پوشاند و از بیمار تشکر کرد.	۱		۲		۳		۴		۵

۱: رویکرد بدون ساختار مشخص
۲: رویکرد ساختارمند، انجام وظایف کلیدی کمتر از ۵۰ درصد مواقع
۳: رویکرد ساختارمند، انجام وظایف کلیدی بیشتر از ۵۰ درصد مواقع
۴: رویکرد ساختارمند، انجام اکثر وظایف کلیدی
۵: رویکرد ساختارمند، انجام تمام وظایف کلیدی

تدوین راهنمای آزمونگر

برای اینکه عملکرد آزمونگران یکسان و استاندارد باشد، نیاز است که برگه راهنمایی برای آنها تدوین شود. علاوه بر خلاصه سناریو و آنچه از دانشجو انتظار می‌رود انجام دهد، باید برای آزمونگر به صورت شفاف بیان شود که چه نقشی دارد، چه اطلاعاتی را باید در اختیار دانشجو قرار دهد و چه مواردی را نباید به دانشجو بگوید. همچنین، اطلاعات بالینی مرتبط با ایستگاه برای ارزیابی درست باید در برگه راهنما ارائه شود. این مسأله مخصوصاً وقتی اهمیت دارد که تخصص آزمونگر دقیقاً از همان دیسیپلین نیست.

تدوینِ راهنمای بیمارنما

قبلاً در مورد استفاده از بیمار واقعی و بیمارنما در OSCE صحبت شد. در صورت استفاده از هر کدام از آنها، مهم است که بیمار/بیمارنما به سؤالات تمام دانشجویان به شکل واحد پاسخ دهد یا در صورت وجود لاین موازی در آزمون، بیمار/بیمارنماهای یک ایستگاه خاص در تمام لاین‌ها اجرای یکسانی از خود نمایش دهند. بنابراین آماده کردن متن سناریو و راهنمای بیمارنما از اهمیت زیادی برخوردار است. سناریوهایی که بر اساس بیمار واقعی نوشته شده‌اند، روایی آزمون را بهبود می‌بخشند (نستل و نیبون[1] ۲۰۱۰).

اطلاعات مربوط به بیمار پس از تدوین چک‌لیست به صورت مشخص و شفاف در می‌آید. برای همین معمولاً تدوین راهنمای بیمارنما به پس از تدوین چک‌لیست موکول می‌شود.

- راهنما، باید به زبان بیمار و افراد معمولی نوشته شود و در آن از ترمینولوژی پزشکی استفاده نشود.
- نحوه لباس پوشیدن، ورود بیمار، راه رفتن، صحبت کردن و ... باید در راهنمای بیمارنما نوشته شود.
- اولین جمله‌ای که باید با آن مکالمه را شروع کند، باید به صورت مشخص نوشته شود.
- سؤالاتی که بیمارنما حتماً باید از دانشجو بپرسد، باید ذکر شود. این سؤالات با این هدف پرسیده می‌شوند که پاسخ دانشجو ارزیابی شود. در واقع متناظر با آنها آیتمی در چک‌لیست وجود دارد که دانشجو آن را برآورده می‌سازد یا نمی‌سازد. در مثال زیر، بیمارنما پرسشی در مورد تشخیص بیماری مطرح می‌کند. در چک‌لیست، آیتمی وجود دارد مبنی بر این که آیا دانشجو اطمینان بخشی زودرس به بیمار داد؟

بیمارنما: «ببخشید، شما فکر می‌کنید که من هم مثل پدرم به سرطان مبتلا شدم؟»
دانشجو: «نه، اصلاً جای نگرانی نیست. این داروها رو مصرف کنید، حتماً سرفه‌تون بهتر میشه»

- علت مراجعه، شرح بیماری، سیر آن، علایم همراه، سابقه خانوادگی، سابقه شخصی، مصرف داروها و ... باید در راهنمای بیمارنما بیاید. مخصوصاً پاسخ آیتم‌هایی که در چک‌لیست آمده است، باید در راهنمای بیمارنما ذکر شود تا بیمارنما بداند در پاسخ به سؤالات دانشجو چه جوابی بدهد. همچنین تظاهرات چهره و بدن و نوع رفتار کلامی و غیرکلامی بیمار باید ذکر شود (اسمی ۲۰۰۳). مثلاً اگر طراح سؤال قصد دارد در سناریو بیماری خجالتی داشته باشد، لازم است مواردی که به این مسأله اشاره دارند، مانند نداشتن تماس چشمی و صحبت کردن با صدای آرام در قالب جملات کوتاه، در راهنمای بیمارنما ذکر شوند.
- طراح سؤال باید به دقت همه مواردی که احتمال می‌رود دانشجو بپرسد، پیش‌بینی کند و پاسخ آنها را در متن بیمارنما بگنجاند.

مرور ایستگاه

پس از پایان طراحی موارد فوق مهم است که خود طراح سؤال یک بار تمام ماتریال را کنترل نماید. از آنجا که همیشه احتمال خطا وجود دارد، مرور ایستگاه توسط همکاران اهمیت بسیاری دارد. همان طور که قبلاً ذکر شد، فرایند بررسی ایستگاه‌ها می‌تواند در کمیته OSCE و با حضور تمام طراحان انجام شود.

در جدول ۱۱–۱۸ اجزای طراحی‌شده برای یک ایستگاه به صورت کامل نشان داده شده است.

1. Nestel & Kneebone

جدول ۱۱-۱۸: اجزای کامل یک ایستگاه شامل شناسنامه ایستگاه، اطلاعات مورد نیاز برای کمیته اجرایی، راهنمای دانشجو، راهنمای آزمونگر، چک‌لیست و راهنمای بیمارنما در یک ایستگاه OSCE پیش کارورزی دانشگاه علوم پزشکی تهران اسفند ۸۸

شناسنامه ایستگاه

شماره ایستگاه: ۱
موضوع ایستگاه: معاینه اعصاب موتور مغزی
تاریخ طراحی: .....
نام طراح: .....
سطح دانشجویان: سال چهارم پزشکی
زمان ایستگاه: ۵ دقیقه
توانمندی مورد ارزیابی: معاینه فیزیکی و تشخیص
نوع مشکل: ☐ مزمن ☐ حاد ☐ تحت حاد
سناریو: بیمار آقای ۳۰ ساله‌ای است که با شکایت دوبینی در هنگام پایین آمدن از پله ها مراجعه نموده است. دانشجو باید در عرض ۵ دقیقه اعصاب موتور مغزی بیمار را معاینه کامل نماید، مشکل وی را تشخیص دهد و به آزمونگر بگوید.

اطلاعات مورد نیاز برای کمیته اجرایی

شماره ایستگاه: ۱
نام ایستگاه: معاینه اعصاب موتور مغزی
مشخصات بیمارنما: آقای ۳۰ ساله
فهرست امکانات
صندلی برای بیمارنما
چراغ قوه
چیدمان ایستگاه
میز و صندلی آزمونگر باید طوری باشد که حرکات دانشجو و صورت بیمارنما را مشاهده کند.
دانشجو به میز و صندلی و قلم نیازی ندارد.

راهنمای دانشجو

شماره ایستگاه: ۱
نام ایستگاه: معاینه اعصاب موتور مغزی
مدت زمان ایستگاه: ۵ دقیقه
دانشجوی گرامی،
در این ایستگاه شما در درمانگاه نورولوژی با آقای ۳۰ ساله‌ای مواجه می‌شوید که با شکایت دوبینی در هنگام پایین آمدن از پله‌ها مراجعه نموده است. شما در این ایستگاه باید در عرض ۵ دقیقه اعصاب موتور مغزی مرتبط را معاینه نمایید، مشکل وی را تشخیص دهید و به آزمونگر بگویید.
به نکات زیر به دقت توجه کنید:

- آن چه در این ایستگاه مورد ارزیابی قرار می گیرد انجام اقدام صحیح با روش درست و رسیدن به تشخیص نهایی می‌باشد.
- گرفتن شرح حال و معاینه کامل عصبی مد نظر نیست.
- فردی که در اتاق حضور دارد عملکرد شما را ارزیابی می‌کند و شما به جز گفتن تشخیص نهایی به او، هیچ تعاملی با وی نخواهید داشت. لذا:
  - از او در مورد این که بیمار چه وضعی دارد یا شما باید چه کاری را انجام دهید سؤال نکنید.
  - مراحل کار را برای او توضیح ندهید و فقط اقدامات را همان گونه که لازم است انجام دهید (همانند شرایط واقعی و گویی که هیچ کس در اطراف شما نیست).
- اقدامات خود را تا زمانی که صدای زنگ را می شنوید ادامه دهید و سپس بلافاصله از اتاق خارج شوید.

راهنمای آزمونگر

شماره ایستگاه: ۱
نام ایستگاه: معاینه اعصاب موتور مغزی
مدت زمان ایستگاه: ۵ دقیقه
هدف از این ایستگاه ارزیابی توانمندی دانشجویان در انجام معاینات اعصاب موتور مغزی مرتبط با دوبینی (۳ و ۴ و ۵) و رسیدن به تشخیص درست می‌باشد.
سناریویی در اختیار داوطلب قرار داده شده است که بر اساس آن از وی خواسته شده است که کلیه معاینات را برای فردی که با دوبینی مراجعه نموده است، انجام دهد. آن چه در این ایستگاه مورد ارزیابی قرار می گیرد انجام اقدام صحیح با روش درست و رسیدن به تشخیص نهایی می باشد.
در این ایستگاه شما هیچ تعامل مستقیمی با فراگیر نخواهید داشت و فقط باید عملکرد او را مورد مشاهده و ارزیابی قرار دهید. از داوطلب اکیداً خواسته شده است که از توضیح و تشریح اقدامات برای شما خودداری کند و همان طور که لازم می داند صرفاً اقدامات را انجام دهد (همانند شرایط واقعی و گویی که هیچ کس در اطراف او نیست). بنابراین، تنها لازم است تشخیص نهایی را به شما بگوید.
شما باید بر اساس عملکرد داوطلب چک لیستی را که در اختیار شما قرار دارد تکمیل نمایید.
لطفاً به نکات زیر به دقت توجه فرمایید:

- چک لیست را به دقت پر کنید.
- به هیچ عنوان با داوطلب صحبت نکنید: به سؤالات وی پاسخ ندهید و در مورد عملکردش به او بازخورد ندهید.
- داوطلب باید تا هنگامی که صدای زنگ را می شنود به کار خود ادامه دهد و سپس بلافاصله از اتاق خارج شود.

چک لیست

شماره ایستگاه: ۱
نام ایستگاه: معاینه اعصاب موتور مغزی
نام دانشجو:
شماره دانشجو:

	بله	خیر	
۱. خود را به بیمار معرفی کرد			(۰/۵)
۲. به بیمار گفت که چه کاری می¬خواهد انجام دهد و از وی اجازه گرفت.			(۰/۵)
۳. در حالت نرمال به چشمهای بیمار نگاه کرد.			(۱)
۴. حرکت چشم‌ها را در محور افقی بررسی کرد.			(۱)
۵. حرکت چشم‌ها را در محور عمودی بررسی کرد.			(۱)
۶. وضعیت تقارب چشم‌ها (convergence) را بررسی کرد.			(۱)
۷. واکنش مردمک‌ها به نور را بررسی کرد.			(۱)
۸. حرکت چشمها به پایین و داخل (نوک بینی) را بررسی کرد			(۲)
۹. فلج زوج ۴ را در بیمار تشخیص داد.			(۲)
جمع نمره			۱۰

در مجموع نظر شما در مورد عملکرد این دانشجو در این ایستگاه چیست؟ (فقط یک مورد را انتخاب کنید)

عالی ☐ خوب ☐ قابل قبول ☐ مرزی ☐ رد ☐

راهنمای بیمارنما

شماره ایستگاه: ۱
نام ایستگاه: معاینه اعصاب موتور مغزی
شما آقای ۳۰ ساله‌ای هستید که به دلیل دوبینی به درمانگاه مراجعه کرده‌اید. افتادگی پلک و انحراف چشم دارید ولی عینکی نیستید. روی صندلی نشسته‌اید و طبق درخواست‌های معاینه کننده عمل می‌کنید ولی اگر از شما خواست که به پایین و داخل چشم یا به نوک بینی خود نگاه کنید، ابراز ناتوانی نمایید.

اجرای OSCE

روز برگزاری آزمون نه تنها برای دانشجویان که برای برگزارکنندگان نیز روز پراسترسی است. هر چقدر برگزارکنندگان در مرحله طراحی OSCE با دقت بیشتری عمل کرده باشند، در این روز تنش کمتری خواهند داشت. اما در هر حال، باید آمادگی برای رویارویی با هر رخدادی را داشته باشند.

دست‌اندرکاران اجرایی باید صبح زود در محل حاضر شوند. باید به دانشجویان اعلام کرد که یک تا دو ساعت زودتر از زمان شروع آزمون در محل حاضر شوند. بیمارنمایان و آزمونگران نیز باید نیم تا یک ساعت قبل از شروع آزمون در محل حاضر شوند.

یک نفر از تیم اجرایی باید مسؤول این باشد که یک بار دیگر تمام ایستگاه‌ها را کنترل کند، مطمئن شود که شماره لاین‌ها و شماره ایستگاه‌ها به درستی نصب شده است، صدای زنگ را امتحان کند، کورکننده آنتن موبایل را آزمایش کند، ماتریال چاپ شده از جمله چک‌لیست و راهنمای آزمونگر و بیمارنما را در ایستگاه کنترل کند، از کار کردن کامپیوترها و نرم افزارها مطمئن شود.

یک نفر از دست‌اندرکاران اجرایی باید مسؤول ثبت نام دانشجویان باشد و فهرست دانشجویان از قبل در اختیار او قرار گرفته باشد. با ورود دانشجویان باید هویت آنها را تایید کند، نام آنان را ثبت کند و برچسب‌های مربوط به دانشجویان را که کد شناسایی مخصوص روی آن نوشته شده است، به آنها تحویل دهد تا در هر ایستگاه به آزمونگر ارائه دهند. دانشجویان باید وسایل شخصی و موبایل خود را تحویل دهند و سپس به سمت قرنطینه هدایت شوند.

با اینکه قبلاً در مورد آزمون به دانشجویان اطلاعات کافی ارائه شده است، یک بار دیگر باید خلاصه‌ای از نحوه برگزاری آزمون را برای آنها شرح داده شود. پیشنهاد می‌شود موارد زیر در این جلسه مطرح شود:

- توضیح در مورد تعداد ایستگاه‌ها، نحوه آغاز حرکت و چرخش در ایستگاه‌ها، ایستگاه استراحت
- نحوه حضور در قرنطینه
- یادآوری قوانین و مقررات از جمله عواقب همراه داشتن تلفن همراه
- تاکید به ارائه برچسب حاوی کد شناسایی در هر ایستگاه به آزمونگر
- تاکید به مطالعه راهنمای دانشجو به دقت در ابتدای هر ایستگاه
- پخش صدای زنگ به صورت امتحانی و تاکید بر لزوم توقف فعالیت‌ها و ترک ایستگاه با شنیدن صدای آن
- تاکید بر تکمیل فرم نظرسنجی بعد از آزمون

همچنین، یک یا چند نفر از برگزار کنندگان OSCE باید در راهرو حاضر باشند تا دانشجویان را راهنمایی کنند. بسیار محتمل است که دانشجویان تحت شرایط استرس نتوانند مسیر خود را به درستی پیدا کنند و سردرگم شوند. استرس ممکن است باعث شود دانشجو هنگام خروج از ایستگاه وسیله‌ای را بردارد و خارج کند. آزمونگران و مراقبان آزمون باید مراقب این امر باشند.

یک نفر از دست‌اندرکاران اجرایی باید مسؤول رسیدگی و سامان‌دهی امور مربوط به بیمارنماها باشد. وی باید اسامی و شماره افراد را در اختیار داشته باشد تا برای بیمارنماهایی که حاضر می‌شود، علامت بزند و اگر کسی دیر کرده است، با او تماس بگیرد. اتاقی برای انتظار بیمارنماها باید در نظر گرفته شده باشد که در آن جمع شوند و توضیحات مهم یک بار دیگر برای آنها ارائه شود. پیشنهاد می‌شود موارد زیر برای آنها ذکر شود:

- اهمیت عملکرد یکسان و استاندارد بین بیمارنمایان مختلف طبق راهنماهای تدوین شده
- تکمیل فرم نظرسنجی بعد از آزمون

سپس شماره ایستگاهی که باید در آن حضور یابند، طبق نقشه‌ای که از قبل مشخص شده است، باید به اطلاع تک تک آنها برسد و بر اساس آن به سمت ایستگاه مورد نظر راهنمایی شوند. پذیرایی در فواصل مناسب باید از ایشان صورت گیرد.

همچنین لازم است یک نفر دیگر مسؤول امور مربوطه به آزمونگران باشد و همان مواردی را که برای بیمارنمایان ذکر شد، انجام دهد. پیشنهاد می‌شود موارد زیر برای آزمونگران ذکر شود:

- هدف آزمون، تکوینی یا تراکمی بودن آن
- مرور ابزار ارزیابی (چک‌لیست یا گلوبال) و نحوه تکمیل و نمره‌دهی در آن
- لزوم دریافت برچسب حاوی کد شناسایی در هر ایستگاه از دانشجو
- لزوم و اهمیت محرمانه ماندن اطلاعات دانشجو
- عدم تعامل و صحبت با دانشجو (مگر مواردی که در سناریوی ایستگاه خواسته شده است یا در مواردی که قرار است بازخورد به دانشجو ارائه شود)
- رفتار با همه دانشجویان به مساوات
- تکمیل فرم نظرسنجی بعد از آزمون

از این بین نحوه نمره‌دهی اهمیت زیادی دارد که باید بر آن تاکید شود.

پس از برگزاری آزمون، چک‌لیست‌ها و پاسخنامه‌ها باید جمع‌آوری و شمارش شود. نمرات چک‌لیست جمع بسته شود و وارد فایلی مانند اکسل شود تا نهایتا با جمع نمرات ایستگاه‌ها، نمرات نهایی محاسبه شوند. در صورتی که تعداد دانشجویان زیاد باشد، این کار بسیار وقت‌گیر است. به همین دلیل بهتر است از قبل تیمی برای این کار در نظر گرفته شوند که بلافاصله بعد از اتمام آزمون شروع به کار کنند. البته در حالتی که از نرم افزارهای کامپیوتری به جای برگه‌های کاغذی استفاده شده باشد، نیازی به انجام این کار نیست زیرا مراحل تصحیح و تحلیل آزمون نیز توسط نرم افزار صورت می‌گیرد.

اگر روش تعیین حدنصاب قبولی از بین روش‌های مبتنی بر آزمون مانند رگرسیون مرزی یا گروه مرزی انتخاب شده است، پس از محاسبه نمرات باید حدنصاب هر ایستگاه و حدنصاب کل آزمون تعیین گردد. به این ترتیب نهایتاً می‌توان نتیجه ردی/ قبولی دانشجویان را اعلام کرد.

پس از اعلام نمرات و نتیجه آزمون هنوز کار تمام نشده است و طبق برنامه‌ای که از قبل برای ارزشیابی و تحلیل آزمون ریخته شده بود، باید آنالیزها و اقدامات لازم صورت گیرد و نتایج آن در اسرع وقت به کمیته OSCE بازخورد داده شود.

پیشامدهای غیرمنتظره و آمادگی برای آنها

در روند اجرای OSCE اتفاقاتی رخ می‌دهد که گاهی قابل کنترل نبوده و نمی‌توان برای آن کاری کرد. از آنجا که این اتفاقات کم و بیش در آزمون‌های مختلف تکرار شده است، مسؤول برگزاری OSCE باید از تجربیات قبلی استفاده کند تا برای رویارویی با آنها آماده باشد. برخی از این اقدامات پیش‌گیری‌کننده هستند و از قبل باید انجام شوند. اما برخی دیگر در همان جلسه آزمون باید رفع گردند. به همین دلیل تا حد امکان بهتر است مسؤول OSCE به عنوان آزمونگر در ایستگاه قرار نگیرد. زیرا مطمئناً موارد زیادی پیش خواهد آمد که نیاز به تصمیم‌گیری فوری و اتخاذ راه حل مناسب دارد.

- دیر رسیدن آزمونگر یا بیمارنما به جلسه یا خسته شدن آنها طی جلسه: داشتن بیمارنما و آزمونگر جایگزین برای مواردی که افراد دیر به جلسه می‌رسند یا در نیمه آزمون خسته می‌شوند، از بی‌نظمی‌های احتمالی و وقوع بحران جلوگیری می‌کند.
- به صدا در آمدن زنگ موبایل آزمونگر یا بیمارنما در جلسه: از آزمونگر و بیمارنما باید خواست تا گوشی خود را خاموش

کنند. علاوه بر اینکه صدای زنگ آن حواس دانشجو را پرت می‌کند، در کار بیمارنما و آزمونگر نیز خدشه وارد می‌شود. مثلاً آزمونگری که گوشی خود را در جلسه در می‌آورد تا آن را خاموش کند، عملاً در این زمان عملکرد دانشجو را مشاهده نکرده است.

- عدم تطابق رفتار بیمارنما با آنچه در سناریو آمده است: مثلا بیماری که باید از کمردرد شاکی باشد، خیلی خوش‌رو و پرحرف باشد. راهنمای بیمارنما، آموزش به آنها، انجام تمرین و ارائه بازخوردهای مکرر از این اتفاق و موارد مشابه آن جلوگیری می‌کند.
- صحبت آزمونگر و بیمارنما طی جلسه: اگر دانشجو در ایستگاه وقت اضافه بیاورد، ایستگاه را ترک نمی‌کند تا صدای زنگ را بشنود. در این حالت ممکن است آزمونگر و بیمارنما با یکدیگر صحبت کنند که چندان درست نیست و بهتر است از این کار اجتناب کنند. اما اگر صحبت در مورد عملکرد دانشجو است، کاری غیر حرفه‌ای است و اصلاً نباید انجام شود.
- دانشجو با صدای زنگ ایستگاه را ترک نکند: دانشجویان علاقه دارند از وقت بیشتری استفاده کنند. آزمونگر باید صریحاً از دانشجو بخواهد که ایستگاه را ترک کند.
- گاهی دانشجو پس از اینکه وظیفه موردنظر را انجام داد، یکی از مراحل را به یاد می‌آورد و به آزمونگر می‌گوید: «راستی، یادم رفته بود. باید دستم را می‌شستم». سؤال این است که در این حالت آیا باید نمره را دریافت کند یا نه. بهتر است که راهکار مورد نظر در موارد این چنینی قبلاً توسط مسؤول OSCE به اطلاع همه آزمونگران برسد که می‌تواند این باشد که چون هر مرحله باید سر جای خودش انجام شود، نمره‌ای تعلق نمی‌گیرد یا می‌تواند این باشد که اگر دانشجو صرفاً گفت، نمره‌ای نمی‌گیرد زیرا در OSCE باید عملا کار را انجام دهد اما اگر کار را انجام داد، نمره را می‌گیرد.
- اضطراب شدید دانشجو: دانشجویان معمولا در OSCE درجاتی از استرس دارند. گاهی دانشجو آنقدر مضطرب است که حتی نمی‌تواند افتالموسکوپ را روشن کنند. در این حالت آزمونگر هرچند که باید سعی کند دانشجو را آرام کند اما نباید در کار او مداخله کند و دستگاه را برای او آماده کند. گاهی دانشجویی که وارد ایستگاه شده اضطراب بسیار شدیدی دارد که البته به ندرت پیش می‌آید. در این مورد، با هماهنگی مسؤول OSCE آزمونگر می‌تواند او را از ایستگاه خارج کند و فرصتی به او داده شود تا پس از مسلط شدن به خود مجددا در ایستگاه حاضر شود.
- از کار افتادن دستگاه: داشتن ذخیره از وسایل مصرفی و امکانات موجود در ایستگاه برای مواقعی که دستگاه خراب می‌شود یا کار نمی‌کند، ضروری است. نبود وسیله‌ای جزئی مانند خودکار یا باتری می‌توان به راحتی روند آزمون را مخدوش کند.

OSCE متوالی

افزایش تعداد ایستگاه‌های OSCE با کاهش خطای اندازه‌گیری و افزایش پایایی آزمون رابطه دارد اما افزایش تعداد ایستگاه‌ها، مستلزم صرف امکانات و منابع بیشتر و هماهنگی با تعداد زیادی آزمونگر و بیمار(نما) و افزایش هزینه است. به منظور مقابله با مشکل هزینه در آزمون OSCE، روشی به نام سنجش متوالی[1] پیشنهاد شده است که در عین حالی که ادعا می‌شود پایایی بالایی دارد، در مصرف منابع صرفه‌جویی می‌کند (گرملی ۲۰۱۱).

در OSCE متوالی تمام دانشجویان ابتدا در یک آزمون با تعداد محدودی ایستگاه شرکت می‌کنند. به عنوان مثال اگر قرار بوده در آزمون اصلی، ۲۰ ایستگاه وجود داشته باشد، آزمون اولیه تنها شامل ۱۰ ایستگاه یا کمتر است. با کاهش تعداد ایستگاه و بیمارنما و آزمونگر، بالتبع هزینه برگزاری این امتحان کمتر خواهد بود. دانشجویی که در این آزمون رد می‌شود،

1. Sequential OSCE

در آزمون تکمیلی شرکت می‌کند و مجموع عملکرد او در دو آزمون ملاک تصمیم‌گیری در خصوص وضعیت او خواهد بود؛ مانند اینکه یک بار در آزمون اصلی شرکت کرده باشد. اما دانشجویانی که در آزمون اولیه قبول می‌شوند، از امتحان بعدی معاف می‌شوند و تنها همان آزمون اول برای ایشان لحاظ می‌شود.

آنچه در سنجش متوالی اهمیت دارد این است که نتایج تست اول باید با دقت قابل قبولی عملکرد دانشجویان را در آزمون اصلی پیش‌بینی کند، به عبارت دیگر توانایی غربال[1] فراگیران توانمند را به درستی داشته باشد. برای اینکه آزمون غربال به صورت موثر و قابل اطمینان اجرا شود، رعایت چند نکته را ضروری دانسته‌اند. از جمله اینکه در ابتدا ایستگاه‌های کل آزمون طراحی شوند و سپس آنچه قرار است در آزمون اول گنجانده شود، از میان آنها انتخاب شود. به عبارت دیگر، آزمون غربال باید در زمینه و متن بلوپرینت کل آزمون در نظر گرفته شود. همچنین باید توجه شود به طور معمول چه تعداد دانشجو در آزمون رد می‌شوند. در اغلب آزمون‌های مشابه انتظار می‌رود تعداد کمی از دانشجویان رد شوند. نکته مهم دیگر اینکه امتحان غربال باید به گونه‌ای طراحی شود که منفی کاذب[2] و مثبت کاذب[3] آن پایین باشد (راتمن و همکاران[4] ۱۹۹۷).

مثبت کاذب یعنی دانشجویانی که در امتحان غربال قبول شده‌اند اما نهایتاً در کل آزمون رد می‌شوند. اگر مثبت کاذب کم باشد، نشان دهنده اشتباهات کم در تصمیم‌گیری و ویژگی[5] بالای آزمون غربال است. زیرا OSCE آزمونی برای تایید صلاحیت بالینی فراگیران است و اگر دانشجویی که توانمند نیست، قبول اعلام شود، تهدید جدی محسوب می‌شود. منفی کاذب یعنی دانشجویانی که در آزمون غربال قبول شده‌اند اما عملکرد آنها مجموعاً در دو آزمون قابل قبول بوده است و نهایتاً قبول اعلام شده‌اند. اگر تعداد موارد منفی یعنی دانشجویانی که از آزمون اول رد شده‌اند، کم باشد، به این معناست که آزمون دوم (تکمیلی) برای تعداد کمی از دانشجویان برگزار می‌شود و مقرون به صرفه است. در حالی که اگر موارد منفی زیاد باشد، آزمون دوم باید برای تعداد زیادی از دانشجویان مجدداً تکرار شود و عملاً مانند است که هدف اصلی که صرفه جویی در هزینه است، اتفاق نیفتاده و حتی بدتر شده است. به صورت ایده‌ال هم مثبت کاذب و هم منفی کاذب باید کم باشند؛ اما کم بودن مثبت کاذب در آزمون غربال اهمیت بیشتری دارد. البته این موضوع به حد نصاب قبولی (استاندارد) آزمون نیز مرتبط است (راتمن و همکاران ۱۹۹۷، موجنس و همکاران[6] ۲۰۰۰).

در یک مقاله مروری در مورد روش اجرای OSCE به شیوه سنجش متوالی آمده است که برای تعیین رد و قبول دانشجویان در آزمون غربال، می‌توان با استفاده از یک متد تعیین استاندارد مانند روش رگرسیون مرزی[7]، حدنصاب قبولی را محاسبه کرد. سپس می‌توان برای اطمینان خاطر بیشتر و کاهش موارد مثبت کاذب، حدنصاب قبولی را به میزان دو خطای استاندارد اندازه‌گیری[8] بالاتر از عدد به دست آمده اعلام کرد. در این حالت، با اطمینان بسیار خوبی می‌توان گفت که این گروه از دانشجویان واقعاً توانمند و شایسته قبولی هستند (گرملی ۲۰۱۱).

می‌توان گفت که در سنجش متوالی، ضمن حفظ پایایی کل آزمون، منابع در جایی استفاده می‌شوند که بیشترین تاثیر را برای تصمیم‌گیری درست در مورد رد و قبول دانشجویان به دنبال دارند. در واقع لازم نیست که پایایی آزمون برای همه دانشجویان بالا باشد. بلکه منابع به منظور افزایش تعداد ایستگاه‌ها و افزایش پایایی، جایی صرف می‌شوند که عملکرد دانشجویان مرزی است و تاثیر خطای اندازه‌گیری بر رد و قبول دانشجویان محسوس و تعیین کننده است (گرملی ۲۰۱۱).

از انجا که روی کاغذ و به صورت تئوری، برگزاری OSCE متوالی منطقی به نظر می‌رسید، پژوهشگران به فکر افتادند که به بررسی‌های تجربی درباره آن بپردازند. اولین مطالعه تجربی در مورد سنجش متوالی در OSCE توسط کولیور و

1. Screening
2. False Negative.
3. False Positive
4. Rothman et al
5. Specificity
6. Muijtjens et al
7. Borderline regression method.
8. Standard Error of Measurement (SEM)

همکاران در دانشگاه ایلینوی جنوبی به صورت گذشته‌نگر انجام شد:

کولیور و همکاران ۱۹۹۱ (۲)

در این پژوهش، از OSCE سه روزه‌ای که برای ۴۰۴ دانشجوی پزشکی برگزار شد، استفاده شد. پژوهشگران روز اول امتحان را به عنوان آزمون غربال در نظر گرفتند و عملکرد دانشجویان در آن روز را با عملکرد کلی ایشان مقایسه کردند. از آنجا که تعداد زیادی از دانشجویان (۲۲۸ نفر معادل ۵۶ درصد) در آزمون غربال قبول شدند، هزینه و زمان آزمون کاهش یافت. همچنین، اکثریت کسانی که در آزمون غربال قبول شدند، در کل آزمون سه روزه نیز قبول شدند. در واقع فقط یکی از این ۲۲۸ دانشجو در آزمون کلی رد شد. بنابراین میزان خطای قبولی، ۰/۴ درصد برآورد شد.

نویسندگان نتیجه‌گیری کردند که آزمون غربال، به خوبی می‌تواند عملکرد دانشجویان در کل آزمون را پیش‌بینی کند اما در عین حال ذکر کردند که تعداد ایستگاه‌های آزمون غربال و همچنین نمره حدنصاب قبولی در آن می تواند بر صحت نتایج به دست آمده از آزمون غربال تاثیر بگذارد که در این مطالعه به آن پرداخته نشده بود (کولیور و همکاران ۱۹۹۱ (۲)). مطالعه بعدی این محققان به منظور بررسی دو مورد فوق صورت گرفت:

کولیور و همکاران ۱۹۹۲

در این مطالعه، چند نوع آزمون غربال با طول متفاوت از دو ایستگاهه تا ده ایستگاهه در نظر گرفته شد. همچنین برای حد نصاب قبولی، که به صورت میانگین ایستگاه‌های مختلف حساب می‌شد، خطای استاندارد اندازه‌گیری نیز محاسبه شد و برای هر آزمون غربال، پنج نمره حدنصاب قبولی با اضافه کردن ضرایبی از SEM (منفی یک، منفی نیم، صفر، مثبت نیم، مثبت یک) نسبت به میانگین تعیین شد (کولیور و همکاران ۱۹۹۲). بنابراین با در نظر گرفتن دو فاکتور تعداد ایستگاه و حدنصاب قبولی، برای هر آزمون غربال، جداول دو در دویی تهیه شد که میزان قبولی دانشجویان در آزمون غربال را با میزان قبولی آنها در کل آزمون نشان می داد. همچنین، برای هر آزمون، یک ROC curve رسم شد که در آن هر چقدر سطح زیر منحنی بیشتر باشد، نشان دهنده غربالگری موفق‌تر است. به طوری که صفر به معنای نبود صحت و ۱ به معنای دقیق‌ترین پیش‌بینی است. ضمنا نزدیک‌ترین نقطه به گوشه چپ و بالای منحنی به عنوان بهترین حدنصاب قبولی در نظر گرفته شد. نتایج نشان داد که همان طور که انتظار می‌رفت افزایش تعداد ایستگاه‌ها صحت آزمون غربال را افزایش می‌دهد اما نویسندگان ذکر کردند صحت آزمون غربال شش ایستگاهه یعنی یک روز از سه روز تفاوت چندانی با آزمون ده ایستگاهه یعنی نصف آزمون کل نداشته است (به ترتیب ۰/۸۱ و ۰/۹). در مورد حد نصاب قبولی نشان داده شد که بهترین نمره از افزودن SEM به میزان نیم برابر به میانگین به دست می‌آید. حساسیت و ویژگی آزمون غربال شش ایستگاهه به ترتیب ۰/۷۹ و ۰/۷۴ محاسبه شد. این اعداد در مورد آزمون غربال ده ایستگاهه به صورت مختصر زیاد شد و ۰/۸۲ و ۰/۸۴ به دست آمد. ارزش اخباری مثبت آزمون شش ایستگاهه ۰/۹۸ بود. به عبارت دیگر، اگر کسی در آزمون اول قبول می‌شد، ۹۸ درصد شانس قبولی در آزمون کل را داشت.

بر اساس یافته‌های پژوهش فوق، نویسندگان به صورت کلی تعداد بهینه ایستگاه‌ها را حدود یک سوم میزان کل برآورد کردند و بیان کردند که برای به حداکثر رساندن حساسیت و ویژگی، نمره حدنصاب باید کمی بالاتر از میانگین در نظر گرفته شود (کولیور و همکاران ۱۹۹۲).

مسأله باقی‌مانده، ویژگی ایستگاه‌هایی بود که برای غربال باید مورد استفاده قرار گیرند. به این منظور پژوهش دیگری توسط همین نویسندگان صورت گرفت که نتایج آن نشان داد یک سوم ایستگاه‌های آزمون اصلی که شاخص همبستگی ایتم با کل آزمون[1] در آنها بالا است، به عنوان تست اولیه مناسب هستند و از این لحاظ شاخص دشواری[2] ایستگاه تاثیری ندارد (کولیور و همکاران ۱۹۹۴).

اما کمی قبل، در سال ۱۹۹۲، کاس و همکاران[3] مقاله‌ای منتشر کرده بودند که در آن توضیح دادند استفاده از شاخص ITC نمی‌تواند معیار خوبی برای انتخاب ایستگاه‌های آزمون اول باشد. زیرا هدف OSCE در امتحان اخذ مجوز، پیش‌بینی دقیق میزان قبولی/ردی است که یک متغیر کیفی دوحالته است؛ و نه تعیین نمره هر فراگیر که یک متغیر پیوسته است.

1. Item-Total Correlation (ITC)
2. Item difficulty
3. Cass et al.

بنابراین استفاده از شاخصی که قادر به پیش‌بینی متغیر دوتایی[1] است، منطقی‌تر است. از طرف دیگر هنگام استفاده از ITC رابطه و همبستگی بین خود ایتم‌ها (ایستگاه‌ها) در نظر گرفته نمی شود و ممکن است این مسأله پیش آید که در واقع ایستگاه‌های منتخب، به سنجش یک واریانس واحد بپردازند و نتوانند در کنار یکدیگر به بهبود پیش‌بینی کمک کنند. این نویسندگان برای مقابله با این دو مشکل، رویکرد رگرسیون با متغیر دوحالته (رد/قبول) را پیشنهاد دادند. با این حال، به دو مشکل در خصوص برگزاری OSCE متوالی اشاره کردند: یکی مسأله روایی محتوایی آزمون که با کاهش تعداد ایستگاه‌ها پیش می‌آید و احتمال دارد که فراگیران با آمادگی کمتری در امتحان شرکت کنند. دیگری این که ممکن است آزمون‌گیرندگان در آزمون دوم نسبت به فراگیران دید منفی داشته باشد و سخت‌گیری بی‌مورد صورت گیرد. پیشنهاد این پژوهشگران برای رفع هر دو مشکل فوق این بود که تعدادی از دانشجویانی که در آزمون اول قبول محسوب می‌شوند، به صورت تصادفی برای شرکت اجباری در آزمون دوم انتخاب شوند (کاس و همکاران ۱۹۹۲).

با وجود اینکه مطالعات فوق، در راستای تایید استفاده از OSCE متوالی بودند، در واقع از لحاظ متدولوژیک باید گفت که عملاً به صورت متوالی برگزار نشده بودند بلکه آنالیز داده‌ها به صورت گذشته‌نگر صورت گرفته بود. اسمی و همکاران[2] مطالعه خود را روی OSCE سازمان پزشکی کانادا انجام دادند که با شرکت تعداد زیادی دانشجو به صورت همزمان و در چندین مرکز مختلف اجرا می‌شود. در این طرح آنها تصمیم گرفتند واقعا یک OSCE غربال برگزار کنند:

اسمی و همکاران ۲۰۰۳

آزمون OSCE سازمان پزشکی کانادا در سال ۱۹۹۷ کلاً ۱۹۵۲ شرکت کننده داشت و در ۱۴ مرکز به صورت سنجش متوالی در دو مرحله صبح و عصر به اجرا درآمد: صبح آزمون غربال با ده ایستگاه و عصر همان روز، آزمون دوم. ایستگاه‌هایی برای آزمون غربال انتخاب شدند که تمایز بالایی داشتند یعنی بر اساس نتایج آنالیز آزمون سال‌های قبل، ITC بالایی با نمره کل آزمون داشتند. ۶۰ درصد شرکت‌کنندگان در آزمون اول قبول محسوب شدند.
در کل، نمره ۵۴۶ نفر به گونه‌ای بود که در آزمون غربال رد شدند و باید در آزمون دوم شرکت می کردند (میانگین نمره ۶۱/۳ در مقایسه با میانگین نمره ۷۱/۴ دانشجویانی که از امتحان اول قبول شده بودند). علاوه بر این، برای جلوگیری از برچسب خوردن دانشجویانی که در امتحان دوم شرکت می‌کردند، ۹۳ دانشجو از میان قبول شدگان نیز به صورت رندوم برای شرکت در آزمون دوم انتخاب شدند (میانگین نمره ۶۹/۵) که البته در نهایت بدون توجه به عملکرد ایشان در آزمون دوم، قبول اعلام شدند. میزان کلی دانشجویان رد شده، ۱/۷ در صد به دست آمد که که کمترین میزان ردی طی چند سال گذشته بود پایایی آزمون غربال، آزمون دوم و آزمون کلی با در نظر گرفتن ۲۰ ایستگاه و با استفاده از آلفای کرونباخ به ترتیب معادل ۰/۶۶، ۰/۵۶ و ۰/۶۵ به دست آمد.
در مقایسه با آزمون سال ۱۹۹۶ که تعداد شرکت‌کنندگان کمتری داشت، مشخص شد که در OSCE متوالی، کاهش نیاز به آزمونگر (تا ۲۵ درصد) و کاهش نیاز به بیمار استاندارد شده (تا ۲۸ درصد) به وجود آمد. همچنین، صرفه‌جویی در هزینه به میزان ۳۵۰ هزار دلار را به دنبال داشت. به این ترتیب از هر دانشجو به جای ۱۲۰۰ دلار، ۱۰۰۰ دلار هزینه شرکت در امتحان اخذ شد.
نویسندگان در انتها از دانشجویان در خصوص سنجش متوالی نظرخواهی کردند. ۵۵ درصد مایل نبودند که این رویه در آینده ادامه پیدا کند. ۳۵ درصد افراد بیان کردند حتی در صورتی که مشکلات مربوط به سرعت تصحیح حل شود، مخالف این روش هستند.

بدیهی است که با چنین طرحی از مطالعه که به اجرای OSCE متوالی واقعی می‌پردازد، بررسی میزان صحت پیش‌بینی آزمون غربال امکان‌پذیر نیست. زیرا آزمون دوم از همه دانشجویان به عمل نیامده و نمی‌توان درباره موارد مثبت کاذب صحبت کرد. اما در مقایسه با سایر آزمون های مشابه که در نهایت تعداد افتاده کمی داشتند و با توجه به اینکه عدم پذیرش ۴۰ درصد دانشجویان حاشیه امنیت وسیعی ایجاد کرده بود، نویسندگان معتقد بودند که موارد مثبت کاذب را به حداقل رسانده‌اند.

نکته دیگر اینکه اگرچه هزینه آزمون کاهش یافت، به علت فاصله زمانی اندک بین دو آزمون و لزوم تسریع در امر تصحیح در همان محل آزمون و اعلام سریع نتایج آزمون صبح، کار اجرایی افزایش یافته بود به این ترتیب که تیم اجرایی از ۲۳۴ نفر در سال ۹۶ به ۵۹۴ نفر در سال ۹۷ افزایش پیدا کرد.

1. Dichotomous
2. Smee et al

همچنین بر اساس نتایج این مطالعه، علی‌رغم اینکه سنجش متوالی از دید برگزارکنندگان آزمون جذاب بود؛ هزینه‌ها را کاهش می‌داد، مشکل بانک سؤالات را تا حدی حل می‌کرد و همچنین مشکل هماهنگی با تعداد زیاد آزمونگر و بیمار را کاهش می‌داد، اما دیدگاه‌های دیگری وجود داشت که با این روش چندان موافق نبودند. از جمله اینکه علی‌رغم تمهیدات اندیشیده شده، روند تصحیح برگه‌ها با سرعت دلخواه پیش نرفت و خطا در آن مشهود بود. از طرف دیگر دانشجویان و اعضای هیات علمی که در مورد هدف و فرایند سنجش متوالی توجیه نشده بودند، پس از آزمون به باورهای نادرست دامن زدند که منجر به بروز نارضایتی و اختلال شد. این موضوع نشان دهنده لزوم و ضرورت توجیه و آموزش کافی افراد قبل از برگزاری آزمون متوالی است (اسمی و همکاران ۲۰۰۳).

مطالعه زیر نیز به بیان تجربه برگزاری یک OSCE به روش ارزیابی متوالی پرداخته است:

کوکسان و همکاران [1] ۲۰۱۱

در این مطالعه، آزمون نهایی سال ۲۰۰۹ دانشکده پزشکی هال یورک [2] که شامل ۱۲ ایستگاه OSCE و ۸ مواجهه با بیمار واقعی (OSLER[3]) بود، به روش سنجش متوالی روی ۱۲۷ دانشجو انجام شد. به این ترتیب که دانشجویانی که عملکردشان در ۶ ایستگاه و ۴ مواجهه واقعی اول، به صورت کاملا رضایت‌بخش ارزیابی شده بود، از ادامه آزمون معاف شدند. بقیه دانشجویان که حدود یک سوم آنها بودند، باید در کل آزمون شرکت می‌کردند و عملکرد آنها که در کل آزمون ملاک تصمیم‌گیری در مورد وضعیتشان بود. قسمت اول آزمون به خوبی عملکرد کلی دانشجویان را پیش‌بینی کرد. میزان صرفه‌جویی در هزینه‌ها معادل ۳۰ هزار پوند بود. هر چند که ذکر کردند کاهش هزینه صرفا به خاطر سنجش متوالی نبود و احتمال داشت ناشی از ترکیب کردن OSCE با OSLER نیز باشد.

1. Cookson et al.
2. Hull York
3. Objective Structured Long Encounter Record.

به صورت کلی، به نظر می‌رسد با برگزاری OSCE به صورت متوالی بتوان ضمن حفظ مشخصاتی مانند پایایی و روایی آزمون، از نظر هزینه‌های مورد نیاز نیز صرفه‌جویی کرد. اما نادیده گرفتن برخی ملاحظات خاص در اجرای آن، صدمات جدی به کیفیت آزمون وارد می‌کند یا حتی ممکن است منجر شود که برگزاری آن از لحاظ هزینه نیز به صرفه نباشد. به این منظور لازم است قبل از اجرای OSCE متوالی، یک آزمون متوالی به صورت فرضی طرح‌ریزی گردد، سپس مشخصات و ویژگی‌های آزمون فعلی با آن مقایسه شود تا مشخص گردد آیا برگزاری OSCE به صورت متوالی موثر و مفید است یا خیر. برخی از موارد چالش‌زا در این خصوص عبارت هستند از:

- تعیین تعداد ایستگاهها برای آزمون غربال
- نحوه انتخاب ایستگاهها برای آزمون غربال
- تعیین حد نصاب قبولی آزمون غربال
- مسائل مربوط به سیاستگزاری، آیین نامه‌ها و اطلاع‌رسانی

پل و همکاران ۲۰۱۳

این مطالعه از داده‌های یک OSCE فارغ‌التحصیلی در انگلستان استفاده کرده بود. به صورت معمول کسانی که در آزمون اصلی رد می‌شوند، پس از گذراندن دوره جبرانی و پس از چند ماه باید در آزمون مجدد شرکت کنند. آزمون اصلی (S_1) که در دو روز برگزار شد، شامل ۱۸ ایستگاه بود و در آن ۲۶۰ دانشجو در چهار مرکز مختلف در دو روز توسط ۵۰۰ آزمونگر مورد ارزیابی قرار گرفتند. از این آزمون به عنوان آزمون غربال در این مطالعه استفاده شد. آزمون مجدد (S_2) نیز شامل ۱۸ ایستگاه بود که به عنوان آزمون تکمیلی در مدل متوالی استفاده شد. به این ترتیب مجموعاً دو آزمون ۳۶ ایستگاه داشتند. مدل متوالی بر اساس تعداد مختلف ایستگاه طرح‌ریزی شد: در مرحله اول تعداد ایستگاه‌های آزمون S_1 و S_2 به ترتیب ۱۶ و ۹ در نظر گرفته شد (مجموعاً ۲۵ ایستگاه). در مرحله بعدی، S_1 شامل ۱۴ و S_2 شامل ۱۱ ایستگاه بود (مجموعاً ۲۵ ایستگاه) و در مرحله بعدی، هر دو از ۱۲ ایستگاه ساخته شدند (مجموعاً ۲۴ ایستگاه). حدنصاب قبولی کل آزمون (یعنی S_1 و S_2) از روش رگرسیون مرزی به اضافه یک SEM به دست آمد. در واقع به اندازه خطای آزمون به حدنصاب اضافه شد تا از قبولی افراد مرزی غیرتوانمند جلوگیری به عمل آید و به اصطلاح موارد قبولی کاذب کاهش یابد. همچنین از مدل نمره‌دهی نیمه‌جبرانی استفاده شد.

به این شکل که دانشجویان علاوه بر کسب حدنصاب قبولی کل آزمون باید در ۶۰ درصد ایستگاه‌ها نیز قبول می‌شدند. برای تعیین حدنصاب آزمون‌های غربال (S۱) تصمیم گرفته شد که نمره به دستآمده از روش رگرسیون مرزی با دو برابر SEM جمع شود تا سختگیری بیشتری در قبولی آزمون غربال لحاظ شود و تمام موارد قبولی کاذب حذف شوند که به نظر نویسندگان نکته کلیدی در سنجش متوالی است و به جلب حمایت ذی‌نفعان و دستاندرکاران کمک می‌کند. البته نویسندگان ذکر کردند که حتی در این حالت نیز احتمال ناچیزی برای بروز قبولی کاذب وجود دارد اما با توجه به خمیدگی نمودار توزیع نمرات در OSCE بسیار نامحتمل است. هنگامی که وضعیت دانشجویان مورد بررسی قرار گرفت مشخص شد اگر آزمون به صورت متوالی برگزار نمی‌شد، ۱۵ نفر (۶/۵ درصد) از آزمون اصلی (۱۸ ایستگاهه) رد می‌شدند و باید پس از چند ماه در آزمون مجدد (۱۸ ایستگاهه) شرکت می‌کردند. در مدل سنجش متوالی، ۳۱ دانشجو (۱۳/۶ درصد) از آزمون غربال ۱۶ ایستگاهه رد شدند. با شرکت در آزمون S۲ و محاسبه جمع عملکرد دانشجویان در دو آزمون اعلام شد که نه نفر از این دانشجوها (۳/۹ درصد) قبول هستند. به این ترتیب کل آزمون در مدل متوالی شش نفر (۲/۶ درصد) ردی داشته است. در این میان یک دانشجو (۰/۴ درصد) هم به صورت کاذب قبول اعلام شده بود. همچنین برای سنجش پایایی OSCE کامل از آلفای کرونباخ استفاده شد. پایایی کل آزمون ۰/۷۹ بود و به صورت قابل‌توجهی در هزینه‌ها صرفه‌جویی شد. به طوری که به جای چهار مرکز برگزاری آزمون، تنها سه مرکز مورد نیاز بود. میزان هزینه که در مدل معمول ۱۲۴ هزار پوند بود، در مدل متوالی حدود ۹۵ هزار پوند برآورد شد که به معنای صرفه‌جویی معادل ۲۹ هزار پوند

کوری و همکاران[1] ۲۰۱۵

در سال ۲۰۱۵ مقاله‌ای منتشر کردند که در آن تاثیر تعداد ایستگاه‌های مختلف و حدنصاب‌های مختلف آزمون غربال را بر پیامدها بررسی کردند. آنها به این منظور از داده‌های OSCE دانشکده پزشکی دانشگاه ابردین[2] در دو سال متوالی (به ترتیب با شرکت ۱۸۵ و ۲۰۳ دانشجو) استفاده کردند. آزمون در هر سال در دو روز متوالی برگزار شد و مجموعاً ۱۵ ایستگاه هشت‌دقیقه‌ای داشت. برای تعیین حدنصاب از روش رگرسیون مرزی استفاده شد که برابر یک SEM به آن اضافه شد و طبق یک مدل نیمه‌جبرانی دانشجویان علاوه بر کسب حدنصاب کل، باید در ۱۰ ایستگاه حدنصاب قبولی لازم را کسب می‌کردند. برای طراحی سنجش متوالی، با استفاده از نرم‌افزار شبیه‌ساز در Stata، پنج هزار OSCE فرضی به عنوان آزمون غربال ساخته شد. هر یک از این آزمون‌ها شامل ۶ تا ۱۴ ایستگاه با انتخاب تصادفی بود و نتیجه عملکرد ۱۰۰ دانشجو را به صورت تصادفی نشان می‌داد. همچنین حدنصاب‌های مختلفی (یک برابر، دو برابر و سه برابر SEM) برای هریک از آزمونهای غربال در نظر گرفته شد. سپس پیامد آنها را در مقایسه با آزمون ۱۵ ایستگاهه اصلی سنجیدند که شامل حساسیت و ویژگی تست بود. شاخص میانه برای حساسیت و ویژگی پنج هزار آزمون غربال محاسبه شد و در هر سال سه نمودار مجزا (به تفکیک سه حدنصاب مختلف) که ویژگی و حساسیت آزمون‌های غربال را نشان میداد، رسم شد. میزان حساسیت آزمون‌های غربال در سال اول بالای ۸۷ درصد و در سال دوم بالای ۸۴ درصد بود و ویژگی بین ۶۰ تا ۱۰۰ درصد متغیر بود. به صورت کلی ویژگی با افزایش تعداد ایستگاه‌ها افزایش یافت و بازه اطمینان ۹۵ درصد آن نیز باریک‌تر شد در حالی که حساسیت بین ۸۴ تا ۹۸ درصد متغیر بود و این یافته‌ها ارتباطی با حدنصاب آزمون نداشت. بیشترین ویژگی در آزمون ۱۱، ۱۳ و ۱۴ ایستگاهه دیده شد اما نویسندگان نتیجه گرفتند که این تعداد ایستگاه از نظر اجرایی به صرفه نیست و آزمون غربال هشت‌ایستگاهه که حساسیت ۸۸ تا ۸۹ درصد و ویژگی ۸۳ تا ۸۶ درصد داشته است، انتخاب مناسبی است.

1. Currie et al
2. University of Aberdeen

مرتاض هجری و همکاران ۲۰۱۶

هدف این مطالعه، طرح‌ریزی مدل بهینه آزمون غربال با استفاده از داده‌های آزمون OSCE پیش‌کارورزی با در نظر گرفتن سه فاکتور اساسی بود. به این منظور خصوصیات سایکومتریک ایستگاه‌های OSCE پیش‌کارورزی دانشگاه علوم پزشکی تهران بر اساس نظریه کلاسیک آزمون و همچنین بر اساس نظریه سؤال پاسخ تعیین گشت. سپس چندین آزمون غربال با تعداد ایستگاه مختلف، خصوصات سایکومتریک متفاوت و حدنصاب‌های قبولی مختلف به صورت فرضی طراحی شدند و نتایج هر یک از آنها با آزمون اصلی مقایسه شد. یک پیامد مطلوب ترکیبی برای هر آزمون غربال تعریف و محاسبه شد که شامل این شاخص‌ها بود: درصد قبولی بیش از ۵۰ درصد، ارزش اخباری مثبت مساوی یک و ارزش اخباری منفی بالای ۰/۲۵.

بر اساس نظریه کلاسیک، تعداد ۲۰ آزمون غربال از ۶۰ آزمون دارای پیامد مطلوب بودند. از نظر حدنصاب می‌توان گفت که ۱۴ مورد از این ۲۰ آزمون دارای حدنصاب سختگیرانه بودند. از نظر تعداد ایستگاه‌ها ۹ آزمون شامل پنج ایستگاه، ۵ آزمون شامل چهار ایستگاه و ۶ آزمون شامل سه ایستگاه بودند. بر اساس نظریه سؤال پاسخ، دو آزمون غربال از ۱۲ آزمون دارای پیامد مطلوب بودند. هر دوی این آزمون‌ها بر اساس پارامتر تمیز تشکیل شده بودند و حدنصاب سختگیرانه داشتند و به ترتیب پنج و چهار ایستگاهه بودند.

نویسندگان نتیجه‌گیری کردند که OSCE متوالی می‌تواند روشی کارآمد و مناسب برای برگزاری OSCE باشد. برای داشتن OSCE غربال موثر با کمترین احتمال خطا، دقت خوب صرفه اقتصادی، انتخاب ایستگاه‌ها بر اساس ضریب تمیز یا ITC (در نظریه کلاسیک) یا پارامتر تمیز (در نظریه سؤال پاسخ) و استفاده از حدنصاب قبولی سختگیرانه باید مدنظر قرار گیرد.

سودمندی OSCE

روایی OSCE

روایی آزمون به این معنا است که آزمون تا چه اندازه واقعاً همان چیزی را ارزیابی کرده که قرار بوده بسنجد. آزمون زمانی روا است که ایستگاه‌ها، نمونه‌ای خوب و معرف از اهداف و توانمندی‌های موردنظر دوره باشند. روایی یک ویژگی ذاتی ابزار نیست. یک ابزار ارزیابی ممکن است در یک موقعیت و شرایط خاص روا محسوب شود اما در شرایط دیگر فاقد روایی مناسب تشخیص داده شود. بنابراین بسته به اینکه OSCE با چه هدفی و برای ارزیابی چه کسانی به کار می‌رود، ممکن است روایی داشته باشد یا نداشته باشد (هاجز[1] ۲۰۰۳).

یک سری اقدامات قبل از برگزاری آزمون لازم است اجرا شود تا روایی خوبی برای آزمون به دنبال داشته باشد. بعد از برگزاری آزمون نیز باید بررسی صوت گیرد و سؤالاتی پرسیده شود تا از روایی OSCE اطمینان یافت. شایان ذکر است که برخی از این سؤالات را می‌توان علاوه بر صاحب‌نظران و مدرسان از خود شرکت‌کنندگان در آزمون نیز پرسید:

- آیا موضوع ایستگاه‌ها بر اساس بلوپرینت و جدول مشخصات آزمون انتخاب شده بودند؟
- آیا آنچه در ایستگاه‌ها خواسته شده بود، مهم و مرتبط با اهداف دوره و توانمندی‌های مورد انتظار بود؟
- آیا سطح ایستگاه‌ها با سطح شرکت‌کنندگان مطابقت داشت؟
- آیا آنچه در ایستگاه‌ها خواسته شده است، آموزش داده شده بود؟
- آیا استادان ذی‌صلاح ایستگاه‌ها را مرور کرده بودند؟
- آیا زمان ایستگاه متناسب با وظیفه خواسته شده بود؟
- آیا موضوع ایستگاه با کار روزمره شرکت‌کنندگان ارتباط داشت؟
- آیا عملکرد بیمارنما قابل باور بود؟

در آزمون‌هایی که برای سنجش مهارت طراحی شده‌اند، یک پیش‌فرض وجود دارد و آن اینکه هر چقدر وظیفه‌ای که در آزمون مورد ارزیابی قرار می‌گیرد، بیشتر شبیه وظیفه فراگیر در دنیای واقعی باشد، آزمون رواتر است. از این حیث آزمون OSCE از بسیاری از آزمون‌های کتبی روایی بهتری دارد اما با وجود اینکه تلاش می‌شود شبیه‌سازی محیط واقعی صورت بگیرد، در عمل چند محدودیت موجب فاصله گرفتن ایستگاه‌ها از دنیای واقعی می‌شوند:

- محدودیت زمان باعث می‌شود در هر ایستگاه فقط یک جنبه ایزوله از مواجهه بالینی مورد ارزیابی قرار گیرد (وس و همکاران ۲۰۰۱، اسمی ۲۰۰۳). این امر موجب می‌شود آزمون، مخصوصاً در حالتی که ایستگاه‌های کوتاه دارد، منعکس‌کننده رابطه واقعی پزشک و بیمار نباشد و روایی پایین داشته باشد. در نتیجه آن را برای یک آزمون تکوینی با هدف بازخورد دادن نامناسب می‌کند. از طرف دیگر، اگر با وجود زمان کم هر ایستگاه، تعداد ایستگاه‌های کل افزایش پیدا کند، از آنجا که می‌توان عملکرد دانشجو را در موقعیت‌های مختلف مشاهده کرد، می‌توان امیدوار بود در قالب یک آزمون تراکمی، تصمیمی که در برای رد یا قبول گرفته می‌شود، پایاتر باشد.
- چک‌لیست‌های OSCE با این فرض نوشته می‌شوند که رابطه پزشک و بیمار را می‌توان به صورت فهرستی از کنش‌ها و واکنش‌ها توصیف کرد. این رویکرد شاید برای ارزیابی دانشجویان کم تجربه مناسب باشد اما هنگامی که تجربه بالینی شرکت‌کنندگان زیاد می‌شود و عملاً با موقعیت‌های پیچیده‌تر سر و کار دارند (مانند OSCE در مقطع دستیاری)، گاهی آیتم‌های چک‌لیست ساده‌انگارانه و نامرتبط به نظر می‌رسند (اسمی ۲۰۰۳).
- گاهی آیتم‌های چک‌لیست آن‌قدر ریز و جزئی نوشته می‌شوند که از مهارت کلی مدنظر فاصله می‌گیرند. در واقع روایی آزمون به خرج افزایش عینیت و پایایی، کاهش پیدا می‌کند (وس و همکاران ۲۰۰۱، نیوبل ۲۰۰۴، پل و همکاران ۲۰۱۰).

1. Hodges

- در OSCE تنها از چیزهایی می‌توان سؤال داد که قابل شبیه‌سازی هستند. در نتیجه عملاً بسیاری از مشکلات بیماران حذف و تعداد معدودی به عنوان نمونه انتخاب می‌شوند. این موضوع هم مخصوصاً هنگامی که تجربه و مقطع دانشجویان بالاتر می‌رود، بیشتر نمایان می‌شود و محتوای ایستگاه‌ها فاصله خود را از دنیای واقعی فراگیران نشان می‌دهد (اسمی ۲۰۰۳).
- این آزمون برای سنجش صلاحیت بالینی دانشجویان در محیطی مصنوعی ساخته شده است. مسائلی مانند رفتارهای حرفه‌ای، نگرش دانشجویان و مراقبت آنها از بیماران در بلندمدت برای OSCE مناسب نیستند و روایی آن را زیر سؤال می‌برند. به همین دلیل، محدود نکردن روش ارزیابی به OSCE و استفاده از ابزارهای متعدد توصیه می‌شود.

بر اساس نتایج یک مطالعه مروری نظام‌مند درباره OSCE، ۶۷ درصد مقالات منتشر شده در مورد OSCE (از ۱۰۴ مقاله) اساساً در مورد روایی آزمونی که برگزار کرده بودند، صحبتی نکرده بودند. پیشنهاد نویسندگان این بود که هر مقاله‌ای که به گزارش اجرای یک OSCE می‌پردازد، باید روایی محتوایی، روایی سازه، روایی همزمان و روایی پیش‌بین آزمون را اندازه‌گیری و گزارش کند (پاتریشیو و همکاران ۲۰۰۹).

برای پاسخ به این سؤال که آیا عملکرد دانشجویان در OSCE می‌تواند موفقیت آتی آنها را پیش‌بینی کند یا ارتباطی با سایر آزمون‌های همزمان آنها دارد، مطالعاتی صورت گرفته است. مقایسه نتایج OSCE با سایر آزمون‌ها از جمله آزمون چندگزینه‌ای، ارزیابی‌های بالینی، آزمون‌های گواهینامه و ... منجر به یافته‌های متناقض شده است. به گونه‌ای که میزان این رابطه از ۰/۱۰ تا ۱/۰۰ گزارش شده است (ترنر و دانکوسکی ۲۰۰۸).

- مطالعه‌ای نشان داد که نتایج OSCE بهترین پیش‌بینی‌کننده عملکرد آتی دانشجویان در دوره دستیاری بود که بعدها توسط سوپروایزرها ارزیابی شده بودند. بالاترین میزان رابطه از ایستگاه مهارت‌های ارتباطی به دست آمد که ۰/۴۲ بود. سایر مهارت‌ها مانند تشخیص افتراقی و تصمیم‌گیری ارتباط کمتری نشان دادند (روتالا و همکاران[1] ۱۹۹۲).
- مطالعه‌ای نشان داده است که نتیجه OSCE ارتباط بالایی با نتایج ارزیابی دستیاران توسط متخصصان دارد. یعنی OSCE ارتباط خوبی با عملکرد پزشک هنگام ارتباط و اداره بیماران، پس از پایان دوره آموزشی دارد (پرابرت و همکاران[2] ۲۰۰۳).
- نتایج یک مطالعه در امریکا نشان داد که عملکرد دانشجویان در OSCE با عملکرد آنها در مرحله اول آزمون USMLE به میزان ۰/۴۱ بوده است (سیمون و همکاران[3] ۲۰۰۲). همچنین در مطالعه دیگری، رابطه نمره OSCE و مرحله دوم USMLE، حدود ۰/۳۹ به دست آمد (سیمون و همکاران ۲۰۰۷).
- مطالعه دیگری در امریکا در یک بررسی ۵ ساله آزمون USMLE طی سال‌های ۲۰۰۷ تا ۲۰۱۱ پرداخت و نتایج OSCE سال دوم را با OSCE سال سوم دانشجویان و با سایر آزمون‌ها مقایسه کرد. یافته‌ها حاکی از آن بود که OSCE ها با یکدیگر و با سایر قسمت‌های USMLE رابطه بالایی نداشتند (دنگ و همکاران[4] ۲۰۱۲).
- در یک مطالعه دیگر در کانادا، ۹۳ درصد دانشجویانی که OSCE را با موفقیت پشت سر گذاشتند، در آزمون گواهینامه پزشکی کانادا[5] نیز قبول شدند. از طرف دیگر، از بین کسانی که در OSCE حدنصاب قبولی را دریافت نکردند، ۶۶ درصد در آزمون گواهینامه قبول شدند (ماویس و هنری[6] ۲۰۰۲).
- در ایران نیز، مطالعه افتخار و همکاران برای سنجش میزان رابطه نتایج OSCE پیش‌کارورزی با معدل دانشجویان و با نمره آزمون کتبی پیش‌کارورزی، روی ۱۰۳ دانشجوی پزشکی سال پنجم انجام شد. یافته‌ها نشان داد که رابطه معدل و نمره کتبی ۰/۷۶ و رابطه معدل و نمره OSCE ۰/۶۸ بود (افتخار همکاران ۲۰۱۲).

در تفسیر نتایج مطالعات فوق، توجه به این نکته حائز اهمیت است که رابطه پایین بین دو آزمون مورد مقایسه لزوماً

1. Rutala et al.
2. Probert et al.
3. Simon et al.
4. Dong et al.
5. Canadian Medical Licensing Examinations
6. Mavis & Henry

به معنای اشکال در OSCE نیست. شاید آزمونی که مقایسه با آن صورت گرفته است، استاندارد نباشد یا اصولاً دو آزمون حیطه‌های متفاوتی را مورد سنجش قرار داده‌اند.

پایایی OSCE

توصیف رایج برای پایایی این است که نتایج چقدر قابل تکرار هستند. به عبارت دیگر، در صورت تکرار اندازه‌گیری تا چه حد نتایج به دست آمده، یکسان و با ثبات هستند. اما شاید توضیح بهتر برای پایایی آزمون این باشد که نتایج به دست آمده چقدر قابل اعتماد هستند. یعنی اگر دانشجویی در آزمون قبول شده است، آیا واقعاً توانمند و شایسته قبولی بوده است؟

آزمون OSCE با توجه به اجزایی که دارد، به صورت کلی آزمونی پایا در نظر گرفته می‌شود اما باید توجه داشت که پایایی خصوصیت ذاتی یک ابزار ارزیابی نیست و به نحوه استفاده و اجرای آزمون برمی‌گردد.

مواردی که باعث کاهش پایایی OSCE می‌شوند (اسمی ۲۰۰۳، ویلکینسون ۲۰۰۳، نیوبل ۲۰۰۴):

- تعداد کم ایستگاه‌ها یا زمان کم آزمون
- یکسان نبودن بازی بیمارنماها یا نقص در سناریوهای آنها
- آزمونگرانی که به صورت غیرقابل پیش‌بینی نمره می‌دهند.
- اشکالات چک‌لیست مانند تعداد بیش از حد آیتم‌های چک‌لیست، جزئی کردن بیش از حد مهارت به ریزمهارت‌های آن، وجود آیتم‌هایی در چک‌لیست که افتراق دهنده نیستند (بسیار آسان یا بسیار دشوار) (توضیحات بیشتر در قسمت طراحی چک‌لیست)
- مشکلات اجرایی مثلاً ایستگاه شلوغ و پر سر و صدا یا به هم ریختگی در کار تیم اجرایی

مواری که به افزایش پایایی OSCE کمک می‌کنند:

- تعداد زیاد ایستگاه‌ها یا زمان طولانی آزمون: از آنجا که پایایی وابسته به مورد است، اگر دانشجو در یک ایستگاه با موضوع معاینه عصبی خوب عمل کند، نمی‌توان آن را به سایر مهارت‌ها مانند معاینه قلبی تعمیم داد. به همین دلیل طرح موارد بالینی متعدد موجب می‌شود بهتر بتوان نتایج آزمون را به سایر موقعیت‌ها تعمیم داد.
- استفاده از چک‌لیست: باعث می‌شود آزمونگران مختلف طبق معیارهای یکسان به ارزیابی دانشجویان مختلفی که در ایستگاه حاضر می‌شوند، بپردازند. مخصوصاً در حالتی که چک‌لیست خوب طراحی شده باشد و آیتم‌های آن بتوانند بین دانشجو با عملکرد قوی و دانشجو با عملکرد ضعیف افتراق قائل شوند، بسیار کمک‌کننده است.
- آموزش بیمارنماها: باعث می‌شود بیمارنماها در برابر دانشجویان مختلف و در ایستگاه‌های مشابه از لاین‌های موازی رفتار یکسانی از خود نشان دهند.
- آموزش آزمونگران: باعث می‌شود آزمونگران در برابر دانشجویان مختلف و در ایستگاه‌های مشابه از لاین‌های موازی رفتار یکسانی از خود نشان دهند. همچنین گفته می‌شود درگیر کردن آزمونگران در روند طراحی و اجرای OSCE به این امر کمک می‌کند.
- تعدد آزمونگران: استفاده از چندین آزمونگر در یک ایستگاه باعث کاهش بروز خطاهای شخصی می‌شود.
- ناهمگونی[1] و عدم تجانس در سطوح توانمندی دانشجویان
- خارج کردن ایستگاه‌های مشکل‌دار بر اساس نتایج تحلیل آزمون
- اعلام نتایج صرفاً به صورت رد و قبول به جای اعلام نمره خام
- ترکیب کردن OSCE با یک آزمون دیگر برای پوشش بهتر محتوای دوره به شرطی که طرح سؤال هر دو بر اساس یک بلوپرینت باشد.

1. Heterogeneity

هدف از برآورد پایایی این است که میزان تاثیر منابع احتمالی خطا بر نمرات دانشجویان سنجیده شود. منابع معمول خطا در OSCE، آزمونگر، ایستگاه و آیتم‌های چک‌لیست هستند. روش‌های متنوعی برای سنجش میزان تاثیر آنها بر اساس نظریه کلاسیک یا نظریه تعمیم‌پذیری پیشنهاد شده است:

- **تاثیر آزمونگران:** اگر دو آزمونگر یا بیشتر در یک ایستگاه به مشاهده عملکرد دانشجو بپردازند، می‌توان از میزان توافق بین آنها به عنوان شاخصی از پایایی نتایج استفاده کرد که از طریق شاخص کاپا قابل محاسبه است. توافق هم می‌تواند در سطح نمره کل ایستگاه مد نظر باشد و هم در مور تک تک آیتم‌ها بررسی شود. مثلاً اینکه آیا آزمونگران با هم توافق داشتند که دانشجو با بیمار با احترام برخورد کرد؟ مسأله این است که از آنجا که در OSCEهای معمول فقط یک آزمونگر در هر ایستگاه حضور دارد، توافق بین آزمونگران بی معنی است و قابل محاسبه نیست.
- **تاثیر ایستگاه‌ها:** برای سنجش تاثیر ایستگاه‌های مختلف در یک آزمون، می‌توان از شاخص ثبات درونی استفاده کرد که با آلفای کرونباخ محاسبه می‌شود. توصیه می‌شود که برای توصیف این نحوه محاسبه، از عبارت «پایایی کل»[1] (پاتریشیو و همکاران ۲۰۰۹) یا «آلفا در سطح ایستگاه‌ها»[2] (برانیک و همکاران[3] ۲۰۱۱) استفاده شود. گفته می‌شود آلفا به دو دلیل برای تخمین پایایی OSCE مناسب نیست: اولاً تنها یک متغیر و منبع خطا را بررسی می‌کند و ثانیاً به جای آنکه آزمون را به صورت معیاری در نظر بگیرد و توانایی واقعی هر دانشجو را ملاک قرار دهد، آزمون را هنجاری فرض کرده و رتبه دانشجویان نسبت به یکدیگر را مورد بررسی قرار می‌دهد. با این وجود، آلفا رایج‌ترین شاخصی است که برای ارزیابی پایایی OSCE به کار می‌رود (برانیک و همکاران ۲۰۱۱).
- تاثیر آیتم‌های چک‌لیست: آلفای کرونباخ به عنوان شاخص ثبات درونی برای سنجش همبستگی بین آیتم‌های یک ایستگاه با یکدیگر نیز به کار می‌رود. در این حالت، برای هر ایستگاه یک آلفا محاسبه می‌شود و سپس از آلفای ایستگاه‌های مختلف میانگین گرفته می‌شود تا عدد مربوط به کل آزمون به دست بیاید. توصیه می‌شود برای عدم تداخل این حالت با مفهوم قبلی (پایایی کل) از عبارت دیگری استفاده شود. پاتریشیو و همکاران به سادگی از عبارت ثبات درونی یا آلفای کرونباخ استفاده می‌کنند و برانیک و همکاران، عبارت «آلفا در سطح آیتم‌ها»[4] را به کار می‌برند.
- تمام منابع خطا: با استفاده از چارچوبی که نظریه تعمیم‌پذیری فراهم می‌کند، می‌توان تمام منابع را به صورت همزمان بررسی کرد و سهم هر یک را در واریانس نمرات برآورد نمود.

استفاده از شاخص‌های ذکر شده به منظور پایش و ارتقای کیفیت OSCE توصیه شده است (GMC ۲۰۰۹، پل و همکاران ۲۰۱۰). اما در مطالعه مروری نظام‌مندی که به بررسی ۱۰۴ مقاله پرداخته بود، ۶۳ درصد مقالات در مورد پایایی OSCE برگزار شده، اطلاعاتی ارائه نداده بودند و در موارد باقی‌مانده تنوع روش‌های اندازه‌گیری و گزارش پایایی و همچنین نام‌گذاری آنها آنقدر زیاد بود که جمع‌بندی را با مشکل مواجه می‌کرد (پاتریشیو و همکاران ۲۰۰۹).

همچنین در مطالعه مروری دیگری که در مورد پایایی آزمون های OSCE انجام شد، نویسندگان به این نتیجه رسیدند که هرچند OSCE روشی استاندارد و نسبتاً عینی برای سنجش مهارت‌های بالینی است، استفاده از آن لزوماً پایایی بالای نمرات و تصمیم‌گیری‌ها را تضمین نمی‌کند. همچنین ارزیابی مهارت‌های ارتباطی به صورت پایا دشوارتر از ارزیابی مهارت‌های بالینی است (برانیک و همکاران ۲۰۱۱).

1. Total reliability
2. Alpha across stations
3. Brannick et al.
4. Alpha across items

برانیک و همکاران ۲۰۱۱

در این مطالعه مروری نظام‌مند، متون و مقالاتی که ضریب تعمیم‌پذیری یا آلفای کرونباخ را در سطح آیتم‌های ایستگاه یا در سطح ایستگاه‌های کل آزمون OSCE گزارش کرده بودند، جمع‌آوری شدند و روی داده‌های به دست آمده در مورد ضریب آلفا متاآنالیز انجام شد. مجموعاً ۳۰ مطالعه به دست آمد که ۱۸۸ آلفا گزارش کرده بودند.

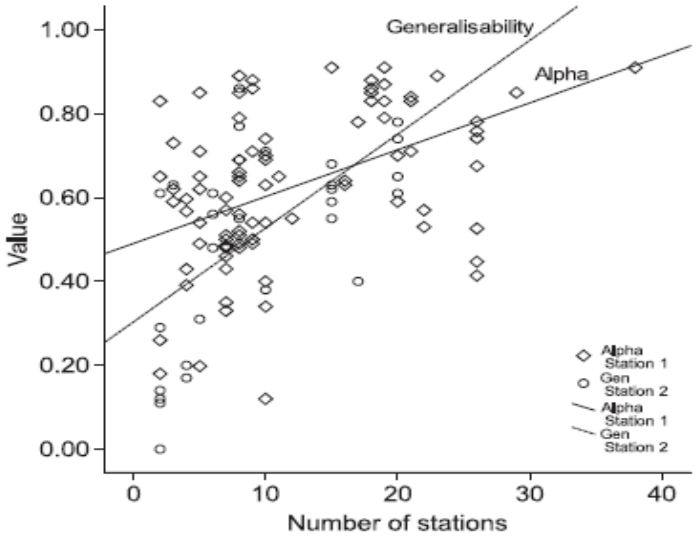
آلفای کل در سطح ایستگاه‌ها ۰/۶۶ (بازه اطمینان ۹۵ درصد ۰/۶۲ تا ۰/۷۰) و آلفای کل در سطح آیتم‌های ایستگاه‌ها ۰/۷۸ (بازه اطمینان ۹۵ درصد ۰/۷۳ تا ۰/۸۲) به دست آمد. به طوری که نویسندگان نتیجه‌گیری کردند که غالباً پایایی نمره کل آزمون چندان بالا نیست ولی در سطح آیتم‌ها قابل قبول است. همچنین مشخص شد در سطح ایستگاه‌ها، ضریب پایایی مهارت‌های بالینی بالاتر از پایایی مهارت‌های بین فردی است (۰/۶۹ در برابر ۰/۵۵) در حالی که در سطح آیتم‌ها برعکس است (۰/۷۵ در برابر ۰/۸۸).

میانگین ضریب آلفای ۲۱ مورد که از مقیاس لیکرت استفاده کرده بودند، در سطح ایستگاه‌ها و در سطح آیتم‌ها به ترتیب ۰/۵۹ و ۰/۸۸ به دست آمد؛ در حالی که میانگین ضریب آلفای ۲۸ مورد که از ارزیابی توسط چک‌لیست استفاده کرده بودند، در سطح ایستگاه‌ها و در سطح آیتم‌ها به ترتیب ۰/۶۹ و ۰/۶۷ بود. در تفسیر این داده‌ها توجه به این نکته ضروری است که اکثر ایستگاه‌های مهارت ارتباطی توسط مقیاس لیکرت مورد سنجش قرار گرفته بودند و اکثر ایستگاه‌های مهارت بالینی توسط چک‌لیست.

پایایی بالای متوسط آزمون با افزایش تعداد ایستگاه‌ها، افزایش تعداد آیتم‌ها و افزایش تعداد آزمونگران در هر ایستگاه رابطه داشت اما نویسندگان ذکر کردند باید توجه داشت که افزایش آیتم‌های یک ایستگاه باعث بروز تکرار[1] می‌شود که بدون افزایش دقت اندازه‌گیری فقط ظاهرا ضریب پایایی را بالا می‌برد. همچنین افزایش تعداد ایستگاه‌ها علاوه بر اینکه هزینه‌بر است، لزوماً و خود به خودی منجر به افزایش پایایی نمی‌شود. کمااینکه در آزمون‌هایی با تعداد ایستگاه زیاد، همچنان واریانس نمرات در آزمون‌های مختلف متغیر بود.

1. Redundancy

از طرفی، آزمون‌هایی که تعداد ایستگاه بیشتر داشتند یا در هر ایستگاه از دو آزمونگر استفاده می‌کردند، ضرایب آلفای بزرگتری داشتند اما همچنان تفاوت زیادی در ضرایب آنها مشهود بود (شکل ۴–۱۸) (برانیک و همکاران ۲۰۱۱).



شکل ۴–۱۸: اسکاترپلات ضریب تعمیم پذیری و ضریب آلفا در OSCEها با تعداد ایستگاه‌های متفاوت (برانیک و همکاران ۲۰۱۱)

تاثیر آموزشی OSCE

آزمون OSCE مانند سایر روش‌های ارزیابی دانشجو نه تنها به ارزیابی میزان یادگیری فراگیران می‌پردازد، بلکه تاثیر به سزایی در یادگیری آنان نیز دارد. تاثیری که آزمون بر رفتار دانشجویان می‌گذارد، می‌تواند مثبت یا منفی باشد. به عنوان

مثال، جلب توجه دانشجویان به اهمیت مهارت‌های عملی و بالینی از جمله موارد تاثیر مثبت این آزمون است اما تمایل دانشجویان به حفظ کردن آیتم‌های چک‌لیست به جای درک عمیق آنها یکی از پیامدهای منفی آن است (بورسیکوت و همکاران ۲۰۱۱). استفاده از رویکرد قطعه‌قطعه‌سازی و نمره‌دهی درجه‌ای که در قسمت طراحی چک‌لیست توضیح داده شد، این مزیت را دارد که دانشجو نمی‌تواند صرفاً با حفظ کردن آیتم‌ها نمره خوبی دریافت کند.

مسأله دیگری که تاثیر آموزشی OSCE را جهت‌دهی می‌کند، این موضوع است که امکان طرح سؤال از برخی از موارد بالینی وجود ندارد. به عنوان نمونه، برخی از علایم بالینی قابل تقلید توسط بیمارنماها نیستند و چنانچه از بیمار واقعی استفاده شود، باز هم استاندارد کردن آنها دشوار است. این مسأله موجب می‌شود دانشجویان قادر به پیش‌بینی موارد بالینی طرح شده در OSCE شوند و موارد دیگر را یاد نگیرند (گرملی و همکاران ۲۰۱۱).

مطالعات به بررسی ارتباط خصوصیات و زمینه قبلی دانشجویان، تجربیات بالینی و سبک یادگیری[1] آنها با میزان موفقیت در OSCE پرداخته‌اند:

- در یک مطالعه آینده‌نگر روی دانشجویان پزشکی سال یک مشخص شد که سبک یادگیری عمیق و ساختارمند[2] با عملکرد دانشجویان در OSCE مرتبط است اما تجربه بالینی این طور نیست (مارتین و همکاران ۲۰۰۰).
- مطالعه دیگر به این نتیجه رسید که محیط و زمینه‌ای که امتحان در آن برگزار می‌شود، مثلاً چرخش‌های بالینی اخیری که دانشجو پشت سر گذاشته است، بر نحوه رویکرد دانشجو به بیمار در جلسه OSCE تاثیر می‌گذارد و همچنین دانشجویی که بخش‌های بالینی بیشتری را گذرانده است، شرح حال گسترده‌تری می‌گیرد (بلاسکویچ و همکاران[3] ۲۰۰۴).
- مهارت‌های ارتباطی نوشتاری مانند نگارش خلاصه شرح حال، با مهارت‌های ارتباطی کلامی ارتباط داشتند اما با مهارت دانشجویان در معاینه فیزیکی ارتباطی نشان ندادند (کیلی و همکاران[4] ۲۰۰۲).

هزینه و قابلیت اجرای OSCE

قابلیت اجرای OSCE به مسایل مختلفی وابسته است: تمایل مسؤولان اجرایی دانشکده، همکاری هیأت علمی، توانمندی هیأت علمی، در دسترس بود بیمارنمایان، زمان، تسهیلات و امکانات از جمله فضای مناسب آزمون و البته منابع مالی.

آزمون OSCE پرهزینه است و نیاز به امکانات و منابع مالی و انسانی متعدد دارد. قسمتی از منابع مالی صرف اجرای آزمون می‌شود؛ مانند اجاره محل آزمون، اقدامات مربوط به تامین حفاظت آزمون، وسایل مصرفی درون ایستگاه‌ها و پرداخت حق‌الزحمه آزمونگران، بیمارنماها و کارمندان و ... قسمت دیگری از هزینه‌ها به اقدامات قبل از شروع آزمون یعنی طراحی آزمون و آمادگی و هماهنگی آن مربوط می‌شوند؛ به عنوان مثال طراحی سؤالات و سناریوها، تدوین چک‌لیست‌ها، چاپ و تکثیر برگه‌ها و توانمندسازی و آموزش آزمونگران و بیمارنماها و ...

برای سازمان‌دهی موارد فوق چارچوب‌های متعددی در مقالات پیشنهاد شده است:

- چهار مرحله شامل هزینه مربوط به طراحی، تولید، اجرا و گزارش (رزنیک و همکاران[5] ۱۹۹۳)
- سه بخش شامل هزینه مربوط به پرسنل، هزینه مربوط به بیمارنما و هزینه‌های اجرایی (کارپنتر[6] ۱۹۹۵)
- دو قسمت شامل هزینه‌های مستقیم (ماتریال و مواد مصرفی) و هزینه‌های غیرمستقیم یا پنهان (ساعت کاری و پرداخت حقوق) (پونارو و همکاران[7] ۱۹۹۷)

1. Learning style
2. Well-organized deep learning style
3. Blaskiewicz et al.
4. Keely et.al
5. Reznick et al.
6. Carpenter
7. Poenaru et al

هر چند موضوع تأمین هزینه در OSCE همیشه مورد توجه بوده است، مقالات اندکی به این مقوله پرداخته‌اند. در یک مطالعه مروری نظام‌مند که با هدف بررسی میزان قابلیت اجرای OSCE انجام شد، تنها ۱۹ مقاله (چهار درصد کل مقالات مورد بررسی) یافت شد که داده‌های مربوط به هزینه و زمان صرف شده را به صورت کمی گزارش کرده بودند. برخی از این مقالات فقط هزینه مستقیم را محاسبه کرده بودند و برخی دیگر، هزینه‌های پنهان و غیرمستقیم را نیز به حساب آورده بودند.

همچنین، برخی از مقالات هزینه کل آزمون را ارائه داده بودند، تعدادی از مطالعات هزینه صرف شده به ازای یک دانشجو[1] و بعضی دیگر، هزینه صرف‌شده به ازای یک دانشجو در یک ایستگاه[2] را بررسی کرده بودند. شاخص آخر، که به نظر می‌رسد مبنای بهتری برای مقایسه بین آزمون‌های مختلف به دست دهد، در مقالاتی که فقط هزینه مستقیم را حساب کرده بودند، بین ۰/۸۸ تا ۶/۹ دلار امریکا متغیر بود. در مقالاتی که هزینه غیرمستقیم را نیز محاسبه کرده بودند، این دامنه بین ۱۳/۶۱ تا ۵۰ دلار بود (پاتریشیو و همکاران ۲۰۱۳). هرچند ذات این موضوع که اصولاً در تخمین هزینه‌ها چه مواردی باید مورد محاسبه قرار گیرند محل مناقشه است و باعث می‌شود استفاده از این آمار محدود شود.

سؤال این است که آیا صرف این هزینه‌ها منطقی است و در مقابل فوایدی به دست می‌آید که در مجموع بتوان گفت برگزاری OSCE هزینه–اثربخش است؟ پاتریشیو و همکاران در مقاله مروری خود ذکر می‌کنند که نباید به خاطر هزینه‌ها از تمام پیامدها و نتایج مثبتی که صرفاً با برگزاری یک امتحان به دست می‌آید، صرف نظر کرد. در مقاله مروری نظام‌مند آنها از بین ۱۹ مقاله که به هزینه OSCE پرداخته بودند، ۱۴ نویسنده عقیده داشتند علی رغم هزینه‌ها برگزاری OSCE باید ادامه پیدا کند (پاتریشیو و همکاران ۲۰۱۳).

همان‌طور که ذکر شد، تمام هزینه‌های OSCE به صورت مالی قابل ارزش‌گذاری نیستند. به عنوان مثال، وقتی که اعضای هیأت علمی برای طراحی سناریوها و همچنین شرکت در جلسه آزمون صرف می‌کنند، عامل مهمی است که باید به آن توجه شود. با درنظر گرفتن فایده و سودی که حضور هیأت علمی به صورت بالقوه در بخش‌های دیگر می‌تواند داشته باشد، صرفاً اعداد و رقم‌های پرداختی نمی‌تواند ملاک باشد.

تعداد مطالعاتی که در مورد زمان صرف شده توسط افراد درگیر در OSCE انجام شده‌اند، زیاد نیست. در مقاله مروری نظام‌مند مذکور هفت مقاله به این مقوله پرداخته‌اند. در یک مطالعه اشاره شده است که تیم اجرایی ده هفته تلاش کرده تا OSCE اجرا شود (باتلز و همکاران[3] ۱۹۹۲). نویسندگان مقاله دیگری ذکر کرده‌اند که فعالیت‌ها ۶ ماه قبل از تاریخ امتحان آغاز گشته است (کوزیمانو و همکاران[4] ۱۹۹۴). مطالعه دیگری زمان صرف شده را به ازای هر دانشجو محاسبه کرده و آن را در یک آزمون چهارساعته با شرکت ۴۲ دانشجو، ۳/۷۵ ساعت برآورد کرده است (هاجز و لفچی[5] ۱۹۹۷). جدول ۱۲–۱۸ ساعات تخمینی برای برگزاری یک OSCE با هشت ایستگاه را نشان می‌دهد (کیسی و همکاران ۲۰۰۹).

1. Per student
2. Per student per station
3. Battles et al
4. Cusimano et al.
5. Hodges & Lofchy

جدول ۱۲-۱۸: تخمین ساعات لازم برای برگزاری OSCE ۸ ایستگاهه به تفکیک وظایف مختلف دست‌اندرکاران

افراد	ساعت	وظایف
مدیر برنامه	۷۰ ساعت	انتخاب ایستگاه‌ها مرور سناریوها و چک‌لیست‌ها دعوت از هیأت علمی و هماهنگی با آنها حضور سر جلسه آزمون نمره‌دهی نهایی
	۷۵ ساعت	آماده‌سازی و برگزاری کارگاه نحوه طراحی ایستگاه برای هیأت علمی
هماهنگ کننده آزمون	۱۶۵ ساعت	آماده کردن ایستگاه‌ها آماده کردن ایستگاه‌ها هماهنگی برگزاری آزمون حضور سر جلسه آزمون
هیأت علمی	۵۶ ساعت	طراحی ایستگاه‌ها حضور سر جلسه آزمون
	۸ ساعت	شرکت در کارگاه نحوه طراحی ایستگاه
بیمارنما	۶۰ ساعت	آمادگی برای ایفای نقش حضور در ایستگاه‌ها
	۱۵ ساعت	شرکت در کارگاه
مسؤول بیمارنما	۲۲۰ ساعت	مرور و پایلوت ایستگاه‌ها هماهنگی با بیمارنماها حضور سر جلسه آزمون
	۱۵ ساعت	کارگاه برای بیمارنماها
پرسنل اجرایی	۳ ساعت	ورود نمرات و آنالیز آماری

هرچه تعداد ایستگاه‌ها بیشتر باشد، امکانات و منابع بیشتری مورد نیاز است و باید از تعداد بیشتر بیمارنما و آزمونگر دعوت به عمل آید (اسمی ۲۰۰۳). به نظر می‌رسد هزینه برگزاری OSCE مستقل از تعداد دانشجویان شرکت کننده در امتحان باشد. برگزاری OSCE دو بار در یک روز، جدا از وقتی که توسط هیأت علمی صرف می‌شود، فقط اندکی هزینه را افزایش می‌دهد (اسمی ۲۰۰۳).

با توجه به مشکلات دانشکده‌ها در تامین منابع مالی و تلاش برای صرفه‌جویی، به حداقل رساندن هزینه‌ها و بهبود هزینه-اثربخشی امتحان OSCE مهم است. اولاً توجه به این نکته حائز اهمیت است که برگزاری OSCE برای بار اول هزینه، زمان و انرژی زیادی می‌برد ولی تمام این موارد در دفعات بعدی به مراتب کمتر می‌شود. ثانیاً یکی از راه‌کارهای پیشنهادی برای کاهش هزینه آزمون OSCE که در حال حاضر بیشتر کاربرد پژوهشی دارد، برگزاری آن به صورت سنجش متوالی است که قبلاً توضیح داده شد. نکته آخر اینکه لزومی ندارد این آزمون حتماً در پیشرفته‌ترین شکل ممکن برگزار می‌شود. می‌توان OSCE را به شکلی طراحی و اجرا کرد که در مجموع هزینه کمتری در برداشته باشد. اجزای مختلف OSCE در دو مدل آزمون ارزان و گران به صورت جدول ۱۳-۱۸ مقایسه شده است (کیسی و همکاران ۲۰۰۹):

جدول ۱۳-۱۸: مقایسه اجزای مختلف OSCE در دو مدل پرهزینه و کم هزینه

اجزا	OSCE با صرف هزینه بالا	OSCE کم هزینه با قابلیت اجرای بالا
تدوین بلوپرینت	هیأت علمی مجرب	هیأت علمی مجرب
طراحی ایستگاه	هیأت علمی مجرب	دستیار/پرستار/دانشجوی پزشکی سال بالا و مرور توسط هیأت علمی
آزمونگر	هیأت علمی مجرب	دستیار/پرستار/دانشجوی پزشکی سال بالا/ بیمارنما
فضای آزمون	مرکز مهارت‌ها یا مرکز شبیه‌سازی مجهز	اتاق‌های اداری، ضبط در صورت امکان
ضبط	مرکزی و دیژیتال	ویدئو
بیمارنما	آموزش‌دیده با پرداخت مشخص	داوطلب
ارزشیابی	کامپیوتری	کاغذی
جلسه توجیهی	هیأت علمی مجرب انفرادی	دستیار/پرستار/دانشجوی پزشکی سال بالا/بیمارنما گروهی یا فقط به دانشجویان ضعیف

مقبولیت OSCE

مهم است که آزمون از طرف تمام ذی‌نفعان پذیرفته شود. بنابراین جمع‌آوری بازخورد و نظرات دانشجویان، آزمونگران، طراحان سؤال و بیمارنماهایی که درگیر آزمون بوده‌اند، اهمیت دارد.

اگر OSCE برای اولین بار اجرا می‌شود، باید مراقب بود و دانشجویان را از قبل آماده کرد زیرا تحت نظر بودن هنگام انجام کار برای بسیاری از دانشجویان تهدیدکننده و استرس‌زا است. توضیح دادن هدف و روند آزمون در پذیرش دانشجویان تاثیرگذار است.

از طرفی، گفته می‌شود آزمون OSCE به علت امکان برقراری عدالت که ناشی از یکسان بودن سؤالات برای همه است، مقبولیت خوبی نزد دانشجویان دارد. این مسأله مخصوصاً در مقایسه با آزمون Long Case که هر دانشجو در مواجهه با بیمار متفاوتی مود ارزیابی قرار می‌گیرد، قابل توجه است (گرملی ۲۰۱۱). اما در مواردی که به علت تعداد بالای دانشجویان، آزمون در لاین‌های موازی یا نوبت صبح و عصر اجرا می‌شود، امکان مخدوش شدن مسأله فوق وجود دارد. در این موارد ضروری است که دست‌اندرکاران به آموزش، توجیه و پایش آزمونگران و بیمارنماها توجه کافی داشته باشند (GMC ۲۰۰۹).

منابع

1. AoME 2012. Academy of Medical Educators. Professional standards. London: Academy of Medical Educators; 2012.
2. Battles JB, Sprankell SJ, Carpenter JL, Bedford JA, Kirk LM. Developing a support system for teaching and assessing clinical competence. J Biocommun 1992;19(4):19–25.
3. Blaskiewicz RJ, Park RS, Chibnall JT, Powell JK. The influence of testing context and clinica rotation order on students' OSCE performance. Acad Med 2004;79:597-601.
4. Brannick MT, Erol-Korkmaz HT, Prewett M. A systematic review of the reliability of objective structured clinical examination scores. Medical Education 2011;45(12):1181–1189
5. Brown g, Manogue M, Martin M. The validity and reliability of an OSCE in dentistry. European Journal of Dental Education 1999;3(3):117–125
6. Carpenter JL. Cost analysis of objective structured clinical examinations. Acad Med 1995; 70(9):828–833.
7. Cass A, Regehr G, Reznick R, Rothman A, Cohen R. Sequential Testing in the Objective Structured Clinical Examination: Selecting Items for the Screen. Academic Medicine 1997:72; S25-7
8. Centre for Medical Education. Queen's University Belfast [Internet]. OSCE examiner training and development. c2010. Queen's University Belfast. Available online from: www.med.qub.ac.uk/OSCE
9. Cohen R, Reznick R, Taylor B, Provan J, Rothman A. Reliability and validity of the objective structured clinical examination in assessing surgical residents. Am J Surg 1990;160:302–305.
10. Chenot JF, Simmenroth-Nayda A, Koch A, et al. Can student tutors act as examiners in an objective structured clinical examination? Med Educ 2007;41:1032-8.
11. Colliver JA, Barrows HS, Vu NV, Verhulst SJ, Mast TA, Travis TA. Test security in examinations that use standardized-patient cases at one medical school. Acad Med 1991; 66(5):279-82.
12. Colliver JA, Mast TA, Vu NV, Barrows HS. Sequential testing with a performance based examination using standardized patients. Academic Medicine 1991; 66 (10) suppl: S64-S66
13. Colliver JA, Vu NV, Barrows HS. Screening test length for sequential testing with a standardized-patient examination: a receiver operating characteristic (ROC) analysis. Academic Medicine 1992; 67:592-5.
14. Colliver JA, Markwell SJ, Travis TA, Schrage JP & Nu NV. Sequential testing with a standardized-patient examination: A ROC analysis of the effects of case-total correlations and difficulty levels of screening test cases. In Rothman AI & Cohen R; 1994;170–173. Toronto: Proceedings of the Sixth Ottawa Conference on Medical Education.
15. Cookson J, Crossley J, Fagan G, McKendree J, Mohsen A. A final clinical examination using a se-

quential design to improve cost-effectiveness. Medical Education 2011;45: 741-7

16. Currie GP, Sivasubramaniam S, Cleland d. Sequential testing in a high state OSCE: Determing number of screening teste. Medical teacher 2016.
17. Cusimano MD, Cohen R, Tucker W, Murnaghan J, Kodama R, Reznick R. A comparative analysis of the costs of administration of an OSCE. Acad Med 1994;69(7):571–576.
18. Dong T, Saguil A, Artino AR Jr, et al. Relationship between OSCE scores and other typical medical school performance indicators: a 5-year cohort study. Mil Med 2012;177(9 Suppl):44-6.
19. Eftekhar H, Labaf A, Anvari P, Jamali A, Sheybaee-Moghaddam F. Association of the pre-internship objective structured clinical examination in final year medical students with comprehensive written examinations. Med Educ Online 2012;17
20. GMC 2009. General Medical Council. Tomorrow's doctors: outcomes and standards for undergraduate medical education. 2nd ed. London: General Medical Council; 2009.
21. GMC 2009 supplementary. General Medical Council. Assessment in undergraduate medical education: advice supplementary to Tomorrow's Doctors. London: General Medical Council; 2009. Available online from: http://www.gmc-uk.org/Assessment_in_undergraduate_web.pdf_38514111.pdf
22. Gormley G. Summative OSCEs in undergraduate medical education. Ulster Med J 2011;80(3):127–132.
23. Gormley GJ, McCusker D, Booley MA, McNeice A. The use of real patients in OSCEs: a survey of medical students' predictions and opinions. Med Teach 2011;33(8):684.

Grand'Maison P, Brailovsky CA, Lescop J. Content validity of the Quebec licensing examination Can Fam Physician 1996;42:254-9. .(OSCE). Assessed by practising physicians

24. Gupta P, Dewan P, Singh T. Objective Structured Clinical Examination (OSCE) Revisited. Indian Pediatr 2010;47(11):911-20.
25. Harden RM. How to Assess Clinical Competence – An overview. Med Teach 1979;1:289–296.
26. Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). Med Educ 1979;1)13): 41-54.
27. Harden RM. Twelve tips for organizing an Objective Structured Clinical Examination (OSCE). Medical teacher 1990;12(3-4):259-264
28. Hodges B. Validity and the OSCE. Med Teach 2003;25(3):250-4.
29. Hodges B, Lofchy J. Evaluating psychiatric clinical clerks with a mini-objective structured clinical examination. Acad Psychiatry 1997;21(4):219–225.
30. Hodges BD, McNaughton N. Who should be an OSCE examiner? Acad Psychiatry 2009;33(4):282-4
31. Homer M, Pell G. The impact of the inclusion of simulated patient ratings on the reliability of OSCE assessments under the borderline regression method. Med Teach 2009; 31(5): 420-5.
32. Jalili M, Mortaz tlejri S, what is an optimal sequantial OSCE Model? Medical Teacher 2016

33. Humphrey-Murto S, Smee S, Touchie C, Wood TJ, Blackmore D. A Comparison of Physician Examiners and Trained Assessors in a High-Stakes OSCE Setting. Acad Med 2005;80:S59–S62.
34. Jefferies A, Simmons B, Tabak D, McIlroy JH, Lee KS, Roukema H, Skidmore M. Using an objective structured clinical examination (OSCE) to assess multiple physician competencies in postgraduate training. Med Teach 2007;29(2-3):183-91.
35. Joorabchi B. Objective structured clinical examination in a pediatric residency program. Am J Dis Child 1991;145(7):757-62.
36. Khan KZ, Ramachandran S, Gaunt K, Pushkar P. The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part I: An historical and theoretical perspective. Med Teach 2013;35(9):e1437-46. doi:10.3109/0142159X.2013.818634.
37. Khan KZ, Gaunt K, Ramachandran S, Pushkar P. The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part II: Organisation & Administration.Med Teach 2013;35(9):e1447-63. doi: 10.3109/0142159X.2013.818635.
38. Keely E, Myers K, Dojeiji S. Can written communication skills be tested in an objectivestructured clinical examination format? Acad Med 2002;77:82-6.
39. Mann KV, Macdonald AC, Nornici JJ. Reliability of objective structured clinical examinations: Four years of experience in a surgical clerkship. Teach Learn Med 1990;2:219–224.
40. Martin IG, Stark P, Jolly B. Benefiting from clinical experience: the influence of learning style and clinical experience on performance in an undergraduate objective structured clinical examination. Med Educ 2000;34:530-4.
41. Mavis BE, Henry RC. Between a rock and a hard place: finding a place for the OSCE in medical education. Med Educ 2002;36:408-9.
42. Mortaz Hejri S, Yazdani K, Labaf A, Norcini JJ, Jalili M, Introducing a model for optimal design of seqvential objective strvetured clinical examinations. Advances in Health Sciencs Education 2016.
43. Mirzazadeh A, Hejri SM, Jalili M, Asghari F, A Labaf, et al. Defining a Competency Framework: The First Step toward Competency-Based Medical Education. Acta Medica Iranica 2014;52(9):210-216
44. Mitchell ML, Henderson A, Groves M, Dalton M, Nulty D. The objective structured clinical examination (OSCE): optimising its value in the undergraduate nursing curriculum. Nurse Educ Today 2009;29(4):398-404.
45. Muijtjens AMM, van Vollenhoven FHM, van Luijk SJ, van der Vleuten PM. Sequential Testing in the Assessment of Clinical Skills. Academic Medicine 2000;75:369-373.
46. Nestel D, Kneebone R. Perspective: authentic patient perspectives in simulations for procedural and surgical skills. Acad Med 2010; 85(5):889-93.
47. Newble DI. Eight years experience with a structured clinical examination. Medical Education 1988 22(3):200–204.
48. Newble D. Techniques for measuring clinical competence: objective structured clinical examina-

tions. Med Educ 2004;38(2):199-203.

49. Patil NG, Saing H, Wong J. Role of OSCE in evaluation of practical skills. Med Teach 2003;25(3):271-2.
50. Patricio M, Juliao M, Fareleira F, Young M, Norman G, Vaz Carneiro A. A comprehensive checklist for reporting the use of OSCEs. Med Teach 2009;31(2):112-24.
51. Pell G, Fvller R, Homer M, Robents T, Advancing the objective structuned clinical examination: seqvential testing in theory and practice Medical Education 2013
52. Patrício MF, Julião M, Fareleira F, Carneiro AV. Is the OSCE a feasible tool to assess competencies in undergraduate medical education? Med Teach 2013;35(6):503-14
53. Payne NJ, Bradley EB, Heald EB, et al. Sharpening the eye of the OSCE with critical action analysis. Acad Med 2008;83(10):900-5
54. Pell G, Fuller R, Homer M, Roberts T. How to measure the quality of the OSCE: A review of metrics – AMEE guide no.49. Med Teach 2010; 32(10): 802-11
55. Pierre RB, Wierenga A, Barton M, Branday JM, Christie CD. Student evaluation of an OSCE in paediatrics at the University of the West Indies, Jamaica. BMC Med Educ 2004;16(4):22.
56. Prislin MD, Giglio M, Lewis EM, Ahearn S, Radecki S. Assessing the acquisition of core clinical skills through the use of serial standardized patient assessments. Academic Medicine 2000;75(5):480–483.
57. Poenaru D, Morales D, Richards A, O'Connor M. Running an objective structured clinical examination on a shoestring budget. Am J Surg 1997;173(6):538–541.
58. Probert CS, Cahill DJ, McCann GL, Ben-Shlomo Y. Traditional finals and OSCEs in predicting consultant and self-reported clinical skills of PRHOs: a pilot study. Med Educ 2003;37:597-602
59. Reznick RK, Smee S, Baumber JS, Cohen R, Rothman A, Blackmore D, Berard M. Guidelines for estimating the real cost of an Objective Structured Clinical Examination. Acad Med 1993;68(7):513–517.
60. Rothman AI, Blackmore DE, Dauphinee WD, Reznick R. Tests of Sequential Testing in Two Years› Results of Part 2 of the Medical Council of Canada Qualifying Examination. Academic Medicine 1997:72; S22-4
61. Rushforth HE. Objective structured clinical examination (OSCE): review of literature and implications for nursing education. Nurse Educ Today 2007;27(5):481-90.
62. Rutala PJ, Fulginiti JV, Mcgeagh AM, Leko EO, Koff NA, Witzke DB. Predictive validity of a required multidisciplinary standardized-patient examination. Academic Medicine 1992;67:S60–S62.
63. Schoonheim-Klein M, Muijtjens A, Habets L, et al. On the reliability of a dental OSCE, using SEM: effect of different days. Eur J Dent Educ 2008;12(3):131-7.
64. Schoonheim-Klein M, Walmsley AD, Habets L, van der Velden U, Manogue M. An implementation strategy for introducing an OSCE into a dental school. Eur J Dent Educ 2005;9(4):143-9.

65. Selby C, Osman L, Davis M, Lee M. Set up and run an objective structured clinical exam. BMJ 1995;310(6988): 1187–1190.
66. Simon SR, Volkan K, Hamann C, Duffey C, Fletcher SW. The relationship between second-year medical students' OSCE score and USMLE Step 1 scores. Med Teach 2002;24:535-9.
67. Simon SR, Bui A, Day S, Berti D, Volkan K. The relationship between second-year medical students' OSCE scores and USMLE Step 2 scores. J Eval Clin Pract 2007;13:901-5.
68. Smee SM, Dauphinee WD, Blackmore DB et al. A Sequenced OSCE for Licensure: Administrative Issues, Results and Myths. Advances in Health Sciences Education 2003; 8: 223–236
69. Swanson DB, Clauser BE, Case SM. Clinical skills assessment with standardized patients in high-stakes tests: a framework for thinking about score precision, equating, and security. Adv Health Sci Educ Theory Pract 1999; 4(1):67–106.
70. van der Vleuten C. Validity of final examinations in undergraduate medical training. BMJ 2000; 321:1217-9.
71. Varkey P, Natt N, Lesnick T, Downing S, Yudkowsky R. Validity evidence for an OSCE to assess competency in systems-based practice and practice-based learning and improvement: a preliminary investigation. Acad Med 2008;83(8):775-80.
72. Walters K, Osborn D, Raven P.The development, validity and reliability of a multimodality objective structured clinical examination in psychiatry. Med Educ 2005;39(3):292-8.
73. Wilkinson TJ, Frampton CM, Thompson-Fawcett M, Egan T. Objectivity in objective structured clinical examinations: checklists are no substitute for examiner commitment. Acad Med 2003; 78:219-23.

۶
ابزارهای ارزیابی مبتنی بر محل کار

فصل ۱۹

خانواده ابزارهای ارزیابی مبتنی بر محل کار

تاریخچه ارزیابی مبتنی بر محل کار

ارزیابی فراگیران در آموزش پزشکی در طول ۵۰ سال گذشته تغییرات بارزی کرده است. روش‌های ارزیابی به لحاظ کمیت و کیفیت رشد کرده است؛ روش‌های جدیدی برای ارزیابی توانمندی‌های[1] پزشکان ابداع شده است و ابزارهایی با روایی و پایایی مناسب بوجود آمده است. با وجود این، ویژگی‌های منحصر به فرد آموزش بالینی و از جمله آموزش دستیاری موجب شد که در ابتدا سرعت تکامل روش‌های ارزیابی در محیط بالینی به اندازه محیط آکادمیک یا شبیه‌سازی پزشکی چشم‌گیر نباشد. دو ویژگی مهم آموزش در محیط بالینی که در کاهش سرعت رشد روش‌های ارزیابی تأثیر گذار بودند عبارتند از: ساختارمند نبودن برنامه درسی و مسؤولیت فراگیران در برابر مراقبت از بیمار. علاوه بر این، در ارزیابی فراگیران در محیط بالینی دامنه وسیعی از مشکلات بیماران از مشکلات حاد و اورژانسی گرفته تا بیماری‌هایی که چند سیستم را درگیر می‌کنند و نیز پروسیجرهای پزشکی را باید درنظر داشت. مسأله دیگری که ارزیابی در محیط بالینی و ارزیابی دستیاران را با چالش مواجه می‌کند این است که در مقایسه با پزشکان در حال طبابت مسؤولیت کامل مراقبت از بیمار بر عهده دستیاران نمی‌باشد و در نتیجه نمی‌توان از محصولات عملکرد آن‌ها برای ارزیابی آن‌ها استفاده کرد (سوانویک[2] ۲۰۱۰).

ارزیابی فراگیران در محیط بالینی به شیوه سنتی مشتمل بر دو بخش است: یک بخش در سالن امتحانات، در قالب آزمون‌های کاغذی برگزار می‌شود که در آنها عمدتاً سنجش محفوظات مربوط به دانش مهارت‌های بالینی مدنظر قرار می‌گیرد. بخش دیگر، به یک قضاوت کلی[3] و ذهنی[4] از عملکرد فراگیران در انتهای دوره بالینی و به دور از محیط واقعی می‌پردازد که در اغلب موارد به صورت تشریفاتی انجام می‌شود یا نهایت کاری که انجام می‌شود برگزاری آزمون‌های شفاهی با روایی و پایایی مشکوک بر بالین بیمار است (سوانویک ۲۰۱۰). از دیگر روش‌های مرسوم می‌توان به «مورد بالینی کامل»[5] اشاره کرد که برای سال‌ها به عنوان یکی از ابزارهای اصلی ارزیابی توانمندی‌های بالینی در بسیاری از دانشکده‌های پزشکی سراسر دنیا مورد استفاده قرار می‌گرفت (توضیحات مربوط به این ابزار در فصل دوم ارائه خواهد شد).

به تدریج به دلیل انتقاداتی که به ویژگی‌های روان‌سنجی این روش‌ها وارد شد، این آزمون‌ها جای خود را به دیگر روش‌های ارزیابی دادند. در سال ۱۹۷۲ بورد طب داخلی آمریکا اجرای آزمون‌های شفاهی را برای آزمون‌های منجر به اخذ مدرک متوقف کرد و از آن پس آزمون CEX[6] را معرفی کرد که در کنار بستر بیمار برگزار می‌شد و آزمون‌گر با مشاهده نحوه تعامل دستیار با یک بیمار، عملکرد وی را در مدیریت بیمار ارزیابی می‌کرد اما به زودی این آزمون هم به دلیل محدودیت‌هایش کنار گذاشته

1. Competencies
2. Swanwick
3. Global
4. Subjective
5. Long case
6. Clinical Evaluation Exercise

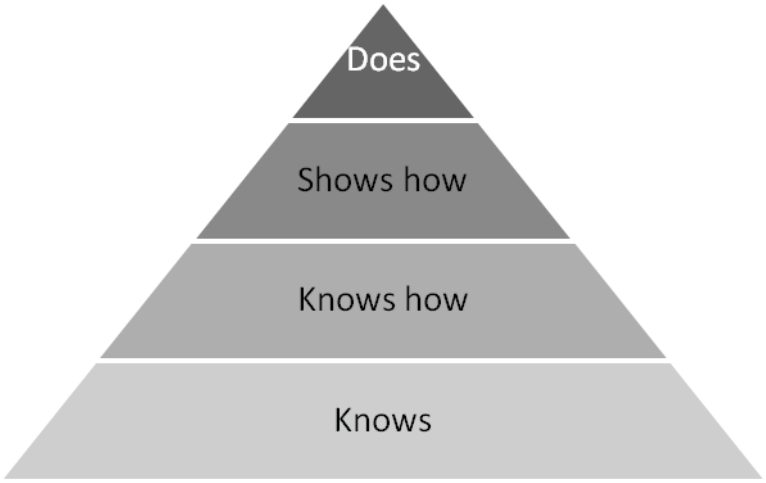
شد و بورد داخلی آمریکا در سال ۱۹۹۵ روش ارزیابی جدیدی تحت عنوان mini-CEX را معرفی نمود (نورسینی و برچ[1] ۲۰۰۷). از طرف دیگر، ارزیابی پروسیجرهای عملی نیز با روش‌هایی صورت می‌گرفت که دارای محدودیت‌هایی بودند. کالج سلطنتی پزشکان انگلیس برای اولین بار در سال ۲۰۰۳ روش ارزیابی DOPS[2] را به منظور ارزیابی پروسیجرهای عملی معرفی کرد و سپس در سال ۲۰۰۵ از آن رسماً در برنامه پیش‌دستیاری استفاده کرد (این دو ابزار به ترتیب در فصول سوم و چهارم ارائه خواهند شد).

علاوه بر موارد فوق که در آنها ارزیابی عملکرد داوطلب بر اساس مشاهده مستقیم صورت می‌گیرد، می‌توان گفت که نقد مستندات و پرونده‌های بیماران در آموزش و ارزیابی بالینی از قدمت زیادی برخوردار است. بسیاری از اعضای هیأت علمی تجربه ارزشیابی دست‌نوشته‌های کارآموزان را به عنوان بخشی از آموزش و ارزیابی دوره کارآموزی دارند. پرونده‌های پزشکی در کنار کاربردهای مهمی که برای مستندسازی تصمیمات پزشکی و بایگانی اطلاعات پزشکی مهم بیماران به منظور استفاده توسط ارائه‌کنندگان خدمات سلامتی دارد، اطلاعات مفیدی جهت ارزیابی عملکرد پزشکان فراهم می‌نماید. از روش‌های ارزیابی مبتنی بر پرونده‌های پزشکی، می‌توان Chart Stimulated Recal (CSR) (در امریکا) و Case Based Dis- (CBD) cussion (در انگلیس) را نام برد (این ابزارها در فصل پنجم این بخش مورد بحث قرار خواهد گرفت).

از طرف دیگر، در دهه‌های ۸۰ و ۹۰ میلادی، مطالعاتی که در مورد کیفیت آموزش پزشکی در سطوح ملی و بین‌المللی انجام شد، نشان دادند که کیفیت آموزش نه تنها در سطح دانشگاه‌های پزشکی مختلف متفاوت است، بلکه دانشجویان مختلف در یک دانشکده پزشکی نیز تجارب آموزشی متفاوتی دارند. زمانی که میزان دستیابی دانشجویان به اهداف فرایندی و پیامدی را ارزیابی می‌کنیم، این سؤال مهم مطرح می‌شود که آیا دانشجو امکان تجربه مواجهه با همه موارد و بیماری‌های مورد نظر را به تعداد مناسب داشته است یا خیر. از ابزارهایی که در بررسی این امر کمک‌کننده هستند، می‌توان لاگ‌بوک[3] و کارپوشه[4] را نام برد که در فصول ششم و هفتم به آنها پرداخته می‌شود. فصل آخر این بخش به ارزیابی ۳۶۰ درجه اختصاص یافته است که خاستگاه آن به صنعت برمی‌گردد. از آنجا که کارفرماها نمی‌توانستند از نزدیک و به طور مکرر بر عملکرد کارکنان نظارت داشته باشند و افراد در گروه‌های کاری در موقعیت بهتری برای مشاهده عملکرد کارکنان بودند، با استفاده از اطلاعات جمع‌آوری شده از افراد مختلف و جمع‌بندی آن، به کارکنان بازخورد داده می‌شد. ارزیابی ۳۶۰ درجه در دنیای صنعتی به این دلیل پذیرفته شده است که اطلاعات را از افراد مختلف و گروه‌های متفاوت جمع‌آوری می‌کند بنابراین سوگیری ذاتی مترتب «به یک فرد، یک دیدگاه» (برای مثال، کارفرما) را کاهش می‌دهد. مشاهده‌گران (ارزیابان) شامل کارفرما، همکاران، زیردستان، مشتریان و خود فرد است.

به طور کلی، اصلاحات اخیر در آموزش دستیاری، سیستم‌های تازه‌ای را برای ارزیابی عملکرد و توانمندی بالینی دستیاران به ارمغان آورده است. یکی از این سیستم‌ها ارزیابی مبتنی بر محل کار[5] است. ارزیابی مبتنی بر محل کار، ارزیابی عملکرد فراگیران در محیط واقعی است. این روش، ارزیابی دانشجو را از سطوح پایین شناختی در هرم میلر به سطوح بالای آن در سطح ارزیابی «آنچه فراگیر انجام می‌دهد»، ارتقا داده است (شکل ۱-۱۹). در محیط‌های آموزشی، ارزیابی مبتنی بر محل کار شامل استفاده از ابزارهای مختلف برای جمع‌آوری اطلاعات در مورد جنبه‌های گوناگون کار فراگیران است که به عنوان ابزاری برای ارائه بازخورد مستقیم، به موقع و مرتبط در مورد عملکرد آن‌ها استفاده می‌شود. از زمان معرفی این ابزارها تاکنون روش‌های متعددی برای ارزیابی مبتنی بر محل کار پیشنهاد و اجرا شده‌اند که در این بخش به آنها خواهیم پرداخت. اما توجه به این نکته ضروری است که روش‌های سنتی کماکان مورد استفاده قرار می‌گیرند و به همین دلیل فصولی از این بخش به آنها نیز اختصاص داده شده است. به عنوان مثال، علی‌رغم مسائلی که در مورد پایایی و روایی «مورد بالینی کامل» گفته می‌شود، همچنان این آزمون در امتحانات مهم به منظور تصمیم‌گیری در مورد عملکرد پزشکان مورد استفاده قرار می‌گیرد و هنوز هم در برنامه‌های آموزشی پزشکی عمومی و دستیاری بسیاری از کشورها از جمله انگلیس، ایرلند، استرالیا، نیوزلند و هندوستان استفاده می‌شود.

1. Norcini & Burch
2. Direct Observation of Procedural Skill (DOPS)
3. Logbook
4. Portfolio
5. Workplace-based assessment



شکل ۱-۱۹: هرم میلر

ضرورت ارزیابی مبتنی بر محل کار

ضرورت ارزیابی مبتنی بر محل کار را می‌توان از زوایای مختلفی مورد بررسی قرار داد. در ابتدا باید مشخص شود با وجود پیشرفت‌های فراوان در فناوری[1]، اهمیت و جایگاه مهارت‌های بالینی در موفقیت مراقبت از بیمار تا چه اندازه است. در این خصوص مطالعات گسترده‌ای نشان داده‌اند که علی‌رغم پیشرفت‌های اخیر در فناوری، کماکان مهارت‌های مصاحبه، معاینه فیزیکی و ارتباط با بیمار سهم بسیار عمده‌ای در تشخیص به موقع و درمان مناسب بیمار دارند. این در حالی است که استفاده از فناوری و بررسی‌های تشخیصی سهم بسیار ناچیزی در پیامدهای بیمار دارد. مطالعات نشان داده‌اند جمع‌آوری اطلاعات به صورت نادرست از بیمار هنوز مهمترین عامل خطاهای پزشکی است (هولمبو[2] ۲۰۰۴).

با وجود پافشاری شواهد بر اهمیت مهارت‌های بالینی، گزارش‌های متعددی مبنی بر ضعف پزشکان در این مهارت‌ها وجود دارد. به طوری‌که پزشکان نمی‌توانند بیش از نیمی از شکایات بیماران را کشف کنند و دستیاران در بیش از ۵۸ درصد موارد حداقل یک خطای معاینه را مرتکب می‌شوند (وری[3] و فریدلند[4] ۱۹۸۳). میزان مطالعات انجام شده در این خصوص فراوان بوده و ذکر آن از دامنه این کتاب خارج است. اما از نتایج این مطالعات می‌توان اهمیت ارزیابی و آموزش مهارت‌های بالینی را دریافت. به دنبال مشخص شدن اهمیت مهارت‌های بالینی مؤسسات اعتباربخشی، آموزش و ارزیابی این مهارت‌ها را از ملزومات اعتبار دانشکده‌های پزشکی قرار دادند (هولمبو ۲۰۰۴).

شاید بتوان گفت یکی از مهمترین مشکلات در ارزیابی مهارت‌های بالینی ناکافی بودن مشاهده مستقیم فراگیران رشته پزشکی توسط استادان بالینی است. در این خصوص مطالعات بسیاری انجام شده است که نتایج برخی از آن‌ها در زیر آورده شده است:

- انجمن دانشکده‌های پزشکی آمریکا[5] در بازدیدی که از ۹۷ دانشکده پزشکی آمریکا بین سال‌های ۱۹۹۳ تا ۱۹۹۸ داشت گزارش کرد که تعامل دانشجویان با بیمار به ندرت توسط اعضای هیأت علمی مورد مشاهده قرار می‌گیرد.

1. Technology
2. Holmboe
3. Wray
4. Friedland
5. Association of American Medical Colleges (AAMC)

به طوری‌که ارزیابی ساختارمند و بر اساس مشاهده در دوره کارآموزی تنها برای ۷/۴ تا ۲۳/۱ درصد دانشجویان رخ می‌دهد (کاسبام و ایگلن[1] ۱۹۹۹).

- در یک مطالعه جدیدتر توسط انجمن دانشکده‌های پزشکی آمریکا در سال ۲۰۰۴ در فارغ‌التحصیلان پزشکی نشان داد که ۱۷ تا ۳۹ درصد دانشجویان در طول کارآموزی‌های اصلی خود در زمان انجام معاینه بالینی مورد مشاهده قرار نگرفته‌اند (نورسینی و برچ ۲۰۰۷).
- در مطالعه دیگر تنها ۲۸ درصد کارآموزان طب داخلی آمریکا مورد ارزیابی ساختارمند بر اساس مشاهده مستقیم قرار گرفتند (کوگان و هاور[2] ۲۰۰۶)
- در خارج از آمریکا نیز نتایج مطالعه دیگری نشان داد که در یک دوره شش ماهه، مشاهده و ارزیابی فراگیران و بازخورد به عملکرد آن‌ها در ۳۵ درصد وقایع آموزشی رخ می‌دهد (دالمنز و همکاران[3] ۲۰۰۴).
- بر اساس گزارش انجمن دانشکده‌های پزشکی آمریکا مشاهده عملکرد دستیاران تخصصی و فوق تخصصی حتی کمتر از دانشجویان رخ می‌دهد. به طوری که ۸۲ درصد دستیاران تنها یک بار در طول سال اول دستیاری در زمان شرح‌حال گیری یا معاینه بیمار مورد مشاهده مستقیم استادان خود قرار گرفته‌اند (دی و همکاران[4] ۱۹۹۰).
- همچنین مطالعات نشان می‌دهد اغلب فرم‌های نمره‌دهی آخر بخش توسط اساتیدی تکمیل می‌شود که عملکرد فراگیران را در زمان ویزیت بیمار مشاهده نکرده‌اند (نورسینی و برچ ۲۰۰۷).

بنابراین، بر اساس نتایج مطالعات فقدان ارزیابی و بازخورد بر اساس مشاهده مستقیم عملکرد در محیط واقعی یکی از جدی‌ترین مشکلات در آموزش بالینی و طبابت است و این موضوع ضرورت و اهمیت ارزیابی ساختارمند عملکرد فراگیران در همه سطوح آموزش پزشکی در محیط واقعی را نشان می‌دهد.

کاربرد ارزیابی مبتنی بر محل کار

هدف هر ارزیابی در طول طیفی از ارزیابی برای یادگیری تا ارزیابی برای پاسخگویی[5] متغیر است. ارزیابی مبتنی برمحل کار نیز می‌تواند در طول این طیف به عنوان محرکی برای یادگیری تا ابزاری برای پاسخگویی به کار گرفته شود. در کشورهای آمریکای شمالی با جمع‌آوری اطلاعات دقیق در مورد پیامدهای بیمار، فرایند و حجم طبابت از این نوع ارزیابی برای پاسخگویی استفاده می‌شود. اما در اکثر موارد ارزیابی مبتنی بر محل کار بخشی از سیستم آموزشی است و به عنوان حلقه اتصال یادگیری و ارزیابی به کار می‌رود. این روش از ارزیابی با فراهم نمودن محیطی واقعی برای ارزیابی و در ادامه، ارائه بازخورد در مورد عملکرد فراگیران نقش مهمی در بهبود یادگیری آن‌ها دارد. در واقع، کاربرد این ارزیابی به عنوان روشی تکوینی و با هدف اصلاح در طول برنامه آموزشی بسیار برجسته است.

به منظور دستیابی هدف تکوینی، بازخورد نقش اساسی و مهمی در روش‌های ارزیابی مبتنی بر محل کار بازی می‌کند. بازخورد در زمینه ارزیابی مبتنی بر محل کار به معنای اطلاعاتی است که توسط یک عامل انسانی که می‌تواند معلم، همکار، خود فرد و دیگران باشند در مورد فهم[6] یا عملکرد[7] فرد داوطلب به وی داده می‌شود (هاتی و تیم‌پرلی[8] ۲۰۰۷). با استفاده از اطلاعات ارائه شده، فراگیر رشته پزشکی می‌تواند اطلاعات قبلی خود را تثبیت کند؛ افزایش دهد؛ بازنویسی یا بازسازی

1. Kassebaum & Eaglen
2. Kogan & Haur
3. Daelmans et al.
4. Day et al.
5. Accountability
6. Understanding
7. Performance
8. Hattie & Timperely

کند. بنابراین هدف اصلی بازخورد کاهش تفاوت بین فهم و عملکرد موجود با فهم و عملکرد مطلوب است.

به منظور دستیابی به این هدف سه سؤال در ذهن فراگیر مطرح می‌شود:

- به کجا می‌روم؟
- چگونه می‌روم؟
- قدم بعدی من چیست؟

برای پاسخ به سؤال اول، مهم‌ترین مسأله این است که «اهداف یادگیری دوره» به وضوح تعریف شده باشند و سپس ارزیابی و ارائه بازخورد بر اساس این اهداف صورت گیرد. اگر اهداف یادگیری به طور شفاف تعیین نشوند فراگیر نمی‌تواند شکاف بین شرایط فعلی و مورد نظر را تشخیص دهد و در نتیجه برای رسیدن به شرایط مطلوب تلاش کند. این مسأله به خصوص در محیط بالینی ارزش زیادی پیدا می‌کند. با توجه به پیچیدگی وظایف در محیط بالینی و عوامل مخدوش‌کننده و مداخله‌گر بسیاری که در این محیط وجود دارد اهداف یادگیری مشخص و واضح می‌تواند فراگیران را به سوی فعالیت‌های یادگیری مناسب هدایت کند. در این صورت فراگیر با وجود مشکلات و عوامل مزاحم، بر یادگیری مهارت یا پروسیجر خاص اصرار می‌ورزد. در مجموع می‌توان گفت در صورتی که اهداف دوره مشخص باشد؛ ارزیابی بر اساس این اهداف انجام شود و بازخورد به فراگیر نیز در راستای این اهداف ارائه شود، فعالیت‌های یادگیری فراگیر نیز به سمت دستیابی به این اهداف جهت می‌گیرد. یکی دیگر از فواید مشخص بودن دقیق اهدف این است که فراگیر برای دستیابی به این اهداف خود به دنبال دریافت بازخورد می‌رود. در این راستا برنامه‌های آموزش پزشکی و از جمله آموزش دستیاری در کشورهای مختلف، سیستمی را به منظور ارزیابی فراگیران تعریف کرده‌اند. در این سیستم پیامدها یا توانمندی‌های مورد انتظار دوره تعیین شده است و بر اساس آن روش‌های ارزیابی دستیاران و آن‌چه که در این روش‌ها مورد ارزیابی قرار می‌گیرد مشخص شده است.

پاسخ به سؤال دوم مستلزم استخراج اطلاعات عینی و دقیق از عملکرد فراگیران مبتنی بر اهداف یادگیری است. به این منظور باید معیارها و شاخص‌هایی تعریف شوند تا به طور دقیق سطح مورد انتظار و مطلوب مهارت یا وظیفه مورد نظر را مشخص کنند. سپس بازخورد بر اساس این معیارها در مورد نقاط مثبت و ضعف عملکرد به فراگیر ارائه می‌شود. تدوین فرم‌های ارزیابی مبتنی بر محل کار که بعداً به تفصیل به آن پرداخته می‌شود در همین راستا است. در این فرم‌ها ابعادی از عملکرد که باید مورد ارزیابی قرار گیرد و سطح مورد انتظار برای هر یک از این ابعاد مشخص می‌شود. در نتیجه ارزیاب می‌تواند بر اساس آن بازخورد دقیقی به فراگیر ارائه دهد و فراگیر نیز از جزئیات عملکرد خود آگاه می‌شود. به عنوان مثال وقتی در فرم آزمون DOPS رعایت شرایط استریل در زمان انجام یک پروسیجر بالینی به عنوان بخشی از عملکرد فراگیر ذکر می‌شود عملکرد ضعیف فراگیر در این بخش و دریافت بازخورد در این ارتباط مسیر یادگیری فراگیر را تعیین می‌کند.

بالاخره مهم‌ترین سؤال از دیدگاه فراگیران این است که چه اقداماتی باید انجام دهد تا بتواند فاصله بین وضع موجود و مطلوب را کاهش دهد. در واقع فراگیران به یک برنامه عملیاتی برای رسیدن به اهداف دوره آموزشی نیاز دارند. لازم به ذکر است که دستیابی به پاسخ سؤال سوم مستلزم رعایت دو مرحله قبل یعنی تعیین اهداف واضح و روشن و تدوین معیارها و شاخص‌های دستیابی به این اهداف است. یکی از چالش‌هایی که در بازخورد در ارزیابی مبتنی بر محل کار وجود دارد این است که در اکثر مواقع مرحله سوم یعنی تدوین یک برنامه عملیاتی اتفاق نمی‌افتد. در حالی که فرمول‌بندی برنامه عملیاتی مهم‌ترین مرحله در ارزیابی تکوینی است و موجب کامل شدن چرخه بازخورد می‌شود.

به طور خلاصه، بازخورد موجب بهبود عملکرد فراگیر و هدایت یادگیری به سمت اهداف مورد نظر دوره می‌شود. به طور معمول بازخورد پس از مشاهده عملکرد فراگیر در یک مواجهه یا در طول عملکرد معمول و روزمره فراگیران داده می‌شود. به عنوان مثالی از نوع اول می‌توان به مشاهده عملکرد فراگیر در انجام یک پروسیجر جراحی و سپس ارائه بازخورد در مورد عملکرد فراگیر در آن مواجهه در آزمون DOPS اشاره کرد. بازخورد ۳۶۰ درجه مثالی از بازخورد در خصوص عملکرد روزمره فراگیران است.

انواع ابزارهای ارزیابی مبتنی بر محل کار

روش‌های متعددی برای طبقه‌بندی ابزارهای ارزیابی مبتنی بر محل کار وجود دارد. نورسینی و هولمبو این ابزارها را از دو بعد، مبنای قضاوت کیفیت عملکرد و نحوه جمع‌آوری داده‌ها، مورد بررسی قرار دادند (نورسینی و هولمبو ۲۰۱۰). آن‌ها اساس قضاوت در مورد عملکرد را در سه جنبه شامل پیامد مراقبت از بیمار[1]، روند مراقبت[2] و حجم کار[3] قرار دادند و روش‌های جمع‌آوری داده‌ها در محیط کار را در چهار گروه پرونده‌ها و مستندات پزشکی[4]، داده‌های مدیریتی[5]، یادداشت‌های روزانه[6] و مشاهده[7] ارائه کردند. تعامل بین عناصر این تقسیم‌بندی را می‌توان در ماتریکسی به صورت جدول ۱-۱۹ نشان داد.

جدول ۱-۱۹: دسته‌بندی ابزارهای ارزیابی مبتنی بر محل کار بر مبنای قضاوت کیفیت عملکرد و نحوه جمع‌آوری داده‌ها (کنتیلون و وود[1] ۲۰۱۰)

	مبنای قضاوت		
روش‌های جمع‌آوری داده‌ها	پیامد مراقبت از بیمار	روند مراقبت از بیمار	حجم کار
پرونده‌ها و مستندات پزشکی	←	←	←
داده‌های مدیریتی	←	←	←
یادداشت‌های روزانه	←	←	←
مشاهده	←	←	←

1.cantillon& wood

مبنای قضاوت

- پیامدهای مراقبت از بیمار: به طور سنتی پیامدهای مراقبت از بیمار را با میزان مرگ و میر بیماران تحت مراقبت یک پزشک ارزیابی می‌کردند. اما هم اکنون موارد دیگری مانند رضایت بیمار، پیامدهای میان‌مدت مانند HbA1c و غلظت چربی در بیماران دیابتی، وضعیت عملکردی، هزینه–اثربخشی و اخیراً میزان خطاهای تشخیصی نیز مورد استفاده قرار می‌گیرند. این نوع قضاوت مورد قبول عموم، بیمار و پزشک است. با وجود اقبال عمومی نسبت به این روش ارزیابی، حداقل پنج مشکل عمده به این روش مترتب است:
  - اِسناد[8]: به منظور یک قضاوت صحیح در مورد عملکرد پزشکان، پیامدهای مراقبت از بیمار باید تنها به فرد مورد ارزیابی قابل انتساب باشد. اما در واقعیت به دلیل این که مراقبت از بیمار در یک سیستم و با مشارکت تیمی از ارائه دهندگان خدمت انجام می‌شود، این امر ممکن نیست. البته اخیراً ارزیابی توانمندی‌های کار گروهی نیز برای پزشکان مورد توجه قرار گرفته است.
  - پیچیدگی بیماری: بیماران با شرایط مشابه به دلیل متفاوت بودن شدت بیماری سرانجام متفاوتی خواهند داشت. بنابراین نمی‌توان عملکرد پزشکان را در مورد یک بیماری یکسان با روشی استاندارد مورد قضاوت قرار داد. هر چند برخی از

1. Outcomes of care
2. Process of care
3. Practice volume
4. Clinical Practice Records
5. Administrative database
6. Diaries
7. Observation
8. Attribution

روش‌های آماری می‌توانند تا حدودی این تفاوت را رفع کنند اما کاملاً موثر نیستند.

- مخلوط شدن موارد بیماری: در اکثر مواقع یک بیمار تنها به یک پزشک مراجعه نمی‌کند و در نتیجه نمی‌توان به صورت استاندارد عملکرد پزشکان را مورد بررسی قرار داد.
- تعداد بیماران: به منظور تخمین عملکرد باید تعداد بیماران قابل ملاحظه باشد. این مسأله ارزیابی پیامدها را به موارد بیماری شایع محدود می‌کند.
- شناسایی پیامدها: برای شناسایی پیامدهایی مانند خطاهای پزشکی باید سیستمی برای تعیین دقیق و طبقه‌بندی خطاها در دسترس باشد که بسیاری از سیستم‌های بهداشتی فاقد آن هستند.

- **فرایند مراقبت از بیمار:** مقیاس‌های عمومی ارزیابی فرایند مراقبت از بیمار شامل غربالگری، خدمات پیشگیری، تشخیص، درمان، تجویز، مشاوره و آموزش بیمار است. علاوه بر این، مقیاس‌های اختصاصی فرایند مانند پایش مداوم HbA1c بیماران دیابتی و معاینه روتین پا در این بیماران نیز به منظور ارزیابی فرایند مورد استفاده قرار می‌گیرد. ارزیابی فرایند نسبت به ارزیابی پیامدها فوایدی دارد. اولاً، فرایند مراقبت بیشتر تحت کنترل پزشک است و در نتیجه مشکل اسناد تا حدودی برطرف می‌شود. ثانیاً، به میزان کمتری تحت تاثیر شدت بیماری قرار می‌گیرد، به عنوان مثال، پزشک بدون توجه به شدت دیابت HbA1c را اندازه می‌گیرد. ثالثاً، برخی از موارد فرایند مانند واکسیناسیون باید برای همه مراجعان خاصی تجویز شوند و این مشکل مخلوط شدن موارد بیماری را کاهش می‌دهد. مهمترین ایراد ارزیابی فرایند این است که فرایند صحیح لزوماً تضمین کننده بهترین پیامدهای بیمار نیست. این‌که پزشک به طور مرتب HbA1c بیمار را بررسی می‌کند به معنی اطمینان از ایجاد تغییرات لازم در مراقبت از بیمار نیست.
- **حجم کار:** سومین مورد از مبانی قضاوت عملکرد، بررسی تعداد مواردی است که یک پزشک در یک فعالیت خاص مانند انجام یک پروسیجر درگیر آن می‌شود. پژوهش‌ها نشان داده است کیفیت مراقبت از بیمار با تعداد دفعات انجام کار ارتباط دارد. هر چند این نوع از ارزیابی مشکلاتی مانند اسناد، پیچیدگی موارد و مخلوط شدن بیماران را ندارد اما دفعات انجام فعالیت تضمین‌کننده کیفیت آن نیست.

روش‌های جمع‌آوری اطلاعات

- **پرونده‌ها و مستندات پزشکی:** یکی از بهترین منابع اطلاعات در مورد پیامد، فرایند و حجم کار مستندات پزشکی است. ممیزی[1] بیرونی این مستندات روشی معتبر در ارزیابی عملکرد پزشکان است. اما جمع‌آوری این اطلاعات گران و وقت‌گیر است و اطلاعات جمع‌آوری شده در بسیاری از موارد ناقص و فاقد شرایط لازم هستند. ثبت الکترونیکی اطلاعات پزشکی می‌تواند کلید حل این مشکل باشد. علاوه بر این، برخی گروه‌ها به مستنداتی که پزشکان خود جمع‌آوری و ارسال می‌کنند اتکا می‌کنند. این روش به همراه ممیزی بیرونی نمونه‌هایی از کار این پزشکان جایگزینی معتبر و قابل اجرا است.
- **داده‌های مدیریتی:** اغلب، اطلاعات مربوط به طبابت پزشکان به منظور مدیریت امور مربوط به مراقبت از بیمار یا پرداخت‌های مالی به پزشکان جمع‌آوری می‌شود. این اطلاعات در دسترس و ارزان هستند و منبع مفیدی برای برخی از جنبه‌های عملکرد مانند هزینه–اثربخشی و تعیین خطاهای پزشکی هستند. هر چند جمع‌آوری این اطلاعات برای اهداف دیگری مانند پرداخت حقوق پزشکان، اهمیت آن را به عنوان تنها منبع گردآوری اطلاعات در خصوص عملکرد پزشکان کاهش می‌دهد.
- **یادداشت‌های روزانه:** پزشکان و به ویژه فراگیران از یادداشت‌برداری یا ابزار لاگ‌بوک برای مستند کردن پروسیجرهایی که انجام می‌دهند استفاده می‌کنند. این منبع اطلاعاتی می‌تواند برای تخمین حجم کار استفاده شود.

1. Audit

- **مشاهده:** یکی از رایج‌ترین روش‌های مشاهده نمره‌دهی توسط سرپرست، همکار، سایر اعضای حرفه‌های پزشکی و یا بیمار است. از موارد دیگر می‌توان به ارزیابی توسط بیمار استاندارد شده در محل طبابت پزشک و مشاهده ویزیت بیماران از طریق نوارهای صوتی و تصویری ضبط شده اشاره کرد.

تقسیم‌بندی نورسینی و هولمبو از ارزیابی مبتنی بر محل کار بیشتر در ارزیابی پزشکان در حال طبابت به کار می‌رود، هر چند می‌تواند برای ارزیابی عملکرد فراگیران در محیط آموزشی نیز مورد استفاده قرار گیرد. سوانویک و چانا[1] روش‌های ارزیابی مبتنی بر محل کار را به صورت زیر تقسیم‌بندی کرده‌اند (سوانویک و چانا ۲۰۰۹):

- مشاهده فعالیت‌های بالینی
- بحث موارد بالینی
- تحلیل داده‌های مربوط به عملکرد
- بازخورد از منابع مختلف

با توجه به این‌که این تقسیم‌بندی در ارزیابی عملکرد فراگیران بیشتر مورد استفاده قرار گرفته است، فصل‌های مربوط به بخش ابزارهای ارزیابی مبتنی بر محل کار در کتاب حاضر نیز بر اساس آن برنامه‌ریزی شده است. به طوری‌که در فصل‌های بعدی کتاب ابزارهای مربوط به هر یک از دسته‌ها به صورت زیر معرفی می‌شوند:

- ابزارهای مشاهده مستقیم عملکرد: mini-CEX و DOPS
- ابزارهای بحث موارد بالینی: CBD[2] یا CSR[3]
- ابزارهای تحلیل داده‌های مربوط به عملکرد: ابزارهای لاگ‌بوک و کارپوشه
- ابزارهای مربوط به دریافت بازخورد از منابع مختلف: ابزار ارزیابی ۳۶۰ درجه که نام دیگر آن MSF[4] است.

لازم به ذکر است که در این کتاب، «مورد بالینی کامل» و «مورد بالینی کوتاه[5]» نیز در همین بخش معرفی می‌شوند. در حقیقت این دو ابزار از این جهت که سطح «نشان می‌دهد چگونه» را در هرم میلر ارزیابی می‌کنند، با دیگر ابزارهای مبتنی بر محل کار متفاوت هستند اما به این دلیل که عملکرد در محیط واقعی را مورد ارزیابی قرار می‌دهند، در خانواده روش‌های ارزیابی مبتنی بر محل کار به آنها پرداخته می‌شود.

گام‌های طراحی و اجرای ابزارهای ارزیابی مبتنی بر محل کار

به منظور طراحی و اجرای بهینه آزمون‌های مبتنی بر محل کار رعایت مجموعه‌ای از قواعد و گام‌ها ضروری است. برخی از این قواعد و گام‌ها به مراحل قبل از اجرای آزمون و برخی دیگر به مراحل اجرا و پس از اجرای آزمون مربوط می‌شود. از جمله مواردی که قبل از اجرای آزمون باید در نظر داشت پیش‌بینی و تأمین منابع و امکانات مورد نیاز شامل نیروی انسانی، منابع مالی و امکانات فیزیکی است. روش‌های ارزیابی مبتنی بر محل کار هزینه‌بر، زمان‌بر و نیازمند نیروی انسانی متخصص و آموزش‌دیده است و در نتیجه فراهم نمودن این امکانات از الزامات طراحی و اجرای موفقیت‌آمیز روش‌های ارزیابی مبتنی بر محل کار است. در فصل‌های مربوط به روش‌های مختلف ارزیابی مبتنی بر محل کار به طور مفصل در مورد هر یک از مراحل ابزارهای مختلف بحث می‌شود. در این‌جا برخی از جنبه‌های عملی مهم ارزیابی مبتنی بر محل کار شرح داده می‌شود. خلاصه این موارد در جدول ۲-۱۹ آمده است.

1. Chana
2. Case-based discussion
3. Chart-stimulated recall
4. Multi-source Feedback
5. Short case

جدول ۲-۱۹: خلاصه گام‌های طراحی و اجرای ابزارهای ارزیابی مبتنی بر محل کار

ردیف	عنوان	توضیح
۱	جلب مشارکت اعضای هیأت علمی	با مشارکت دادن اعضای هیأت علمی در فرایند برنامه‌ریزی و طراحی ارزیابی مبتنی بر محل کار و توجیه ایشان در مورد هدف و فرایند مشاهده و بازخورد می‌توان مشارکت این افراد در اجرای ارزیابی مبتنی بر محل ما را افزایش داد. همچنین در نظر گرفتن سیستم پاداش و نیز زمان جداگانه‌ای در برنامه کاری اعضای هیأت علمی و در نظر گرفتن یک گروه از اعضای هیأت علمی که کار آموزشی آن‌ها ارزیابی و ارائه بازخورد باشد از دیگر راهکارهای جلب مشارکت خواهند بود.
۲	آموزش اعضای هیأت علمی	آموزش و توانمندسازی اعضای هیأت علمی یکی از راهکارهای مهم در بهبود کیفیت ارزیابی مبتنی بر محل کار است که می‌تواند قبل و حین اجرای ارزیابی به عمل آید. مدل‌های مختلفی برای آموزش ارزیابان در روش‌های ارزیابی مبتنی بر محل کار ارائه شده است.
۳	استفاده از توانمندی‌ها و اهداف برنامه به عنوان راهنمایی برای مشاهده مستقیم عملکرد	ضروری است چارچوب توانمندی‌ها و اهداف مربوط به مهارت‌های بالینی مورد انتظار از فراگیران، سطح عملکرد مورد انتظار از فراگیران در مقاطع زمانی مختلف و اهداف رفتاری مورد انتظار برای مهارت‌های بالینی اصلی مشخص شوند تا هدایت کننده ارزیابان و داوطلبان در مورد رفتارهایی که باید مورد مشاهده و ارزیابی قرار گیرد باشد.
۴	انتخاب ابزارهای ارزیابی مناسب	طیفی از ابزارهای ارزیابی عملکرد در محیط کار وجود دارند که می‌توان با در نظر گرفتن شرایط و با رعایت موارد مربوط به سودمندی ابزار و هدف آزمون از آن‌ها استفاده نمود.
۵	ترویج فرهنگ ارزش‌گذاری ارزیابی مبتنی بر محل کار به ویژه مشاهده مستقیم عملکرد	مشاهده مستقیم عملکرد و ارائه بازخورد باید جزئی از فرهنگ مؤسسه باشد. به این منظور تأکید بر ضرورت مشاهده مستقیم با ارائه شواهدی از تأثیر سوء ضعف در مهارت‌های بالینی بر مراقبت از بیمار و تاثیر بازخورد بر کاهش خطاهای تشخیصی و ارتقای خبرگی، استفاده از افراد الگو در توانمندسازی اعضای هیأت علمی، و پاداش، تامین مالی و کاهش بار درمان افرادی که به طور مستمر عملکرد فراگیران را مشاهده می‌کنند از راهکارهای مؤثر می‌باشند.
۶	طراحی فرم‌های ارزیابی	از موارد دیگری که قبل از اجرای آزمون باید در مورد آن چاره‌ای اندیشید، تهیه فرم‌های ارزیابی عملکرد، نمره‌دهی و ارائه بازخورد است. فرم‌های متعددی متناسب با انواع روش‌های ارزیابی مبتنی بر محل کار طراحی شده است که نحوه نمره‌دهی آن به سه صورت مقیاس نمره‌دهی، چک‌لیست و روبریک است.
۷	قضاوت در مورد عملکرد فراگیران	در آزمون‌های مبتنی بر محل کار، قضاوت بر اساس مشاهده عملکرد در یک مواجهه خاص یا عملکرد روزانه داوطلب صورت می‌گیرد و در خصوص رخداد رفتار و کیفیت و تناسب آن قضاوت می‌شود.
۸	ارائه بازخورد بر اساس عملکرد مشاهده شده	یکی از بخش‌های مهم آزمون‌های مبتنی بر محل کار ارائه بازخورد است. ضروری است زمان و مکان مناسبی برای بازخورد در نظر گرفته شود و به طور اختصاصی در مورد رفتارهای داوطلب بازخورد و محیطی توأم با احترام و غیرتهدیدآمیز بازخورد ارئه شود.

جلب مشارکت اعضای هیأت علمی

یکی از مهم‌ترین عوامل در موفقیت اجرای روش‌های ارزیابی مبتنی بر محل کار، اعضای هیأت علمی بالینی و مشارکت ایشان است. تصور کنید سیستم خوبی برای روش‌های ارزیابی مبتنی بر محل کار طراحی کرده‌اید و فرم‌های روا و پایایی برای آزمون‌های خود تهیه کرده‌اید، در صورتی که استادان این سیستم ارزیابی را پیاده نکنند؛ دانشجویان را در محیط واقعی مشاهده نکنند؛ فرم‌ها را کامل نکنند و بازخورد ارائه ندهند زحمات شما به هدر رفته است! در سال ۱۹۸۱ لاندری و فار[1] جستجوی وسیعی را به منظور یافتن یک فرم ارزیابی «ایده‌ال» آغاز کردند و در انتها نتیجه گرفتند که به جای صرف وقت و هزینه برای این کار باید انرژی خود را بر آموزش اعضای هیأت علمی به منظور استفاده مناسب و مؤثر از فرم‌ها متمرکز می‌کردند. با وجود این، پس از گذشت سال‌ها جستجو برای یافتن ابزارها و فرم‌های ارزشیابی ایده‌ال همچنان ادامه دارد (هولمبو و هاوکینز[2] ۲۰۰۸).

1. Landry & Farr
2. Hawkins

به منظور جلب مشارکت اعضای هیأت علمی راهبردهایی به کار گرفته شده است و مطالعات معدودی کارآیی این راهبردها را بررسی کرده‌اند و به نتایج متغیری دست یافته‌اند. از جمله این راهبردها می‌توان به موارد زیر اشاره کرد:

- مشارکت دادن اعضای هیأت علمی در فرایند برنامه‌ریزی و طراحی ارزیابی مبتنی بر محل کار
- توجیه اعضای هیأت علمی در مورد هدف و فرایند مشاهده و بازخورد
- یادآوری منظم و مداومِ نقش مهم ارزیابی تکوینی برای اعضای هیأت علمی و در نتیجه فعال نگهداشتن ارزیابی مبتنی بر محل کار
- در نظر گرفتن سیستم پاداش برای افرادی که در این فعالیت‌ها شرکت می‌کنند
- در نظر گرفتن زمان جداگانه‌ای در برنامه کاری اعضای هیأت علمی به منظور مشارکت بیشتر ایشان در ارزیابی مبتنی بر محل کار
- در نظر گرفتن یک گروه از اعضای هیأت علمی که کار آموزشی آن‌ها ارزیابی و ارائه بازخورد است

هولمبو و همکاران ۲۰۰۱

پژوهشگران در مطالعه‌ای با روش کارآزمایی تصادفی کنترل شده تأثیر توجیه استادان در مورد اهمیت مشاهده و بازخورد را بررسی کردند. آن‌ها برگه‌های نمره‌دهی طراحی کردند که در آن ابعاد مورد سنجش عملکرد و ارائه بازخورد مشخص شده بود. استادان گروه مداخله قبل از شروع چرخش بالینی در یک جلسه توجیهی در مورد اهمیت مشاهده مستقیم عملکرد فراگیران و ارائه بازخورد به مدت ۲۰ دقیقه شرکت کردند. نتایج مطالعه نشان داد کمیت بازخورد ارائه شده در گروه مداخله بیشتر از گروه کنترل نبود اما فراگیران از کیفیت بازخورد ارائه شده توسط این گروه (مداخله) رضایت بیشتری داشتند.

ون‌درهماستوکروس و همکاران [1] ۲۰۰۴

در مطالعه‌ای در هلند، روش‌های ارزیابی فراگیر در دوره کارآموزی جراحی با هدف افزایش مشاهده عملکرد فراگیران و ارائه بازخورد توسط اعضای هیأت علمی با تجربه، بازنگری شد. تغییرات دوره شامل معرفی ابزار ارزیابی لاگ‌بوک، فرم مشاهده مستقیم عملکرد فراگیران و ارزیابی فردی توسط استادان با تجربه بود. به اعضای هیأت علمی در مورد تغییرات اطلاع‌رسانی شده بود اما این افراد هیچ آموزشی در مورد روش‌های ارزیابی و ارائه بازخورد دریافت نکردند. نتایج مطالعه هیچ افزایش بارزی در میزان مشاهده عملکرد فراگیران و ارائه بازخورد نشان نداد. محققان فقدان تأثیر مداخله را به عدم جلب مشارکت اعضای هیأت علمی در فرایند بازنگری روش‌های ارزیابی فراگیر در دوره کارآموزی نسبت دادند.

1. Van der hem-stokroos et al.

مطالعه دالمنز و همکاران ۲۰۰۵

در این مطالعه نیز مداخله‌ای تحت عنوان «ارزیابی حین آموزش» در محیط بالینی اجرا شد، اما میزان مشاهده عملکرد و ارائه بازخورد به فراگیران تغییری نکرد. در ابتدای دوره به اعضای هیأت علمی در خصوص «ارزیابی حین آموزش» اطلاع‌رسانی شد و برگه‌هایی حاوی مطالبی در مورد انواع روش‌های «ارزیابی حین آموزش» توزیع شد. پژوهشگران پیشنهاد دادند یادآوری مستمر «ارزیابی حین آموزش» مانند جلسات توجیهی روزانه با استادان در مورد اهمیت و تأثیرات آموزشی بالقوه سیستم جدید می‌تواند در موفقیت آن کمک کننده باشد.

همه این مطالعات اهمیت به کار بستن راهبردهایی برای جلب مشارکت اعضای هیأت علمی را به موازات تلاش برای تغییر سیستم ارزیابی در محیط بالینی مورد تأکید قرار داده است. در واقع جلب مشارکت اعضای هیأت علمی یکی از عوامل مهم تضمین موفقیت اجرای ارزیابی مبتنی بر محل کار است. علاوه بر اعضای هیأت علمی فراگیران نیز باید در مورد فرایند آزمون توجیه شوند. بازخورد یک عمل دو طرفه است، داوطلبان باید اهمیت مشاهده و بازخورد در کسب مهارت را درک کنند و انتظارات را بدانند. این موارد باعث کاهش اضطراب آن‌ها نیز می‌شود.

آموزش اعضای هیأت علمی

هر چند اجرای ارزیابی مبتنی بر محل کار به نوبه خود دستاورد مهمی محسوب می‌شود، کیفیت ارزیابی عملکرد و بازخورد ارائه شده نیز اهمیت به‌سزائی دارد. همان‌طور که قبلاً نیز اشاره شد، مطالعات بسیاری از دقت پایین اعضای هیأت علمی در تعیین خطاهای داوطلبان و همچنین کیفیت پایین بازخوردهای ارائه شده حکایت می‌کند. به نظر می‌رسد آموزش و توانمندسازی اعضای هیأت علمی یکی از راهکارهای مهم در بهبود کیفیت ارزیابی مبتنی بر محل کار باشد که می‌تواند قبل و حین اجرای ارزیابی به عمل آید. در این قسمت مدل سه مرحله‌ای آموزش ارزیابان که توسط هولمبو و هاوکینر (۲۰۰۸) ارائه شده است بیان می‌شود:

- **آموزش مشاهده رفتاری:** آموزش مشاهده رفتاری[1] بر ارتقای توانمندی ارزیابان در تعیین، درک و بازگو کردن عملکرد در محیط واقعی متمرکز است. در این روش بر سه راهبرد تأکید می شود:
  - فراهم نمودن فرصت‌های بیشتر برای مشاهده مستقیم عملکرد فراگیران، این روش باعث تمرین مهارت مشاهده و یادگیری آن توسط ارزیابان می شود.
  - تهیه وسایل کمک مشاهده‌ای که ارزیابان بتوانند با آن مشاهده خود را ثبت کنند که گاهی به آن یادداشت‌های روزانه می‌گویند. مطالعات نشان داده‌اند که حتی تهیه یک کارت کوچک ایندکس ۳ در ۵ اینچ برای یادداشت کردن مشاهده، کیفیت اطلاعات وارد شده در فرم‌های ارزیابی را افزایش می‌دهد. اگر اعضای هیأت علمی به صورت مداوم یادداشت بردارند این موضوع کمک می‌کند تا در زمان انجام مشاهده عینی‌تر عمل کنند. مثلاً افراد می‌توانند هر روز یک نقطه قوت و یک نقطه قابل اصلاح از فراگیران را یادداشت کنند. افراد می‌توانند این اطلاعات را در مدت زمان کوتاهی وارد کامپیوتر شخصی خود کنند و به عنوان یک منبع برای ارزیابی پایان ترم در اختیار داشته باشند. علاوه بر این کارت‌ها، فرم‌های نمره‌دهی کلی[2] و چک‌لیست (فهرست‌وارسی) نیز می‌توانند به این منظور استفاده شوند.
  - به ارزیابان کمک کنیم تا یاد بگیرند که چگونه برای جلسه مشاهده آماده شوند. به این صورت که قبل از شروع مشاهده اهداف مشاهده را تعیین کنند. به عنوان مثال اگر هدف، ارزیابی مهارت معاینه فیزیکی است باید در جریان مشکل و شرح‌حال بیمار باشند تا بدانند جنبه‌های مهم معاینه فیزیکی بیمار مورد نظر چیست. مصاحبه کوتاه با بیمار قبل از شروع مشاهده یا ارائه بیمار توسط آزمون‌دهنده به این امر کمک می‌کند. علاوه بر این، موقعیت مشاهده‌گر در زمان مشاهده داوطلب به طوری که بهترین کارآیی و کمترین تداخل با مواجهه را داشته باشد از موارد دیگر است. آماده کردن داوطلب و بیمار و به حداقل رساندن آسیب یا تداخل در طول مواجهه نیز از دیگر موضوعات مورد آموزش است (جدول ۳-۱۹).

جدول ۳-۱۹: چهار قانون ساده برای مشاهده در ارزیابی مبتنی بر محل کار (هولمبو و هاوکینز ۲۰۰۸)

قانون	توصیف
موقعیت صحیح	به عنوان ارزیاب سعی کنید در جلوی دید بیمار و داوطلب به ویژه زمانی که در حال برقراری ارتباط هستند نباشید. اصل مثلث سازی را رعایت کنید. البته مطمئن باشید در زمان انجام معاینه اشراف کامل برای بررسی انجام صحیح معاینه توسط داوطلب را دارید.
کاهش مداخله بیرونی	به پرسنل بخش اطلاع دهید که دقایقی را باید با دانشجوی خود بگذرانید. از پاسخ به تلفن بپرهیزید و موارد دیگر ...
پرهیز از دخالت	از مداخله در مواجهه پزشک و بیمار پرهیز کنید. زیرا ممکن است روند آن را تغییر دهد. البته در مواردی که داوطلب مسیر را اشتباه رفته است یا هدف آزمون را به خوبی درک نکرده است مجاز به دخالت هستید.
آمادگی	قبل از ورود به اتاق از اهداف و تمرکز آزمون اطلاع کسب کنید.

1. Behavioral observation training
2. Global rating

- **آموزش ابعاد عملکرد[1]:** در این روش اعضای هیأت علمی با حیطه‌های مورد ارزیابی در سیستم ارزیابی آشنا می‌شوند. این آشنایی شامل مرور تعاریف هر حیطه عملکرد، معیارهای قضاوت در مورد آن و رفتارهای معرف انجام موفقیت آمیز آن حیطه توسط داوطلب است. به عنوان مثال، در سیستمی که بر اساس ارزیابی توانمندی‌های ACGME[2] بنا نهاده شده است، ابتدا این توانمندی‌ها معرفی می‌شوند و سپس تعاریف و معیارهای هر یک بازگو می‌شود. نکته مهم این است که ارزیابان این معیارها و تعاریف را به کار ببرند و بر آن بازاندیشی کنند تا به یک اجماع در مورد این تعاریف و معیارها برسند. این مرحله بسیار مهم است زیرا موجب استاندارد شدن قضاوت ارزیابان و در نتیجه ارتقای روایی و پایایی و نیز رعایت عدالت در آزمون می‌شود. از دو روش زیر می‌توان در آموزش ابعاد عملکرد استفاده کرد:
  - ارائه سناریو به منظور تسهیل تعامل ارزیابان با توانمندی‌ها: در این روش سناریو ساده‌ای از یک موقعیت بالینی بر اساس یکی از توانمندی‌های مورد ارزیابی ارائه می‌شود و سپس از ارزیابان سؤال می‌شود به نظر آن‌ها یک مواجهه مؤثر در برخورد با موقعیت مطرح شده چگونه است. شرکت‌کنندگان در دوره آموزشی باید رفتار فراگیران را برای یک مواجهه مطلوب توصیف کنند. تمرکز بر رفتار بسیار مهم است به این دلیل که آنچه که ارزیابان مشاهده می‌کنند رفتار داوطلبان است. بهتر است این کار در گروه‌های کوچک انجام شود و سپس گروه‌ها نتایج کار خود را با هم به اشتراک بگذارند. بحث در مورد تفاوت‌های موجود باعث شکل‌گیری بهتر معیارها و رفتارهای متناسب آن در فراگیران می‌شود. معمولاً در این نوع آموزش دو موقعیت بالینی در یک ساعت مورد بحث قرار می‌گیرد.
  - ارائه یک موقعیت بالینی واقعی و ایجاد واکنش در ارزیابان: در این روش آموزشی یک موقعیت بالینی واقعی که معمولاً به صورت نوار ویدئویی ضبط شده است ارائه می‌شود و شرکت‌کنندگان در برنامه آموزشی در مورد رفتارهای مناسب و نامناسب داوطلب بحث می‌کنند.
- **آموزش چارچوب مرجع[3]:** در این بخش از آموزش، اعضای هیأت علمی برای بهبود دقت و قدرت تمیز خود در قضاوت و کاهش تفاوت در سخت‌گیری آموزش می‌بینند. در واقع، آموزش چارچوب مرجع ادامه مرحله قبل است و ارزیابان با استفاده از معیارهای مرحله قبل به تمایز سطوح مختلف عملکرد در داوطلبان می‌پردازند. سطوح مختلف عملکرد را می‌توان به صورت «نامطلوب»، «مرزی»، «مطلوب» و «عالی» تقسیم‌بندی کرد. به عنوان مثال، ابتدا باید تعریف کنیم معیارهای رفتار عالی از جنبه پیامدهای مطلوب برای بیمار چیست. سپس حداقل رفتار مورد نظر برای داوطلب با عملکرد مطلوب تعریف می‌شود. سپس رفتار داوطلب مرزی تعیین می‌شود و هر آنچه کمتر از این عملکرد باشد در سطح نامطلوب تعریف می‌شود. در این مرحله، شرکت‌کنندگان می‌توانند از شواهد موجود در این زمینه برای رسیدن به اجماع استفاده کنند. به عنوان مثال، مطالعات نشان می‌دهند در شروع شرح‌حال‌گیری سؤالات باز پاسخ مناسب‌تر هستند. همچنین مدل‌های مختلفی از استانداردهای مهارت‌های بالینی در اختیار هستند. از جمله این موارد می‌توان به الگوی کالگری–کمبریج و SEGUE[4] در مصاحبه با بیمار، الگوی تصمیم‌گیری آگاهانه[5] در مشاوره و همچنین مدل‌های در دسترس برای سودمندی مهارت‌های جراحی اشاره کرد. نمونه‌ای از تمرین دو مرحله آموزش ابعاد عملکرد و چارچوب مرجع در زیر آمده است.

1. Performance dimension training
2. Accreditation Council for Graduate Medical Education (ACGME)
3. Frame of Reference Training (FORT)
4. Set the stage, Elicit information, Give information, Understand the patient's perspective, and End the encounter
5. Informed decision making

نمونه تمرین برای تدوین معیارهای مربوط به یکی از حیطه‌های توانمندی (مشاوره)

دستیار، بیماری را ویزیت می‌کند که تشخیص وی مسجل شده است. او باید داروی جدیدی را برای بیمار شروع کند.
معیارهای یک مشاوره و آموزش بیمار در سطح «عالی و کاملا کارآمد» چیست؟ شما از چه معیارهایی استفاده می‌کنید تا در مورد عملکرد مشاوره‌ای این دستیار قضاوت کنید؟ معیارهای یک مواجهه مشاوره مؤثر و اجزای آن را بر اساس مدل دانش، مهارت و نگرش (KSA) مشخص کنید. به یاد داشته باشید که شما در موقعیت مشاهده رفتار عمل می‌کنید و باید معیارها مربوط به رفتار قابل مشاهده باشند.
پس از این‌که معیارها را مشخص کردید رفتارهایی که داوطلب باید نشان دهد تا در برابر این معیارها نمره مطلوب بگیرد را مشخص کنید.

در راستای توانمندسازی اعضای هیأت علمی در ارزیابی بر اساس مشاهده، هولمبو و همکاران (۲۰۰۴) کارگاهی را طراحی کردند و سه روش آموزش ذکر شده در بالا را در آن به کار بردند. این کارگاه در چندین مورد در محیط‌های مختلف مورد استفاده قرار گرفته است و تاثیر آن مورد مطالعه قرار گرفته است:

هولمبو و همکاران ۲۰۰۴

مدت کارگاه اجرا شده توسط پژوهشگران چهار روز بود. ۴۰ عضو هیأت علمی از ۱۱ برنامه دستیاری در ۵ دانشگاه مختلف آمریکا در آن شرکت کردند. روش آموزش کارگاه شامل سخنرانی‌های کوتاه، بحث گروه کوچک، تمرین ارزیابی با فیلم ویدئویی و تمرین با فراگیر و بیمار استاندارد شده بود. ارزیابی عملکرد از طریق نوارهای ویدیوئی ضبط شده صورت گرفت به این صورت‌که دو سری ۹ تایی سناریو تهیه شد که در آن بیمار استاندارد شده و دستیار استاندارد شده نقش بازی می‌کردند. یک سری از سناریوها برای ارزیابی پایه و سری دیگر برای ارزیابی بعد از آموزش بودند. حیطه‌های مورد ارزیابی مصاحبه، معاینه و مشاوره بود. مواجهه‌ها در هر سه سطح توانمندی تهیه شدند و هر کدام دارای خطاهایی از عملکرد بودند. به عنوان مثال «سطح ۱» نشان‌دهنده عملکرد ضعیف بود و متوسط خطای عملکرد ۱۲ مورد بود. «سطح ۲» نشان‌دهنده عملکرد متوسط و با ۶ مورد خطا و «سطح ۳» نشان‌دهنده عملکرد عالی و با ۲ مورد خطای عملکرد بود. نمونه‌ای از خطاها در حیطه مصاحبه شامل عدم معرفی خود به بیمار، سؤال نکردن از عوامل خطر بیماری‌های ترومبولیتیک در بیمار با تورم ساق پا بودند. دو روز از کارگاه به آموزش مشاهده مستقیم اختصاص داشت. هشت ماه پس از برگزاری کارگاه پژوهشگران به ارزشیابی تاثیر کارگاه با سه معیار رضایت از آموزش ارائه شده، راحتی اعضای هیأت علمی با مشاهده عملکرد و تغییر رفتارهای نمره‌دهی اعضای هیأت علمی پرداختند. اعضای هیأت علمی کارگاه را بسیار عالی ارزشیابی کردند و اشاره داشتند آن را به همکاران خود توصیه می‌کنند. گروه مداخله احساس راحتی بیشتری در مشاهده مستقیم داشتند که با گروه کنترل تفاوت معنی‌دار داشت. هیچ تفاوت معنی‌داری در کیفیت قضاوت گروه کنترل و شاهد مشاهده نشد و تنها سخت‌گیرانه‌تر ارزیابی کردند. این لزوماً بدان معنی نیست که نمره‌دهی آن‌ها دقیق‌تر شده بود، اما می‌تواند منجر به بازخورد مؤثرتری شود. گروه مداخله با دقت بیشتری توانستند دانشجویان با عملکرد ضعیف را تشخیص دهند اما در کل دقیق‌تر نبودند. این موضوع از جهت یافتن زودتر دستیاران با عملکرد ضعیف و اندیشیدن چاره‌ای برای آنان ارزش دارد. پژوهشگران توصیه به بررسی تأثیر این نوع کارگاه‌ها از طریق بررسی تعامل ارزیابان در محیط واقعی و همچنین بر تمرکز بیشتر بر آموزش دقت مشاهده نمودند.

استفاده از توانمندی‌ها و اهداف برنامه به عنوان راهنمایی برای مشاهده مستقیم عملکرد

روش‌های ارزیابی مبتنی بر محل کار باید به عنوان جزئی از برنامه آموزشی و ارزیابی مهارت‌های بالینی باشند. مسؤولان برنامه‌های آموزشی باید چارچوبی را برای توانمندی‌ها و اهداف مربوط به مهارت‌های بالینی مورد انتظار از فراگیران تدوین کنند. سپس اهداف و توانمندی‌ها را بر اساس سطح عملکرد مورد انتظار در مقاطع زمانی خاص دوره تعریف کنند. در ادامه لازم است اهداف رفتاری مورد انتظار برای مهارت‌های بالینی اصلی[1] مشخص شوند. این مرحله در پیاده سازی ارزیابی مبتنی بر محل کار

1. Core

بسیار ضروری است زیرا ارزیابان و داوطلبان را در مورد رفتارهایی که باید مورد مشاهده و ارزیابی قرار گیرد هدایت می‌کند. علاوه بر این اهداف رفتاری در انتخاب روش مناسب برای هر موقعیت بالینی خاص بسیار کمک کننده است.

انتخاب ابزارهای ارزیابی مناسب

تاکنون ابزارهای متعددی با هدف ارزیابی مبتنی بر محل کار طراحی شده است (پلگریم و همکاران[1] ۲۰۱۱). بنابراین، دست‌اندرکاران ارزیابی طیفی از گزینه‌ها را برای ارزیابی عملکرد در محیط کار در اختیار دارند. توصیه بر آن است که از ابزارهای موجود استفاده شود و در صورت لزوم با توجه به شرایط اصلاحات مورد نیاز انجام شود. انتخاب ابزار با توجه به رعایت موارد مربوط به سودمندی ابزار و هدف آزمون صورت می‌گیرد. در مواردی که هدف ارزیابی تکوینی است تاثیر آموزشی و ارائه بازخورد بر روایی و پایایی آن اولویت دارد.

ابزارهای خانواده ارزیابی مبتنی بر محل کار متنوع بوده و در ادبیات آموزش پزشکی با نام‌های مختلفی به کار رفته‌اند. در زیر نتایج دو مطالعه مروری که این ابزارها و ویژگی‌های آن‌ها را بررسی کرده است اراه شده است:

پلگریم و همکاران ۲۰۱۱

پلگریم و همکاران در مقاله مروری ابزارهای این دسته از ارزیابی مبتنی بر محل را بررسی کردند و ابزارهای با نام‌های مختلفی یافتند. برخی از این ابزارها توانمندی‌های با محتوای وسیع‌تری را مورد ارزیابی قرار می‌دادند، در نتیجه برای اکثر محیط‌های بالینی مناسب بودند. در حالی‌که دیگر ابزارها توانمندی‌های محدودی را مورد ارزیابی قرار می‌دادند و برای موقعیت‌های بالینی ویژه‌ای به عنوان مثال کارآموزی روان‌پزشکی مناسب بودند. در تمام ابزارها از فرم‌های نمره‌دهی کلی استفاده شد اما برخی محتوا را با جزئیات بیشتری مورد ارزیابی قرار می‌دادند به عنوان مثال ارزیابی مهارت‌های ارتباطی در مقابل پرسیدن سؤالات باز یکی از این موارد است. ابزارها محتوای مختلفی را مورد ارزیابی قرار می‌دادند و به آسانی قابل انطباق با شرایط خاص محیطی بودند. اکثر این ابزارها با هدف ارزیابی تکوینی و ارائه بازخورد طراحی شده‌اند.
سیستم نمره‌دهی مورد استفاده در این ابزارها از حالت دوتایی (مطلوب/نامطلوب) تا لیکرت ۱۱ تایی متغیر بود. در ۸ مورد از این ابزارها معیار مرجع مشخصی برای نمره‌دهی تعریف شده بود. در ۵ مورد معیار مرجع در سطح انتهای دوره (مقطع) و در سه مورد در سطح انتهای دوره (درس) تعریف شده بود. البته در هیچ یک از موارد تعریف چارچوب مرجع بر اساس بررسی متون نبوده و نویسندگان برای حمایت از انتخاب چارچوب مرجع و سیستم نمره‌دهی خود شواهد بسیار ناچیزی را ارائه کرده بودند. برای اکثر ابزارهای این دسته ضریب تعمیم‌پذیری بیش از ۰/۸ با بیش از ۱۰ مواجهه به دست آمد. در اکثر موارد قبل از اجرای این نوع از آزمون‌ها آموزش ارزیابان صورت گرفته بود که بیشترین روش مورد استفاده توضیحات شفاهی و پس از آن برگزاری کارگاه بوده است. تنها یک مطالعه تأثیر این آموزش‌ها را مورد بررسی قرار داده است. اکثر این ابزارها پس از معرفی mini-CEX در سال ۱۹۹۵ میلادی طراحی شدند. اما نکته جالب این است که Clinical skills assessment form خیلی قبل‌تر از mini-CEX در سال ۱۹۸۴ در ارزیابی کارآموزی روان‌پزشکی در دانشگاه مک‌مستر استفاده شد اما احتمالاً به دلیل مساعد نبودن شرایط آن زمان با اقبال عمومی مانند mini-CEX مواجه نشد. در زیر ۱۸ ابزار مطرح شده در ۳۹ مقاله مرور شده توسط پژوهشگران فهرست شده است:

1. Mini clinical evaluation exercise
2. Ophthalmic clinical evaluation exercise
3. Palliative care clinical evaluation exercise
4. Professionalism mini evaluation exercise
5. Competence based assessment, rheumathology
6. Structured clinical observation
7. Patient evaluation assessment form
8. Global rating form in anaesthesiology
9. Ward rating form (in clinical work sampling approach to in-training assessment) (WRF or CWS)
10. Clinical-performance biopsy instrument
11. Clinical evaluation exercise (in emergency medicine training programme)
12. *Clinical skills assessment form, direct observation exercise*
13. *Standardized direct observation assessment tool*
14. *Evaluation of consulting skills (of trainee general practitioners)*
15. *Longitudinal evaluation of performance*
16. *Minicard*
17. *Clinical encounter card (CEC)*
18. *Bedside formative assessment*

1. Pelgrim et al

کوگان و همکاران ۲۰۰۹

این پژوهشگران در مطالعه مروری با هدف بررسی روش‌های ارزیابی مبتنی بر مشاهده مستقیم عملکرد، ۵۵ ابزار موجود را شناسایی کردند. ۳۸ درصد این ابزارها در مقطع پزشکی عمومی، ۵۸ درصد در مقطع دستیاری تخصصی و فوق تخصصی و ۲ مورد در هر دو مقطع به کار رفتند. بخش عمده‌ای از ابزارها در بخش داخلی استفاده شدند. آزمون mini-CEX بیشترین مطالعات را به خود اختصاص داد. این ابزارها هم در بخش بیماران بستری و هم درمانگاه بیماران سرپایی، در برنامه‌های رشته‌های داخلی و جراحی و در مقاطع مختلف پزشکی عمومی و دستیاری تخصصی و فوق تخصصی استفاده شدند. بیشتر ابزارها مهارت‌های شرح‌حال‌گیری، معاینه بالینی و مهارت‌های ارتباطی را مورد ارزیابی قرار دادند. ۱۱ ابزار از ۵۵ مورد توصیفات رفتاری از نمره‌دهی داشتند. ۲۰ ابزار فضایی را برای بازخورد نوشتاری و ترسیم برنامه عملیاتی درنظر گرفته بودند. ابزارها در ۵۸ درصد موارد با اهداف تکوینی، ۱۳ درصد با اهداف تراکمی و در ۵ درصد مورد با هدف تراکمی و تکوینی به کار رفتند. در بسیاری موارد ابزارها تنها یک مرتبه (یک مواجهه) برای هر داوطلب به کار رفتند اما در مواردی هم بیش از ۱۰ بار (ده مواجهه) مورد استفاده قرار گرفتند.

ترویج فرهنگ ارزش‌گذاری ارزیابی مبتنی بر محل کار به ویژه مشاهده مستقیم عملکرد

مشاهده مستقیم عملکرد امر رایجی نیست و ممکن است فرهنگ مؤسسه آموزشی به عنوان یک عنصر لازم به آن اهمیت ندهد. توجه به فرهنگ سازمانی برای موفقیت هر برنامه ارزیابی ضروری است. به منظور تغییر فرهنگ مؤسسه می‌توان از مدل‌های موجود برای تغییر نگرش سازمانی مانند مدل کاتر[1] که به طور موفقیت آمیزی در حوزه علوم پزشکی بکار رفته‌اند استفاده کرد. اقداماتی که در این زمینه می‌توان انجام داد عبارت است از:

- تأکید بر ضرورت مشاهده مستقیم با ارائه شواهدی از تأثیر سوء ضعف در مهارت‌های بالینی بر مراقبت از بیمار و تاثیر بازخورد بر کاهش خطاهای تشخیصی و ارتقای خبرگی[2]
- استفاده از افراد الگو[3] در توانمندسازی اعضای هیأت علمی
- ترسیم و اشاعه یک دورنمای[4] واضح و دقیق در خصوص مشاهده مهارت‌های بالینی به منظور همسو نمودن انتظارات ذی‌نفعان برنامه ارزیابی
- پاداش، تامین مالی و کاهش بار درمان افرادی که به طور مستمر عملکرد فراگیران را مشاهده می‌کنند به منظور افزایش انگیزه این افراد و رفع موانع

طراحی فرم‌های ارزیابی

از موارد دیگری که قبل از اجرای آزمون باید در مورد آن چاره‌ای اندیشید، تهیه فرم‌های ارزیابی عملکرد، نمره‌دهی و ارائه بازخورد است. فرم‌های متعددی متناسب با انواع روش‌های ارزیابی مبتنی بر محل کار طراحی شده است که نحوه نمره‌دهی آن به سه صورت مقیاس درجه‌بندی[5]، چک‌لیست و روبریک[6] است. مقیاس درجه‌بندی و چک‌لیست بیشتر در انواع روش‌های ارزیابی از طریق مشاهده مستقیم و روبریک در نمره‌دهی داده‌های عملکردی مانند کارپوشه به کار می‌رود. به فرم‌های اختصاصی هر ابزار و مزایا و معایب هر یک در فصل‌های مربوط پرداخته می‌شود. در ادامه این فصل در مورد نمره‌دهی روش‌های مشاهده مستقیم با جزئیات بیشتری بحث می‌شود.

قضاوت در مورد عملکرد فراگیران

از مسائل مهمی که در زمان اجرای ارزیابی مبتنی بر محل کار باید در نظر داشت دقت و صحت قضاوت ارزیابان در مورد عملکرد فراگیران و کاهش خطاهای مختلف احتمالی است که در این فرایند رخ می‌دهد. دقت و صحت قضاوت

1. Kotter
2. Expertise
3. Role models
4. Vision
5. Rating scale
6. Rubric

و نمره‌دهی ارزیابان به ویژه در روش‌های ارزیابی از طریق مشاهده مستقیم اهمیت زیادی دارد و بخش زیادی از آن به کیفیت فرم‌های طراحی شده و آموزش اعضای هیأت علمی بستگی دارد که به طور مفصل در فصل‌های مربوط به آن می‌پردازیم. یک سیستم نمره‌دهی روا باید عدم توافق بین ارزیابان را کاهش و قدرت تمایز و افتراق ایشان را افزایش دهد. مطالعات نشان می‌دهد که ارزیابان در ارزیابی عملکرد فراگیران نمی‌توانند همه خطاها را تشخیص دهند و گزارش کنند. نوئل و همکاران[1] (۱۹۹۲) متوجه شدند که اعضای هیأت علمی نتوانستند ۶۸ درصد خطاهای دستیاران مرزی را با مشاهده نوارهای ویدئویی عملکرد آن‌ها تشخیص دهند. زمانی که از چک‌لیست استفاده شد استادان توانستند خطاهای بیشتری را پیدا کنند و این میزان از ۳۲ درصد به ۶۴ درصد رسید. هر چند تشخیص بیشتر خطاها دقت قضاوت ارزیابان را بهتر نکرد. به طوری که دو سوم اعضای هیأت علمی هنوز دستیاران مرزی را مطلوب یا عالی ارزیابی کردند.

قضاوت در مورد عملکرد فراگیران در ابزارهای ارزیابی مشاهده مستقیم را می‌توان از دو بعد زمینه قضاوت[2] و ماهیت قضاوت[3] بررسی کرد:

- **زمینه قضاوت:** زمینه قضاوت به قضاوت ارزیاب بر اساس یک مواجهه خاص یا مواجهه‌های متعددی که طی عملکرد روزمره اتفاق می‌افتد اشاره دارد:
  - یک مواجهه: در این مورد ارزیاب بر اساس مشاهده یک مواجهه در مورد عملکرد فراگیر قضاوت می‌کند. به عنوان مثال عضو هیأت علمی عملکرد فراگیر را در تعامل با یک بیمار یا انجام یک پروسیجر به خصوص بر روی بیمار مشاهده می‌کند و سپس در مورد آن قضاوت می‌کند. از معایب این روش همان ویژگی محتوا[4] یا ویژگی مورد[5] در ارزیابی مهارت‌های بالینی است.
  - عملکرد روزمره: در این مورد ارزیاب بر اساس مشاهداتی که در طول زمان از عملکرد فراگیر به عمل آورده است قضاوت می‌کند. این نوع از قضاوت شایع‌ترین نوع ارزیابی در کشورهای آمریکا و انگلیس است. به این صورت که اعضای هیأت علمی به صورت دوره‌ای فرم‌های نمره‌دهی مربوط به ارزیابی توانمندی‌های فراگیران خود را کامل می‌کنند. مهم‌ترین مزیت این ارزیابی این است که قضاوت بر اساس مشاهده مواجهه‌های گوناگون صورت می‌گیرد. در نتیجه تا حدودی مشکل ویژگی مورد را در ارزیابی عملکرد کاهش می‌دهد. اگر چه گاهی ارزیابان جنبه‌هایی از عملکرد را بدون مشاهده آن ارزشیابی می‌کنند. برای مثال پولیتو و همکاران[6] (۲۰۰۶) نشان دادند که اصولاً اعضای هیأت علمی فقط مهارت‌های شناختی و تعهد حرفه‌ای را مشاهده می‌کنند و جنبه‌های دیگر توانمندی را مشاهده نمی‌کنند. سیلبر و همکاران[7] (۲۰۰۴) نیز متوجه شدند که اعضای هیأت علمی با استفاده از فرم نمره‌دهی کلی تنها دو حیطه از توانمندی‌های ACGME شامل دانش پزشکی و مهارت‌های بین فردی را می‌سنجند.
- **ماهیت قضاوت:** ماهیت قضاوت به قضاوت در خصوص رخداد[8]، کیفیت[9] یا تناسب[10] رفتار مورد ارزیابی اشاره دارد:
  - رخداد: گاهی اوقات مشاهده رخداد یا عدم رخداد یک رفتار مورد سؤال است. در این صورت ارزیابان با استفاده از یک چک‌لیست وقوع رفتار را مشخص می‌کنند. در چک‌لیست فقط وجود، عدم وجود یا کمیت ویژگی مورد نظر ثبت می‌شود. نمونه آن چک‌لیست مورد استفاده در آزمون OSAT[11] است که به طور معمول به منظور ارزیابی مهارت‌های تکنیکی به کار می‌رود. این نوع از ارزیابی عینی و ساختارمند است و از ثبات و پایایی در مشاهدات مختلف و برای ارزیابان مختلف برخوردار است.

---

1. Noel et al.
2. Grounds for Judgement
3. Nature of the Judgement
4. Content specificity
5. Case specificity
6. Pulito et al.
7. Silber et al.
8. Occurance
9. Quality
10. Fitness
11. Objective Structured Assessment of Technical Skill (OSATS)

- کیفیت: در این روش قضاوت ارزیاب به صورت مقیاس درجه‌بندی ثبت می‌شود و قضاوت به صورت درجه یا مقدار رفتار یا ویژگی مورد نظر تعیین می‌گردد. نمونه آن فرم‌های با مقیاس درجه‌بندی لیکرت یا دیگر انواع مقیاس نمره‌دهی است که به عنوان مثال در آزمون mini-CEX مورد استفاده قرار می‌گیرد.
- تناسب: گاهی از ارزیاب خواسته می‌شود در مورد رضایت بخش بودن عملکرد یا تناسب آن با هدف قضاوت کند. در اغلب موارد ارزیابان ابتدا در مورد کیفیت عملکرد قضاوت کرده و سپس در مورد تناسب عملکرد با هدف مورد نظر مانند رد یا قبولی فراگیران تصمیم‌گیری می‌کنند.

به طور خلاصه، فرم‌های مختلفی به منظور ساختارمند نمودن قضاوت ارزیابان و در نتیجه افزایش دقت و صحت آن طراحی شده است. با وجود این، در روش‌های ارزیابی مبتنی بر محل کار قضاوت ارزیابان نسبت به عملکرد داوطلب یک قضاوت انسانی است. در نتیجه عوامل بسیاری از ویژگی‌های خود ارزیاب به عنوان یک فرد تا محیط اجتماعی که در آن ارزیابی رخ می‌دهد بر قضاوت صورت گرفته تأثیرگذارند. به عنوان مثال، بر اساس مدلی که کوگان و همکاران (۲۰۱۱) ارائه دادند ارزیابان که در اکثریت موارد اعضای هیأت علمی هستند مجموعه‌ای از ویژگی‌های فردی شامل سن و جنس، تجارب و توانمندی بالینی، آموزشی و ارزیابی، و همچنین نگرش و احساسی که نسبت به فرایند مشاهده و بازخورد دارند را با خود به محیط ارزیابی می‌آورند. هر یک از این خصوصیات فردی به نوعی فرایند قضاوت ارزیابان را تحت تأثیر قرار می‌دهند که در مورد تاثیر برخی از آن‌ها نتایج پژوهش‌ها شفاف بوده (به فصل mini-CEX بخش نمره‌دهی مراجعه شود) و در موارد دیگر نیاز به پژوهش‌های بیشتری وجود دارد. به عنوان مثال، گوارتز و همکاران[1] (۲۰۱۱) در مطالعه خود به نقش تجربه قبلی در ارزیابی بر قضاوت ارزیابان پرداختند و نشان دادند ارزیابان باتجربه مشکلات عملکردی داوطلبان را سریع‌تر پیدا می‌کردند و در این مورد استنتاج‌های بیشتری می‌کردند. علاوه بر این، به احتمال بیشتری عوامل محیطی را در قضاوت در نظر می‌گرفتند. آن‌ها نتیجه گرفتند که ارزیابان با تجربه از شِمای[2] ذهنی با جزئیات بیشتری برخوردار هستند.

کوگان و همکاران (۲۰۱۱) در ادامه مدل خود نشان دادند که ارزیابانی که با یکسری ویژگی‌ها و تجربیات فردی وارد محیط ارزیابی می‌شوند مواجهه دستیار با بیمار را از طریق دو لنز مورد مشاهده قرار می‌دهند. یکی از لنزها چارچوب مرجعی[3] است که اعضای هیأت علمی برای مقایسه عملکرد دستیار از آن استفاده می‌کنند و دیگری معنی و تفسیری است که آن‌ها به این مشاهده می‌دهند. به نظر می‌رسد آنچه که ارزیابان به عنوان مرجع برای مقایسه عملکرد داوطلب استفاده می‌کنند یکسان نیست که به نوبه خود منجر به تغییرپذیری قضاوت ایشان می‌شود. در اغلب موارد ارزیابان از یکی از سه منبع زیر به عنوان استاندارد برای مقایسه عملکرد داوطلب استفاده می‌کنند:

- عملکرد خود
- عملکرد دیگر داوطلبان و پزشکان در حال طبابت
- عملکرد استاندارد مورد نیاز برای مراقبت از بیمار

علاوه بر آنچه که ارزیابان به عنوان معیاری برای مقایسه انتخاب می‌کنند تحلیل و تفسیرهای ایشان از رفتار داوطلب یا به عبارت دیگر استنباط[4] ایشان از عملکرد داوطلب نقش بسیار مهمی در روند قضاوت ایفا می‌کند. به این معنا که ارزیابان رفتارهای داوطلب را می‌بینند و از میان آن مواردی را به صورت آگاهانه و ناآگاهانه انتخاب می‌کنند. به این رفتارها معنا می‌دهند و آن را تفسیر می‌کنند. مشخص شده است که ارزیابان مختلف با مشاهده رفتار یکسان از یک داوطلب در برخورد با بیمار (شرایط کاملاً یکسان) تفاسیر متعددی می‌کنند. اغلب تفسیرهایی که ارزیابان از عملکرد داوطلب به عمل می‌آورند در ارتباط با ویژگی‌های شخصیتی داوطلب، مهارت و احساسات (مثلاً، اعتماد به نفس و راحتی) وی، انگیزه داوطلب برای پیشرفت، تجربه قبلی و میزان آمادگی وی

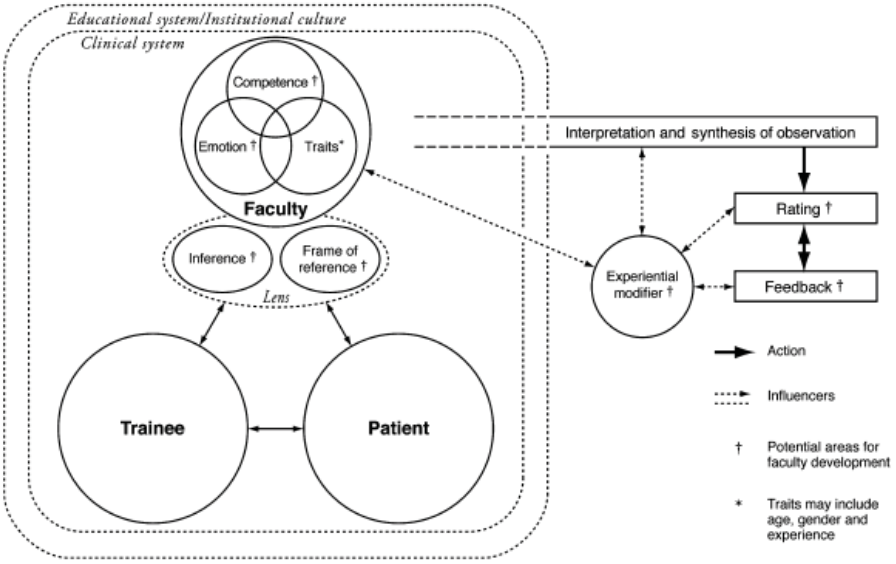
1. Govaerts et al.
2. Schema
3. Frames of reference
4. Inferences

است. نکته اینجاست که همه این مشاهدات و تفسیرها در یک سیستم بالینی و فراتر از آن در بستر یک موسسه آموزشی با فرهنگ منحصر به فرد خود رخ می‌دهد و در نتیجه شرایط محیطی مذکور بر مشاهده و تفسیر عملکرد داوطلب تاثیر می‌گذارد.

در مرحله بعد اعضای هیأت علمی مشاهده و قضاوت مبتنی بر آن را به نمره تبدیل می‌کنند. در این مرحله نیز ارزیابان از رویکردهای متفاوتی استفاده می‌کنند. برخی از حیطه‌های مختلف توانمندی درج شده در فرم‌های نمره‌دهی میانگین می‌گیرند در حالی‌که دیگران به صورت غیرجبرانی[1] نمره می‌دهند. برخی وزن بیشتری از نمره را به توانمندی مربوط به تمرکز مواجهه بالینی می‌دهند. به عنوان مثال اگر تمرکز مواجهه بر مهارت ارتباطی باشد و داوطلب مهارت ارتباطی مناسبی نداشته باشد، از نظر این دسته از ارزیابان نمی‌تواند پزشک باصلاحیتی باشد و در نتیجه عملکرد او را ضعیف ارزیابی می‌کنند، هر چند در دیگر توانمندی‌ها خوب عمل کرده باشد. در مواردی هم ارزیابان در تبدیل قضاوت خود به نمره با ابهام مواجهه هستند و قادر نیستند بین طیف اعداد تمایز قائل شوند.

لازم به ذکر است در روند تبدیل قضاوت به نمره هم عوامل دیگری به غیر از عملکرد داوطلب نمره‌دهی را تحت تأثیر قرار می‌دهد که از جمله آن می‌توان به شرایط محیطی (ارتباط ارزیاب و داوطلب، پیچیدگی مواجهه و میزان آشنا بودن داوطلب با مورد بالینی) و پاسخ به بازخورد (از سوی داوطلب، خودِ ارزیاب و مؤسسه) اشاره کرد. به عنوان مثال، اگر ارزیاب با داوطلب (دستیار) در طولانی مدت در ارتباط باشد و داوطلب در زمان مواجهه توصیه‌ها و آموخته‌های معلم خود را به کار نبندد این موقعیت موجب سخت‌گیری در قضاوت ارزیاب می‌شود. در حالی‌که یک ارتباط مثبت قبلی با داوطلب منجر به سهل‌گیری در قضاوت می‌شود. در خصوص بازخورد نیز گاهی ارزیابان با توجه به برخورد احساسی داوطلبان نمره آن‌ها را تغییر می‌دهند (معمولاً افزایش می‌دهند). فضای حاکم بر مؤسسه نیز بر نمره‌دهی ارزیابان تأثیرگذار است. در صورتی‌که جو حاکم به نحوی باشد که افرادی که نمره پایین می‌دهند را حمایت نکند و حتی مورد نکوهش قرار دهد، ارزیابان از دادن نمره پایین امتناع می‌کنند.

خود این فرایند تجربه‌ای برای مشاهدات و ارزیابی‌های بعدی ارزیاب فراهم می‌آورد. به طور خلاصه قضاوت و نمره‌دهی ارزیابان تحت تاثیر عوامل متعددی قرار می‌گیرد (شکل ۲-۱۹). ماهیت و نحوه تاثیر این عوامل هنوز به خوبی شفاف نشده است و نیازمند پژوهش‌های بیشتر در این زمینه و در محیط‌های آموزشی مختلف است. تعیین این عوامل از این جهت اهمیت دارد که عواملی که قابل اصلاح و آموزش هستند شناسایی و کنترل شوند.



شکل ۲-۱۹: عوامل مؤثر بر قضاوت ارزیابان در مشاهده مستقیم عملکرد (کوگان و همکاران ۲۰۱۱)

1. non-compensatory

ارائه بازخورد بر اساس عملکرد مشاهده شده

یکی از بخش‌های مهم آزمون‌های مبتنی بر محل کار ارائه بازخورد است. بازخورد بخش مهمی از فرایند ارتقای مهارت‌های بالینی است. اکثر استادان بالینی با فنون و مفاهیم ارائه بازخورد آشنا هستند اما اغلب از آن به عنوان یک ابزار آموزشی با ارزش استفاده نمی‌کنند. نتایج مطالعات نشان می‌دهد دستیاران پزشکی در آمریکا تقریباً هیچ نوع بازخوردی دریافت نمی‌کنند. از عواقب این وضعیت کاهش اعتماد به نفس برخی از فراگیران و شکل‌گیری نادرست توانمندی‌ها در دیگر فراگیران است (نورسینی و برچ ۲۰۰۷). علل احتمالی استفاده کم از بازخورد در محیط بالینی شامل موارد زیر است:

- نیاز به مشاهده عملکرد فراگیر
- دغدغه تأثیر بازخورد منفی بر فراگیر و بر رابطه فراگیر و آموزش دهنده
- لزوم آشنایی با اصول یادگیری بزرگسالان و فنون ارائه بازخورد

چند توصیه برای بازخورد مؤثر شامل موارد زیر است:

- زمان مناسبی برای بازخورد در نظر بگیرید (زمانی برابر ۵ تا ۱۰ دقیقه بلافاصله بعد از جلسه ارزیابی).
- مکان مناسبی برای بازخورد در نظر بگیرید (در مکانی خصوصی و به دور از رفت و آمد و احتمال مزاحمت).
- قبل از ارائه بازخورد بر افکار و احساسات خود غلبه کنید.
- جوی دوستانه، توأم با احترام و غیرتهدیدآمیز ایجاد کنید.
- بر اساس اهداف تعریف شده و اطلاع‌رسانی شده بازخورد دهید.
- از زبان غیرقضاوتی استفاده کنید.
- به طور اختصاصی در مورد رفتارهای داوطلب بازخورد دهید؛ نه عملکرد کلی وی.
- در مورد تصمیمات و اعمال داوطلب بازخورد دهید؛ نه تفسیر شخصی خود از باورها و انگیزه‌های وی.
- در هر بار بازخورد توصیه‌های قابل‌فهم کوتاه ارائه دهید.
- فراگیر را در فرایند ارائه بازخورد سهیم کنید.
- در انتها پیشنهادهایی به منظور اصلاح ارائه دهید.

سودمندی ابزارهای ارزیابی مبتنی بر محل کار

در این بخش سودمندی روش‌های ارزیابی مبتنی بر محل کار بر اساس معیارهای ون‌درولوتن[1] و به صورت کلی بررسی می‌شود و سودمندی هر روش در فصل مربوط با جزئیات بیشتر بحث می‌شود. لازم به ذکر است اکثر روش‌های ارزیابی مبتنی بر محل کار جدید هستند و شواهد زیادی در مورد سودمندی آن وجود ندارد. اگر پژوهشی هم در این ارتباط صورت گرفته است در بسیاری از موارد دارای حجم نمونه پایین بوده و شواهد محکمی در اختیار ما قرار نمی‌دهد.

پایایی ارزیابی مبتنی بر محل کار

مفهوم پایایی در آزمون‌های مبتنی بر محل کار به این صورت است که اگر فراگیران یکسان در مواجهه‌های مختلف با بیماران متفاوت و توسط ارزیابان متفاوت ارزیابی شوند، ما انتظار داریم نتایج ایشان مشابه باشد. این عبارت نمایانگر مفهوم پایایی یا تکرارپذیری است. سه عامل اصلی بر پایایی مشاهده عملکرد فراگیر در محیط واقعی تأثیرگذار است:

- تعداد مشاهدات (هم در قضاوت مبتنی بر یک مواجهه و هم در عملکرد روزمره)
- تعداد ارزیابان
- تعداد جنبه‌های مورد ارزیابی عملکرد

1. Van der Vleuten

یکی از مسائلی که روایی و پایایی آزمون‌های مربوط به ارزیابی توانمندی‌های بالینی را با چالش مواجه می‌کند، ویژگی مورد یا ویژگی محتوا است. ویژگی مورد یا محتوا به این معنی است که عملکرد در یک رخداد یا وظیفه[1]، پیش‌گویی کننده عملکرد در موقعیت‌های دیگر نیست. در واقع، ارزیابی عملکرد پزشکان وابسته به مورد بیماری یا وظیفه در حال انجام است. به منظور غلبه بر این مشکل و به دست آوردن تخمین قابل‌تعمیمی از عملکرد داوطلبان لازم است مواجهه هر فراگیر با بیماران مختلف مورد مشاهده قرار گیرد تا در مورد نتایج ارزیابی وی اطمینان حاصل شود. در نتیجه در برنامه‌های ارزیابی معتبر مانند برنامه پیش‌دستیاری[2] کشور انگلیس از چندین مواجهه برای ارزیابی عملکرد فراگیر استفاده می‌شود.

در انواع روش‌های ارزیابی مبتنی بر محل کار، بیماران مختلف یا پرونده‌های بیماران مختلف مورد استفاده قرار می‌گیرد که از لحاظ پیچیدگی و سطح دشواری متفاوت هستند. همچنین اعضای هیأت علمی و افراد مختلف به عنوان ارزیاب استفاده می‌شوند که از لحاظ سخت‌گیری متفاوت هستند. علاوه بر این، قضاوت ارزیابان یکسان نیز در موقعیت‌های مختلف از تغییرپذیری قابل ملاحظه‌ای برخوردار است. در نتیجه مشخص نیست که تفاوت نتایج ارزیابی فراگیران به دلیل تفاوت در توانایی آن‌ها است یا به دلیل تفاوت در دشواری مواجهه‌ها و قضاوت ارزیابان است. البته این مسأله در زمانی که حساسیت آزمون کمتر است و سنجش تراکمی در اولویت قرار ندارد، مشکل‌زا نیست. در هر حال، استفاده از فهرست مشکلات بالینی یکسان، افزایش تعداد و تنوع ارزیابان برای هر فراگیر و برگزاری برنامه‌های توانمندسازی برای اعضای هیأت علمی تأثیر این مشکلات را کمتر می‌کند.

آخرین نکته این که هر چه تعداد جنبه‌های عملکرد که مورد ارزیابی قرار می‌گیرند، بیشتر باشد، پایایی آزمون افزایش می‌یابد. البته افزایش جنبه‌های عملکرد تا حد مشخصی منجر به بهبود پایایی می‌شود و افزایش بیش از حد آن نه تنها پایایی را آنچنان تغییر نمی‌دهد بلکه قابلیت اجرای آزمون را نیز کاهش می‌دهد. در واقع تعداد دقیق سؤالات بستگی به خصوصیات عملکرد و ماهیت قضاوت دارد. مثلاً قضاوت در مورد کیفیت و تناسب عملکرد به سؤالات بیشتر و قضاوت برای رخداد به سؤالات کمتری نیاز دارد. اما به صورت کلی ۵ تا ۱۰ سؤال کافی است.

عوامل دیگری مانند واژه‌های مورد استفاده در سؤالات و تعداد گزینه‌های مقیاس نمره‌دهی تأثیر کمی بر پایایی دارد. اما به این دلیل که این عوامل به آسانی قابل اصلاح هستند، برخی از استفاده‌کنندگان فرم‌های ابزارهای مبتنی بر محل کار تلاش زیادی را صرف این موضوع می‌کنند. در صورتی که توصیه می‌شود این زمان و تلاش صرف تربیت و آموزش ارزیابان شود.

روایی ارزیابی مبتنی بر محل کار

روش‌های ارزیابی مبتنی بر محل کار به این دلیل که تعامل فراگیر با بیمار را در محیط واقعی و در طول یک دوره زمانی مورد مشاهده قرار می‌دهند دارای روایی قابل ملاحظه‌ای هستند. همان‌طور که در مبحث پایایی عنوان شد به منظور غلبه بر مشکل ویژگی موارد، لازم است نمونه‌گیری وسیع از مواجهه‌های بالینی مختلف و در شرایط گوناگون صورت پذیرد. بنابراین، استفاده از چندین ارزیاب در یک دوره زمانی طولانی و قضاوت در مورد چندین مواجهه توصیه می‌شود. علاوه بر این، به منظور اطمینان از یکپارچگی ارزیابی با برنامه درسی باید پوشش مناسبی از توانمندی‌ها صورت پذیرد. مسلم است که یک روش ارزیابی واحد نمی‌تواند به تنهایی این مشکل را حل کند و در نتیجه «جعبه‌ای از ابزارها»[3] مورد نیاز است. در واقع، تلفیقی از انواع روش‌های ارزیابی مبتنی بر محل کار و نیز دیگر روش‌های ارزیابی مورد نیاز است تا توانمندی‌های بالینی را پوشش دهد. در این میان، برخی از توانمندی‌ها با روش‌های دیگر به نحو مطلوبی ارزیابی می‌شوند. به عنوان مثال، دانش بالینی را می‌توان با آزمون‌های کتبی ارزیابی کرد. در حالی که ارزیابی جنبه‌هایی از توانمندی‌ها مانند کار گروهی،

1. Task
2. Foundation programme
3. Tools Box

رهبری و تعهد برای ارتقای مداوم حرفه‌ای عملاً با روش‌های دیگر به غیر از ارزیابی مبتنی بر محل کار امکان‌پذیر نیست.

یکی از مشکلاتی که در عمل در ارزیابی توانمندی‌ها بوجود می‌آید این است که به منظور افزایش عینیت آزمون دچار تقلیل‌گرایی[1] می‌شویم و توانمندی‌ها را بسیار ریز و جزئی می‌کنیم. این موضوع موجب صدمه به روایی آزمون می‌شود. بنابراین، ضروری است در طراحی ابزارها و فرم‌های آزمون‌های مبتنی بر محل کار توانمندی‌های مورد ارزیابی به صورت عبارت‌های کلی نوشته شوند تا روح حاکم بر آن آسیب نبیند.

به طور خلاصه، اگر چه در ارزیابی مبتنی بر محل کار با انتقال ارزیابی دانشجو از سالن‌های امتحانات به محیط واقعی سعی شده است روایی ارزیابی فراگیر افزایش یابد اما هنوز پژوهش‌هایی که این موضوع را تأیید کند کافی نبوده و در واقع مسائل مبهم زیادی در این ارتباط وجود دارد. عوامل بسیاری در محیط واقعی رخ می‌دهند که بسیاری از آن‌ها قابل پیش‌بینی یا در اختیار ارزیابان نبوده اما بر قضاوت و ارزیابی ایشان تأثیر می‌گذارد. به برخی از این عوامل در قسمت قضاوت در ارزیابی مبتنی بر محل کار اشاره شد و به برخی دیگر متناسب با هر ابزار در فصل‌های بعدی اشاره می‌شود. به هر حال، پژوهش‌های بیشتری مورد نیاز است تا ماهیت این عوامل و نحوه تأثیر آن را شناسایی کند.

تأثیر آموزشی ارزیابی مبتنی بر محل کار

همان‌طور که شواهد نشان می‌دهند، ارزیابی محرک یادگیری است و ارزیابی در محیط واقعی باعث تحریک فراگیران به یادگیری از طریق بیمار می‌شود. ارزیابی مبتنی بر محل کار دارای اثر آموزشی بالا اما پایایی کمتر از آزمون‌های ساختارمند مانند سؤالات چندگزینه‌ای است. بر خلاف آزمون‌های استاندارد که در آخر دوره اجرا می‌شوند، ارزیابی مبتنی بر محل کار فرصتی را برای ادغام فعالیت‌های تدریس، یادگیری و ارزیابی فراهم می‌آورد. البته این هدف در صورتی محقق می‌شود که اولاً عملکرد فراگیر مورد مشاهده قرار گیرد و سپس بر اساس آن بازخورد مناسب ارائه شود.

اکثر مطالعات انجام شده تأثیر آموزشی این آزمون‌ها را با بررسی نگرش فراگیران و اعضای هیأت علمی بررسی نموده‌اند و مطالعات بسیار کمی به بررسی اثر آزمون‌ها در ارتقای مهارت‌های بالینی و مراقبت از بیمار پرداخته‌اند. در مطالعه مروری که اوریم و همکاران[2] (۲۰۰۷) انجام دادند، ۱۹ مطالعه اثر مثبت آموزشی روش‌های ارزیابی مبتنی بر محل کار را گزارش کردند و ۲ مطالعه هیچ اثر آموزشی را گزارش نکردند. هشت مطالعه تأثیر آموزشی را در سطح یک هرم کرک‌پاتریک یعنی «رضایت فراگیران» گزارش کردند. چهار مطالعه اثر مثبت آموزشی را در سطح دوم کرک‌پاتریک به صورت بهبود «یادگیری» گزارش کردند. ۱۲ مطالعه بهبود «عملکرد فراگیران» یعنی سطح سوم هرم کرک‌پاتریک را گزارش کردند و هیچ مطالعه‌ای پیامدهای «مراقبت از بیمار و اثر بر سیستم بهداشتی درمانی» یعنی سطح چهار هرم کرک‌پاتریک را بررسی نکرده بود. لازم به ذکر است ۱۲ مطالعه مربوط به سطح سه هرم کرک‌پاتریک نیز بهبود عملکرد را با بررسی نظر فراگیران نسبت به عملکرد خود ارزیابی کردند.

قابلیت اجرا و هزینه ارزیابی مبتنی بر محل کار

بیشتر مطالعاتی که به بررسی قابلیت اجرای این آزمون‌ها پرداخته‌اند، میزان پر کردن فرم‌ها و میزان رضایت استفاده‌کنندگان و در مواردی نیز زمان صرف‌شده را مدنظر قرار داده‌اند. اکثر مطالعات قابلیت اجرای نسبتاً خوبی را گزارش کرده‌اند. به عنوان مثال، ویلکینسون و همکاران[3] (۲۰۰۸) سه روش ارزیابی mini-CEX، DOPS و MSF را با هم مقایسه کردند و قابلیت اجرای مناسب را برای این روش‌ها گزارش کردند. در این میان mini-CEX کمترین میزان قابلیت اجرا را داشت و طولانی‌ترین آزمون بود. با وجود این، در برخی موارد شواهد حمایت‌کننده قابلیت اجرای مناسب ابزارهای مبتنی

1. Reductionist Approach
2. Overeem et al.
3. Wilkinson et al.

بر محل کار نیست. یکی از دلایل تفاوت نتایج مطالعات آن است که از معیارهای یکسانی به منظور بررسی قابلیت اجرای آزمون‌ها استفاده نشده است. به عنوان مثال، برخی از مطالعات زمان اجرای آزمون را مجموع زمان مشاهده و ارائه بازخورد در نظر گرفته‌اند و برخی دیگر زمانی که صرف تدوین فرم‌ها، هماهنگی، اجرای آزمون و مستندسازی نتایج می‌شود را نیز در نظر گرفته‌اند. مسأله دیگر این است که اکثر مطالعات به قابلیت اجرای این آزمون‌ها در کوتاه‌مدت پرداخته‌اند، در حالی که حفظ و اجرای با کیفیت آنها در طولانی‌مدت اهمیت بسیار دارد.

در مجموع، معرفی و اجرای ارزیابی مبتنی بر محل کار زمان‌بر، منوط به صرف منابع و نیازمند فرهنگ‌سازی است. آزمون‌های مختلف از این نظر به طور قابل ملاحظه‌ای با هم تفاوت دارند. در مطالعه‌ای مروری که اوریم و همکارانش (۲۰۰۷) انجام دادند ارزیابی توسط همتایان[1] از بالاترین اثربخشی با صرف زمان یک ساعت برای هر پزشک و کارپوشه و ممیزی از پایین‌ترین اثربخشی با صرف زمان ۴۰ ساعت برای هر پزشک برخوردار بودند. همچنین در این مطالعه کمترین هزینه برای ارزیابی توسط همتایان و بیشترین هزینه برای کارپوشه، ممیزی و ضبط ویدئویی گزارش شدند.

مقبولیت ارزیابی مبتنی بر محل کار

درصورتی که یک روش ارزیابی، مورد پذیرش دانشجویان و اعضای هیأت علمی واقع نگردد، سرانجام مطرود می‌شود. این مسأله به ویژه در مورد ارزیابی‌های مبتنی بر محل کار که ارزش ارزیابی به استفاده‌کنندگان آن وابسته است تا خود ابزار، نمود بیشتری دارد. درنتیجه کسب اطلاعات در مورد عقاید و دیدگاه‌های ارزیابان و آزمون شوندگان یکی از راهبردهای مهم در پیاده‌سازی این ابزارها است. در این مورد نیز ابزارهای ارزیابی مبتنی بر محل کار متفاوت هستند که در بخش سودمندی مربوط به هر ابزار به آن پرداخته می‌شود.

ابزارهای خانواده ارزیابی مبتنی بر محل کار متنوع بوده و در ادبیات آموزش پزشکی با نام‌های مختلفی به کار رفته‌اند. در زیر نتایج دو مطالعه مروری که این ابزارها و ویژگی‌های آن‌ها را بررسی کرده است ارائه شده است:

پلگریم و همکاران ۲۰۱۱

پلگریم و همکاران در مقاله مروری ابزارهای این دسته از ارزیابی مبتنی بر محل را بررسی کردند و ابزارهای با نام‌های مختلفی یافتند.
در مورد مقبولیت و قابلیت اجرای این ابزارها مطالعات کیفیت خوبی را گزارش کرده‌اند اما اکثر آن‌ها معیار مشخصی را ذکر نکرده‌اند و تنها به بررسی میزان تکمیل فرم‌ها و رضایت استفاده‌کنندگان پرداخته‌اند. نتایج آن‌ها بسیار متنوع بوده است و از ۹۶ و ۱۰۰ درصد در برخی از مطالعات تا ۲۳ درصد در مورد روش ارزیابی Clinical work sampling (CWS) متغیر بوده است. در مجموع نتایجی که در مورد قابلیت اجرای این آزمون‌ها در دسترس است در مورد روش‌های ارزیابی غیر از mini-CEX تنها به یک مطالعه اکتفا شده است.
در مورد mini-CEX هم که مطالعات بیشتری در اختیار است، نتایج گزارش شده ضد و نقیض است.
پژوهشگران نتایج مربوط به پایایی مطالعات را با هم ادغام کردند و با استفاده از ضریب اسپیرمن پایایی کل را به دست آوردند. نتایج نشان داد که با بیش از ده مواجهه ضریب پایایی بیش از ۰/۸ به دست می‌آید. پژوهشگران نشان دادند که مطالعات سیستماتیک بیشتری مورد نیاز است تا منبع واریانس در پایایی این نوع آزمون‌ها را نشان دهد تا بر اساس آن بتوان در مورد تعداد مناسب ارزیاب و مواجهه تصمیم‌گیری کرد. علاوه بر این نتایج مطالعات در خصوص تأثیر آموزش ارزیابان بر پایایی بین ارزیابان متناقض بود که پژوهشگران پیشنهاد کردند مطالعات بیشتری برای بررسی اثر آموزش بر پایایی بین ارزیابان و تولید نتایج متقاعد کننده مورد نیاز است. در مورد بیشتر ابزارهای این گروه بررسی از نظر روایی انجام نشده بود. در این میان، روایی mini-CEX به ویژه روایی پیش‌بین آن مورد بررسی قرار گرفته است که ارتباط خوبی با دیگر آزمون‌های ارزیابی مهارت‌های بالینی داشت. روایی پیش‌بین clinical encounter cards (CEC) نیز مناسب گزارش شده است. در مورد برخی از ابزارها هم روایی سازه خوبی گزارش شد. تأثیر آموزشی این ابزارها تنها در سطح رضایت ارزیابان و فراگیران بررسی شده است. با توجه به این‌که هدف اصلی این آزمون‌ها ارتقای یادگیری است ضروری است در خصوص تأثیر این آزمون‌ها بر یادگیری و مهارت‌های بالینی پژوهش‌های بیشتری انجام شود.

1. Peer assessment

کوگان و همکاران ۲۰۰۹

این پژوهشگران در مطالعه مروری با هدف بررسی روش‌های ارزیابی مبتنی بر مشاهده مستقیم عملکرد، ۵۵ ابزار موجود را شناسایی کردند. در ۳۶ درصد موارد روایی محتوا مورد بررسی قرار گرفت که از طریق مرور توانمندی‌ها و مرور متون توسط متخصصان و رسیدن به اجماع بود. در ۴۷ درصد موارد آموزش ارزیابان صورت گرفت که اغلب کوتاه، از ۱۰ دقیقه تا ۳ ساعت بود. آموزش از طریق پست الکترونیکی، کارگاه و برنامه های روتین موسسه در زمینه توانمندسازی صورت گرفت. در مورد ۸ ابزار، جلسات آموزشی به صورت مشاهده فیلم‌های ویدئویی تعامل فراگیر در سطوح مختلف عملکرد با بیمار بود. در زمینه پایایی بیشترین پایایی مورد اندازه‌گیری پایایی بین ارزیابان بود که در اکثر موارد زیر ۰/۷ گزارش شد. پایایی درون ارزیابان برای یک ابزار و آزمون-آزمون مجدد نیز برای یک ابزار دیگر مورد بررسی قرار گرفت. ثبات درونی فرم‌ها بالا گزارش شدند. در مورد پیامدها اکثر ابزارها در حد رضایت استفاده‌کنندگان گزارش کردند اما دو مطالعه تغییر و بازنگری کوریکولوم را بر اساس نتایج آزمون گزارش کردند.

سؤالات رایج در ارزیابی مبتنی بر محل کار

آیا بازخورد بر یادگیری فراگیران تأثیر مثبت دارد؟

بیشتر پژوهش‌هایی که در این زمینه وجود دارد مربوط به محیط کلاس درس است. هاتی (۱۹۹۹) در مطالعه خود نشان داد که اندازه تأثیر[1] بازخورد بسیار بزرگ است (ES =۰/۷۹) و در میان چهار عامل اصلی تأثیرگذار بر پیشرفت تحصیلی فراگیران قرار دارد. وی همچنین نشان داد این تأثیر بسته به نوع بازخورد بسیار متغیر است و بیشترین تأثیر مربوط به ارائه اطلاعات در مورد یک وظیفه و فعالیت خاص است.

اطلاعات برای پاسخ به این سؤال در حوزه آموزش پزشکی محدودتر است. در متاآنالیزی که ولوسکی و همکاران[2] (۲۰۰۶) بر روی ۴۱ مطالعه در این حوزه انجام دادند، ۷۴ درصد مطالعات اثر مثبت بازخورد را بر یادگیری نشان دادند.

هاتی ۱۹۹۹

هاتی ۱۹۹۹
وی اطلاعات ۵۰۰ متاآنالیز را که شامل ۱۸۰۰ مطالعه و ۲۵ میلیون دانش‌آموز بود سنتز کرد و نشان داد که اندازه اثر نوع سیستم آموزشی بر پیشرفت دانشجویان در حدود ۰/۴۰ است (به عنوان مثال، سیستم آموزشی نوآورانه موجب افزایش میانگین نمره آزمون پیشرفت تحصیلی تا ۰/۴۰ یک انحراف معیار می‌شود). اگر این عدد را به عنوان استاندارد طلائی برای قضاوت در مورد عوامل مختلف تأثیرگذار بر عملکرد در نظر بگیریم، هاتی در مرحله بعد ۱۲ متاآنالیز که به طور خاص به تأثیر بازخورد می‌پرداختند را بررسی کرد. نتایج این بررسی نشان داد که اندازه تأثیر بازخورد نسبت به سیستم آموزشی بسیار بزرگ (ES=۰/۷۹) است و در میان چهار عامل اصلی تأثیرگذار بر پیشرفت تحصیلی فراگیران قرار دارد. وی همچنین نشان داد این تأثیر بسته به نوع بازخورد بسیار متغیر است و بیشترین تأثیر مربوط به ارائه اطلاعات در مورد یک وظیفه و فعالیت خاص است

آیا تعداد مواجهه مورد نیاز برای همه فراگیران یکسان است؟

در بیشتر برنامه‌های ارزیابی فراگیر، برای همه فراگیران تعداد مواجهه یکسان و ثابتی در نظر گرفته می‌شود. به عنوان مثال، در برنامه پیش‌دستیاری انگلیس برای هر فراگیر تعداد چهار تا شش مواجهه mini-CEX، DOPS و CBD در طول یک سال در نظر گرفته می‌شود. تخمین تعداد مواجهه در چنین برنامه‌هایی از همان روش سنتی تخمین با در نظر گرفتن تعادل پایایی در برابر قابلیت اجرا ناشی می‌شود. هر چند، بسته به هدف ارزیابی، ممکن است تعداد مواجهه‌ها به این روش برای یک فراگیر مشخص زیاد یا کم باشد. استفاده از خطای معیار اندازه‌گیری[3] (SEM) باعث می‌شود بتوانیم راهبرد بهتری برای تصمیم‌گیری در این زمینه به کار بندیم. توضیحات کامل در مورد این مفهوم در بخش هشتم کتاب آمده است اما در اینجا سعی می‌کنیم با ارائه یک مثال، موضوع را روشن کنیم:

1. Effect size (ES)
2. Veloski
3. Standard error of measurement (SEM)

داده‌های مربوط به یک مطالعه در مورد آزمون mini-CEX استخراج شده است. این داده‌ها بر اساس فرم‌های نمره‌دهی نه‌تایی به دست آمده است که در آن ۱ تا ۳ «نامطلوب»، ۴ تا ۶ «مطلوب» و ۷ تا ۹ «عالی» در نظر گرفته شدند. اکنون سؤال این است که وضعیت فراگیر با نمره چهار چگونه است. اگر صرفاً نمره خام وی در نظر گرفته شود، وضعیت مطلوبی دارد اما این نمره، نمره واقعی[1] وی نیست و خطای اندازه‌گیری نیز باید لحاظ شود. پس از محاسبه SEM، فاصله اطمینان را مشخص می‌کنیم. آنالیزها نشان می‌دهد که ۹۵ درصد فاصله اطمینان برای نمره کلی توانمندی بالینی بعد از دو مواجهه ۱/۲± و برای چهار مواجهه ۰/۸± است و با افزایش مواجهه‌ها این رقم به میزان کمتری کاهش می‌یابد. بنابراین می‌توانیم نتیجه بگیریم که نمره واقعی یک فراگیر با نمره چهار در این آزمون با دو مواجهه بین ۲/۸ تا ۵/۲ متغیر است. پس نمی‌توان در مورد وضعیت مطلوب یا نامطلوب این فراگیر اظهارنظر کرد. در نتیجه در این مورد باید از چهار مواجهه برای تشخیص فراگیر با عملکرد نامطلوب استفاده کرد. این در حالی است که افرادی که نمرات بالاتر از شش کسب نموده‌اند، به احتمال قوی فقط به دو مواجهه نیاز دارند زیرا نمرات واقعی آن‌ها بین ۴/۸ و ۷/۲ قرار دارد که به طور مشخص در طیف نامطلوب قرار نمی‌گیرد.

در واقع، با افزایش مواجهه، فاصله اطمینان کمتر می‌شود و بهتر می‌توان افراد با عملکرد خوب را از افراد با عملکرد ضعیف تشخیص داد. فایده این کار این است که با کاهش تعداد مواجهه‌ها برای داوطلبان قوی در منابع ارزیابی صرفه‌جویی می‌کنیم. حسن دیگر این راهبرد تأثیر آموزشی آن است. به این ترتیب که دانشجویان مرزی با مواجهه‌های بیشتری روبه‌رو می‌شوند که به دنبال آن با بازخورد همراه است و در نتیجه این افراد که نیاز بیشتری به آموزش دارند از اقدامات بیشتری نیز در این ارتباط برخوردار می‌شوند.

اصطلاحی که در ادبیات ارزیابی برای توضیح فرایند فوق به‌کار می‌رود «سنجش متوالی»[2] است که در بخش پنجم کتاب به آن پرداخته شد. به این ترتیب که تمام دانشجویان ابتدا در یک آزمون با تعداد محدودی مواجهه (برای روش‌های ارزیابی مبتنی بر محل کار) یا ایستگاه (برای OSCE[3]) شرکت می‌کنند. دانشجویی که در این آزمون رد می‌شود، در آزمون تکمیلی شرکت می‌کند و مجموع عملکرد او در دو آزمون ملاک تصمیم‌گیری در خصوص وضعیت او خواهد بود؛ اما دانشجویانی که در آزمون اولیه قبول می‌شوند، از امتحان بعدی معاف می‌شوند. به نظر می‌رسد با برگزاری آزمون‌ها به صورت متوالی بتوان از نظر هزینه‌های وارده صرفه‌جویی کرد. اما آنچه اهمیت پیدا می‌کند، این است که نتایج تست اول باید قادر باشد با دقت قابل قبولی عملکرد دانشجویان در آزمون اصلی را پیش‌بینی کند.

1. True score
2. Sequential testing
3. Objective Structured Clinical Examination

منابع

1. Amin Z, Seng CY, Eng KH. Practical guide to medical student assessment. World Scientific Publishing; 2006.
2. Boulet JR, Mckinley DW, Norcini JJ, Whelan GP. Assessing the comparability of standardized patient and physician evaluations of clinical skills. Adv Health Sci Educ. 2002;7:85–97.
3. Brown N, Doshi M. Assessing professional and clinical competence: the way forward. Adv Psychiatr Treat. 2008;14(2):122–30.
4. Cantillon P, Wood D. ABC of Learning and Teaching in Medicine. 2nd ed. West Sussex: John Wiely & Sons; 2010.
5. Carr S. The Foundation Programme assessment tools: An opportunity to enhance feedback to trainees? Postgrad Med J. 2006;82(971):576–9.
6. Dent JA, Harden RM. A Practical guide for medical teachers. Edinburg: Elsevier; 2009.
7. Daelmans HE, Overmeer RM, van der Hem-Stokroos HH. Reliability of the clinical teaching effectiveness instrument. Med Educ. 2005;39(9):904–10.
8. Day SC, Grosso LG, Norcini JJ, Blank LL, Swanson DB, Horne MH. Residents' perceptions of evaluation procedures used by their training program. J Gen Inter Med. 1990;5(5):421–6.
9. Driessen E, Scheele F. What is wrong with assessment in postgraduate training? Lessons from clinical practice and educational research. Med Teach. 2013;35(7):569–74.
10. Fitch C, Malik A, Lelliott P, Bhugra D, Andiappan M. Assessing psychiatric competencies: what does the literature tell us about methods of workplace-based assessment? Adv Psychiatr Treat. 2008;14(2):122–30.
11. Govaerts MJB, Schuwirth LWT, Van der Vleuten CPM, Muijtjens AMM. Workplace-based assessment: Effects of rater expertise. Adv Health Sci Educ. 2011;16(2):151–65.
12. Hauer KE, Holmboe ES, Kogan JR. Twelve tips for implementing tools for direct observation of medical trainees' clinical skills during patient encounters. Med Teach. 2011;33(1):27–33.
13. Holmboe ES, Fiebach NF, Galaty L, Huot S: The effectiveness of a focused educational intervention on resident evaluations from faculty: A randomized controlled trial. J Gen Intern Med. 2001;16(7):427–34.
14. Holmboe ES. Faculty and the Observation of Trainees' Clinical Skills: Problems and Opportunities. Acad Med. 2004;79(1):16–22.
15. Holmboe ES, Hawkins RE. Practical guide to the evaluation of clinical competence. Philadelphia: Mosby/Elsevier; 2008.
16. Holmboe ES, Ward DS, Reznick RK, Katsufrakis PJ, Leslie KM, Patel VL, *et al*. Faculty Development in Assessment: The Missing Link in Competency-Based Medical Education. Acad Med.

2011;86(4):460–7.

17. Kassebaum DG, Eaglen RH. Shortcoming in the evaluation of studnets' clinical skills and behaviours in medical school. Acad Med. 1999;74(7):841–9.
18. Kogan JR, Conforti L, Bernabeo E, Iobst W, Holmboe E. Opening the black box of clinical skills assessment via observation: A conceptual model. Med Educ. 2011;45(10):1048–60.
19. Kogan JR, Hauer KE. Brief report: use of the mini-clinical evaluation exercise in Internal Medicine core clerkships. J Gen Inter Med.2006;21(5):501–2.
20. Kogan JR, Holmboe ES, Hauer KE. Tools for direct observation and assessment of clinical skills of medical trainees. JAMA. 2009; 302(12):1316–26.
21. Miller GE. The assessment of clinical skills/competence/performance. Acad Med. 1990;65(9):S63-7.
22. Noel GL, Herbers JE, Caplow MP, Cooper GS, Pangaro LN, Harvey J. How well do Internal Medicine faculty members evaluate the clinical skills of residents? J Gen Inter Med. 1992;117(9):757–65.
23. Norcini J, Burch V. Workplace-based assessment as an educational tool: AMEE Guide No. 31. Med Teach. 2007;29(9-10):855–71.
24. Overeem K, Faber MJ, Arah OA, Elwyn G, Lombarts KM, Wollersheim, HC, et al. Doctor performance assessment in daily practise: does it help doctors or not? A systematic review. Med Educ. 2007;41(11):1039-49.
25. Pelgrim EAM, Kramer AWM, Mokkink HGA, van den Elsen L, Grol RPTM, van der Vleuten CPM. In-training assessment using direct observation of single-patient encounters: A literature review. Adv Health Sci Educ. 2011;16:189–99.
26. Swanwick T. Uderstanding Medical Education: Evidence, Theory and Practice. West Sussex: John Wiely & Sons; 2010.
27. Schuwirth L. Making the horse drink: use of mini-CEX in an assessment for learning view. Adv in Health Sci Educ. 2013;18:1–4.
28. Swanwick T, Chana N. Workplace-based assessment. BJHM. 2009;70(5):290–3.
29. van der Hem-Stokroos NH, Daelmans HE, van der Vleuten CP, Haarman HJ, Scherpbier AL. The impact of multi-faceted educational structuring on learning effectiveness in a surgical clerkship. Med Educ. 2004;38(8):879–86.
30. Veloski J, Boex JR, Grasberger J, Evans A, Wolfson DB. Systematic review of the literature on assessment, feedback, and physicians' clinical performance: BEME Guide No 7. Med Teach. 2006;28(2):117–28.
31. Wilkinson JR, Crossley JGM, Wragg A, Mills P, Cowan G, Wade W. Implementing workplace-based assessment across the medical specialties in the United Kingdom. Med Educ. 2008;42(4):364–73.
32. Wray NP, Friedland JA. Detection and correction of house staff error in physical diagnosis. JAMA. 1983;249(8):1035–7.

33. Yeates P, O'Neill P, Mann K, Eva K. Seeing the same thing differently. Adv Health Sci Educ. 2013;1–17.

۳۴. سیف ع ا. اندازه‌گیری، سنجش و ارزشیابی آموزشی. ویرایش ششم. تهران: نشر دوران، ۱۳۹۰.

فصل ۲۰

آزمون «مورد بالینی کامل»

ساختار آزمون «مورد بالینی کامل» سنتی

همان‌گونه که در فصل یک گفته شد، «مورد بالینی کامل» از این جهت که سطح «نشان می‌دهد چگونه» را در هرم میلر ارزیابی می‌کند، با دیگر ابزارهای مبتنی بر محل کار متفاوت است اما چون عملکرد فراگیران را در برخورد با بیمار واقعی مورد ارزیابی قرار می‌دهد، به عنوان یک روش ارزیابی در محل کار شناخته می‌شود.

به طور معمول، در «مورد بالینی کامل» سنتی به داوطلب ۳۰ تا ۴۵ دقیقه (طبق برخی از منابع ۳۰ تا ۶۰ دقیقه) فرصت داده می‌شود تا بدون اینکه تحت مشاهده و مداخله قرار گیرد، بیمار را در بخش یا درمانگاه مورد مصاحبه، مشاهده و ارزیابی قرار دهد. سپس آزمون‌شونده یافته‌های خود را به یک یا تعداد بیشتری ارزیاب ارائه می‌دهد و در مدت ۲۰ تا ۳۰ دقیقه به سؤالات شفاهی بدون ساختار آنها پاسخ می‌دهد. گاهی لازم می‌شود آزمون‌شونده و ارزیاب برای نشان دادن نشانه‌های بالینی دوباره به بالین بیمار مراجعه کنند. بیماری که در این آزمون استفاده می‌شود به این منظور آموزش ندیده است. در نوع سنتی این ابزار، سیستم نمره‌دهی کاملاً بدون ساختار است و از چک‌لیست یا مقیاس درجه‌بندی که دارای شاخص‌های مرتبط با مهارت داوطلب باشد، استفاده نمی‌شود.

البته با توجه به مزایا و محدودیت‌های این روش که در قسمت بعدی به آن خواهیم پرداخت، انواع اصلاح‌شده و تغییریافته‌ای از «مورد بالینی کامل» پیشنهاد شده است.

مزایا و محدودیت‌های آزمون «مورد بالینی کامل» سنتی

مزایای آزمون «مورد بالینی کامل» سنتی

- **ارزیابی عملکرد فراگیر در مواجهه با بیمار واقعی و در محیط واقعی:** نقطه قوت اصلی مورد بالینی کامل این است که ارزیابی عملکرد داوطلب در تعامل با بیمار واقعی[1] سنجیده می‌شود. مواجهه با بیمار واقعی فرصت برخورد با طیف وسیعی از شرایطی را که در محیط بالینی دیده می‌شود، فراهم می‌کند. این در حالی است که بیمار استاندارد شده یا شبیه‌سازی شده که بخشی از OSCE است در نشان دادن تنوع و پیچیدگی مشکلات واقعی پزشکی محدودیت دارد.
- **ارزیابی جامع و یکپارچه از تعامل فراگیر و بیمار:** در این آزمون داوطلب با یک مسأله بالینی واقعی و کامل روبرو می‌شود. درنتیجه باید همه اطلاعات مورد نیاز را از طریق شرح‌حال‌گیری و معاینه فیزیکی از بیمار کسب کند، یافته‌ها را تحلیل کند و چگونگی اداره بیمار را طرح‌ریزی کند.

1. Authentic

- **تأثیر آموزشی مناسب:** با توجه به اینکه این آزمون در محیط واقعی انجام می‌شود، بازخورد تشخیصی برای دانشجو و مدرس فراهم می‌کند. «مورد بالینی کامل» به عنوان یک روش ارزیابی تکوینی باعث می‌شود نقاط ضعف دانشجویان، بخش‌های فراموش شده آموزش یا نقص‌های تدریس بر اساس پیامدهای دوره مشخص شود. البته در نوع سنتی این آزمون که ارزیابی عملکرد فراگیر بدون مشاهده و به صورت ذهنی و بدون معیارهای از پیش‌تعریف‌شده انجام می‌شود، امکان ارائه بازخورد دقیق وجود ندارد و در نتیجه تأثیر آموزشی آن مورد تردید است.

محدودیت‌های آزمون «مورد بالینی کامل» سنتی

- **عدم تعمیم‌پذیری نتایج و پایایی پایین:** عدم تعمیم‌پذیری نتایج به این معنی است که نتایج یک مورد بالینی کامل به دلیل ویژگی مورد، معیار قابل تعمیمی از توانایی داوطلب برای طیفی از دیگر موارد و موقعیت‌های بالینی نمی‌باشد. در طول سه دهه اخیر چندین مطالعه بر پایایی پایین آزمون مورد بالینی کامل تأکید کرده‌اند. سه عامل به عنوان دلایل عدم تعمیم پذیری یا تکرار پذیری نتایج و پایایی کم مورد بالینی کامل ذکر می‌شود. این دلایل به ترتیب کاهش اهمیت عبارتند از: ویژگی مورد، تفاوت ارزیابان و تغییرپذیری جنبه‌های مواجهه مورد ارزیابی (نورسینی و ون‌درولوتن ۲۰۰۴).

ون‌درولوتن ۱۹۹۶

ما به طور شهودی گمان میکنیم وقتی توانمندی فردی را در برخورد با یک بیمار اندازه گرفتیم، میتوانیم توانمندی او را در برخورد با دیگر بیماران پیش‌بینی کنیم. متأسفانه این پیش‌بینی ضعیف است و در نتیجه، مورد بالینی کامل پایایی ضعیفی دارد.

- **روایی پایین:** مهمترین علت روایی پایین این آزمون تعداد کم موارد بیماری مورد استفاده در آزمون، عدم مشاهده عملکرد داوطلب، عدم وجود ساختار در ارزیابی و عدم استانداردسازی بیمار است.

داگدال [1] ۱۹۹۶

دانستن این مسأله که پزشکی در آزمون مورد بالینی کامل با بیمار مولتیپل اسکلروزیس عملکرد خیلی خوبی داشته است به من اطمینان خاطر نمی‌دهد که می‌تواند فرضاً کارسینوم پروستات مرا نیز تشخیص دهد.

1. Dogdull

- **عدم مشاهده عملکرد داوطلب در طول مواجهه:** یکی از معایب «مورد بالینی کامل» سنتی عدم مشاهده عملکرد داوطلب در مواجهه با بیمار است. در نتیجه نمی‌تواند به طور مستقیم کیفیت مهارت‌های شرح‌حال، معاینه فیزیکی، تعامل پزشک با بیمار و مدیریت زمان را ارزیابی کند. مطالعات نشان می‌دهد ارزیابی مهارت‌های ارتباطی، شرح‌حال‌گیری و معاینه فیزیکی بدون مشاهده مستقیم امکان‌پذیر نیست. این امر روایی این آزمون را زیر سؤال می‌برد.
- **عدم وجود ساختار در ارزیابی و نمره‌دهی:** در نوع سنتی این آزمون عملکرد داوطلب بدون معیار و شاخص‌های از پیش‌تعیین‌شده و بر اساس ملاک‌های ذهنی ارزیابی می‌شود.
- **عدم استانداردسازی بیمار:** استفاده از بیماران با سطوح مختلف پیچیدگی و دشواری برای فراگیران مختلف یکی دیگر از معایب «مورد بالینی کامل» است.
- **قابلیت اجرای نامناسب:** به دلیل زمان‌بر بودن «مورد بالینی کامل» قابلیت اجرای آن با مشکل مواجه است.

در مجموع می‌توان مزایا و محدودیت‌های «مورد بالینی کامل» را به صورت جدول ۱-۲۰ خلاصه کرد.

جدول ۱-۲۰: مزایا و محدودیت‌های «مورد بالینی کامل» سنتی

مزایا	معایب
ارزیابی عملکرد فراگیر در مواجهه با بیمار واقعی در محیط واقعی	پایایی پایین و عدم تعمیم‌پذیری نتایج
ارزیابی جامع از تعامل فراگیر و بیمار	عدم استانداردسازی بیمار
تأثیر آموزشی مناسب در صورت بازخورد	روایی پایین
	عدم وجود ساختار در ارزیابی و نمره‌دهی
	قابلیت اجرای پایین

انواع آزمون‌های مرتبط با «مورد بالینی کامل»

با توجه به آنچه پیشتر ذکر شد بدون انجام اصلاحات در آزمون «مورد بالینی کامل» سنتی، این آزمون نباید به تنهایی مبنای تصمیم‌گیری‌های مهم در مورد عملکرد فراگیران قرار گیرد. برای بهبود کیفیت این روش ارزیابی لازم است موارد زیر را مدنظر قرار داد:

- **موارد بیماری یا مواجهه:** افزایش تعداد مواجهه فراگیر با بیمار، مهمترین اقدام در جهت ارتقای آزمون مورد بالینی کامل است. افزایش هر مواجهه بدون تغییر تعداد ارزیابان و مهارت‌های مورد سنجش، افزایش قابل‌توجهی در تکرارپذیری نمرات ایجاد می‌کند. در حالی که تغییر در تعداد ارزیابان و مهارت‌های مورد سنجش، بدون تعداد کافی مورد بیماری تأثیر کمی بر پایایی نمرات آزمون دارد. در مواردی که امکان افزایش موارد بالینی وجود ندارد روش‌های جایگزین می‌تواند در نظر گرفتن مورد بالینی کوتاه، افزایش ایستگاه بیمار واقعی به آزمون OSCE یا استفاده از روش‌های دیگر برای افزایش تعداد مواجهه دانشجو و بیمار باشد. البته استفاده از بیمار واقعی به همراه OSCE و تأثیر آن بر بهبود تکرارپذیری نتایج نیاز به تحقیقات بیشتر دارد.
- **ارزیابان:** با سه روش می‌توان عوامل مربوط به ارزیابان را به منظور بهبود تکرارپذیری نتایج آزمون مورد بالینی کامل دستکاری کرد. استفاده از هر سه روش با هم افزایش قابل‌قبولی در تکرارپذیری نتایج ایجاد می‌کند. از طرف دیگر افزایش تعداد ارزیابان بیشتر از حد مشخصی (احتمالاً چهار یا پنج نفر) تأثیر بسیار کمی بر تکرارپذیری نتایج دارد:
  - افزایش تعداد ارزیابان
  - آموزش ارزیابان
  - به کار بردن روش‌های آماری به منظور تعدیل تفاوت بین ارزیابان
- **مهارت‌های مورد ارزیابی:** راهبردهای مربوط به این قسمت نیز اثر متوسطی بر بهبود تکرارپذیری نتایج دارد. مثلاً افزایش تعداد مهارت‌های مورد سنجش بیش از یک حد معین (پنج تا ده مورد) اثر بسیار کمی بر بهبود تکرار پذیری نتایج دارد. در این ارتباط موارد زیر را باید رعایت نمود:
  - افزایش تعداد مهارت‌های مورد سنجش
  - مشاهده عملکرد فراگیر در زمان مواجهه با بیمار
  - تهیه فهرستی از مهارت‌های مورد سنجش قبل از آزمون توسط ارزیابان

در این قسمت با توجه به مواردی که برای بهبود روایی و پایایی این روش ارزیابی بیان شد به نمونه‌هایی عملی که بخشی از عوامل اصلاح مورد بالینی کامل در آن‌ها به کار رفته است اشاره می‌کنیم:

مورد بالینی کامل ساختارمند مشاهده‌شده

یکی از ابزارهای مشخص در این زمینه OSLER[1] است. اولین بار گلیسون[2] این روش ارزیابی را در زمانی که مورد بالینی کامل سنتی به دلیل ارزیابی تصادفی فراگیر مورد انتقاد بود، به کار برد. در این روش از سه راه‌کار زیر برای رفع مشکلات مورد بالینی کامل سنتی استفاده می‌شود:

- مشاهده تعامل داوطلبان با بیمار
- ارزیابی ساختارمند بر اساس فرم ده سؤالی، شامل مهارت‌های مورد سنجش
- استفاده از سیستم نمره‌دهی ساختارمند

مورد بالینی کامل ساختارمند با چند ارزیاب

در این آزمون از دو راه‌کار برای رفع مشکلات مورد بالینی کامل سنتی استفاده می‌شود (السون[3] ۲۰۰۰):

- ارزیابی ساختارمند بر اساس جدول سؤالات
- استفاده از دو ارزیاب برای ارزیابی یک مورد بالینی

مورد بالینی کامل ساختارمند مشاهده‌شده با چند ارزیاب

در این آزمون نیز از دو راه‌کار برای رفع مشکلات مورد بالینی کامل سنتی استفاده می‌شود (وس و جالی[4] ۲۰۰۱):

- مشاهده تعامل داوطلبان با بیمار
- استفاده از سیستم نمره‌دهی ساختارمند

نورسینی معتقد است انجام این اصلاحات مانند آنچه در مطالعه وس و جالی برای بهبود پایایی آزمون به کار رفته است، کماکان نمی‌تواند مورد بالینی کامل را به عنوان یک آزمون با کاربرد تراکمی حمایت کند.

ابزارهای مشخص در این زمینه DOCEE[5] و IDOCEE[6] است. در این ابزارها از راه‌کارهای زیر برای رفع مشکلات مورد بالینی کامل سنتی استفاده شده است:

- استفاده از چندین ارزیاب با تخصص‌های مختلف
- استفاده از چندین بیمار (چهار تا شش)
- استفاده از سیستم نمره‌دهی ساختارمند

در DOCEE هر داوطلب با چهار مورد بالینی و دو جفت از ارزیابان ارزیابی می‌شود. هر دو نفر از ارزیابان عملکرد داوطلب را در برخورد با دو بیمار بررسی می‌کنند. این روش ارزیابی اولین بار در کشور بحرین به عنوان ارزیابی تراکمی دوره پزشکی عمومی به جای مورد بالینی کامل سنتی به کار گرفته شد که ضریب پایایی بر اساس تئوری تعمیم‌پذیری ۰/۸۴ به دست آمد. (حمدی و همکاران[7] ۲۰۰۳).

روش دیگر یعنی IDOCEE نیز از لحاظ ساختار و اجرا بسیار شبیه DOCEE است. تنها تفاوت آن در تعداد بیماران (چهار یا شش) و تعداد ارزیابانی (دو پانل هر کدام شامل دو یا سه ارزیاب) است که هر داوطلب با آن رو به رو می‌شود. هر پانل از ارزیابان عملکرد داوطلب را در مواجهه با دو یا سه مورد بالینی ارزیابی می‌کردند. داوطلبان و ارزیابان از ساختار، سازمان‌دهی و کارآیی این آزمون رضایت داشتند (حمدی و همکاران ۲۰۰۳).

1. Objective Structured Long Examination Record (OSLER)
2. Gleeson
3. Olson
4. Wass & Jolly
5. Direct Observation Clinical Encounter Examination (DOCEE)
6. Integrated Direct Observation Clinical Encounter Examination (IDOCEE)
7. Hamdy et al.

مورد بالینی کوتاه

در «مورد بالینی کوتاه»، داوطلب با تعدادی بیمار دارای شرایط متنوع روبه‌رو می‌شود. معمولاً ارزیابی در این روش شامل استفاده از سه تا چهار (گاهی تا شش) بیمار واقعی غیراستاندارد است. تعامل داوطلب با بیمار واقعی مورد قضاوت قرار می‌گیرد و حیطه‌های مورد ارزیابی در این آزمون عبارتند از:

- معاینه بالینی یک سیستم یا بخشی از بدن
- ارائه تشخیص افتراقی یا نشان دادن نشانه‌های بالینی روی بیمار
- گاهی، تعهد حرفه‌ای و مهارت‌های ارتباطی به جای معاینه بالینی

در «مورد بالینی کوتاه» داوطلب مدت زمان پنج دقیقه را صرف معاینه بخشی از بدن یا یک سیستم به عنوان مثال سیستم قلبی و عروقی می‌کند و سپس حدود سه دقیقه برای سازمان‌دهی اطلاعات و رسیدن به تشخیص افتراقی اختصاص می‌یابد. عملکرد داوطلب توسط یک یا دو ارزیاب مورد مشاهده و قضاوت قرار می‌گیرد. در این روش ارزیابی معمولاً از تدابیر نمره‌دهی یکسانی برای تمامی موارد بالینی استفاده می‌شود.

«مورد بالینی کوتاه» از این نظر که در آن عملکرد داوطلب بر اساس مواجهه با چندین بیمار ارزیابی می‌شود، به آزمون OSCE شباهت دارد. اما این دو روش ارزیابی تفاوت‌های مهمی نیز دارند. از جمله این تفاوت‌ها می‌توان به موارد زیر اشاره نمود:

- عدم مواجهه فراگیران مختلف با بیماران مشابه
- تفاوت سطح دشواری بیماران
- واقعی بودن بیماران
- عدم ارزیابی مهارت‌های ارتباطی
- ساختارمند نبودن آزمون و نمره‌دهی
- عدم تعیین سطح مورد انتظار استاندارد عملکرد

مزایای آزمون مورد بالینی کوتاه عبارت است از:

- فرصت ارزیابی عملکرد فراگیران در مواجهه با بیماران واقعی را فراهم می‌کند.
- امکان ارائه نمونه وسیعتری از موارد بالینی را در مقایسه با ارائه یک مورد فراهم میکند.
- امکان ارزیابی مهارت‌های بالینی فراگیران را با جزئیات بیشتری فراهم میکند.
- از روایی سازه خوبی برخوردار است.

محدودیت‌های آزمون مورد بالینی کوتاه عبارت است از:

- میزان پایایی درون فردی برای یک آزمون مشابه متغیر است.
- موارد بالینی ارائه شده به صورت سنتی، در مقایسه با اشکال نوین ارزیابی از قبیل OSCE کمتر استاندارد شده هستند.

توصیه می‌شود در آزمون‌هایی از قبیل ارزیابی عملی مهارت‌های بالینی و OSCE، از چندین «مورد بالینی کوتاه» استانداردشده استفاده شود. انجام این کار منجر به افزایش پایایی و استانداردسازی بیشتر آزمون میشود (زبیر امین ۲۰۰۶). مواردی برای این آزمون انتخاب شوند که دربرگیرنده چندین مهارت و دامنه وسیعی از مشکلات بالینی موجود باشند. انجام این کار منجر به افزایش روایی و تعمیم‌پذیری بیشتر نتایج آزمون میشود.

کاربرد آزمون «مورد بالینی کامل»

آزمون «مورد بالینی کامل» مانند دیگر روش‌های ارزیابی مبتنی بر محل کار هم با هدف تکوینی و هم با هدف تراکمی مورد استفاده قرار می‌گیرد. البته با توجه به انتقاداتی که به پایایی و تعمیم‌پذیری نتایج این آزمون وارد است، استفاده از آن به عنوان

یک ارزیابی تراکمی چالش برانگیز است و معمولاً با هدف تکوینی در سیستم‌های آموزش دستیاری و در مواردی در دوره آموزش پزشکی عمومی به کار می‌رود. به عنوان مثال، در دوره کارآموزی طب داخلی در دانشگاه ریج[1] هلند از آزمون «مورد بالینی کامل» به عنوان ابزاری برای بازخورد در مورد عملکرد دانشجویان در کنار چند روش ارزیابی دیگر استفاده می‌شود.

نورسینی (۲۰۰۱) پیشنهاد می‌کند که بهتر است آزمون «مورد بالینی کامل» سنتی به عنوان یک ابزار غربالگری[2] با هدف تعیین دانشجویان با عملکرد مرزی یا نامناسب استفاده شود و سپس عملکرد این دانشجویان در آزمون‌های دقیق‌تری مانند OSCE ارزیابی شود.

در صورتی‌که آزمون مورد بالینی کامل با هدف تراکمی استفاده شود، نباید به تنهایی مبنای تصمیم‌گیری در مورد عملکرد فراگیران مخصوصاً در آزمون‌های سرنوشت‌ساز، قرار گیرد. دو پیشنهاد در این رابطه مطرح شده است:

- استفاده از چند روش ارزیابی مبتنی بر محل کار دیگر همراه «مورد بالینی کامل»: این کار منجر به تلفیق مزایای روش‌های ارزیابی مختلف و کاهش معایب آن‌ها می‌شود. نمونه‌ای از آن، آزمون معاینه بالینی است که کالج سلطنتی پزشکان استرالیا[3] هر سال یک بار جهت ارزیابی دستیاران طب اطفال و داخلی در استرالیا و نیوزلند برگزار می‌کند و عملکرد هر داوطلب در دو «مورد بالینی کامل» و چهار «مورد بالینی کوتاه» ارزیابی می‌شود.
- همراه ساختن آن با سایر روش‌های ارزیابی از سطوح مختلف هرم میلر: به عنوان مثال در کشور بحرین برای آزمون تراکمی دوره پزشکی عمومی از روش‌های ارزیابی مختلف استفاده می‌شود: آزمون نوشتاری (شامل سؤالات چندگزینه‌ای، کوته‌پاسخ و PMP[4])، OSCE و آزمون «مورد بالینی کامل تغییریافته» تحت عنوان DOCEE.

نمونه استفاده شده در استرالیا

یکی از انواع آزمون‌هایی که کالج سلطنتی پزشکان استرالیا (RACP) برگزار می‌کند، آزمون معاینه بالینی است که به همین نام RACP معروف است. در این آزمون که هر سال یک بار برای ارزیابی دستیاران طب اطفال و داخلی در استرالیا و نیوزلند برگزار می‌شود، عملکرد هر داوطلب در مقابل دو «مورد بالینی کامل» و چهار «مورد بالینی کوتاه» ارزیابی می‌شود. به منظور غلبه بر تفاوت بین ارزیابان دو راهبرد زیر را به کار بسته شده است:

- طراحی سیستم نمره‌دهی ساختارمند
- طراحی سیستم آموزش ارزیابان و هماهنگ نمودن آنها

مورد بالینی کامل: همه بیماران قبل از شروع ارزیابی توسط ارزیابان بدون مراجعه به پرونده‌های بیمار بررسی می‌شوند تا ارزیابان انتظارات واقعی از شرح‌حال و معاینه به دست آورند. دو ارزیاب ۴۰ دقیقه را صرف بررسی بیمار می‌کنند و سپس به مدت ۲۰ دقیقه در مورد یافته‌ها و مسائل اصلی بیمار که باید در ارزیابی به آن توجه شود، توافق می‌کنند. سپس داوطلب به بررسی بیمار می‌پردازد و به مدت ۶۰ دقیقه بیمار را در پنج حیطه بررسی می‌کند: اخذ شرح‌حال، معاینه بیمار، تشخیص بیمار، طراحی مدیریت بیماری و تعیین اثر بیماری بر بیمار و خانواده‌اش. در ادامه، دو ارزیاب دو داوطلب را با صرف زمان ۲۵ دقیقه برای هر کدام ارزیابی می‌کنند. با در نظر گرفتن ۱۰ دقیقه زمان استراحت، مدت زمانی که هر ارزیاب برای هر داوطلب صرف می‌کند، ۶۰ دقیقه است.

«مورد بالینی کوتاه»: این آزمون، شامل مشاهده تعامل بیمار و داوطلب می‌شود که در آن پنج حیطه مورد ارزیابی قرار می‌گیرد: تعامل بین داوطلب و بیمار، توانایی معاینه مفصل یک سیستم خاص، دقت در تعیین نشانه‌ها، فرمول‌بندی تشخیص افتراقی و تفسیر مناسب. قبل از آزمون دو ارزیاب، بیمار را به مدت ۲۰ دقیقه برای تعیین پیچیدگی علائم و استانداردهای تعیین سطح عملکرد مطلوب بررسی می‌کنند. سپس چهار داوطلب به صورت جداگانه و هر یک به مدت ۱۵ دقیقه هر بیمار را معاینه می‌کند. کل زمان صرف شده ۸۰ دقیقه (۲۰ دقیقه برای هر داوطلب) است. در نتیجه زمانی که توسط ارزیاب برای سه «مورد بالینی کوتاه» صرف می‌شود، برابر یک «مورد بالینی کامل» است. در آزمون معاینه بالینی RACP عملکرد داوطلب در پنج حیطه و با مقیاس لیکرت هفت‌تایی مورد ارزیابی قرار می‌گیرد. علاوه بر این پنج حیطه، عملکرد کلی داوطلب در تعامل با بیمار نیز مورد ارزیابی قرار می‌گیرد. نمره این آیتم به عنوان نمره داوطلب در آن مواجهه در نظر گرفته می‌شود. در این آزمون عملکرد رضایت‌بخش نمره چهار محسوب می‌شود.

عملکرد هر داوطلب توسط دو ارزیاب مورد سنجش قرار می‌گیرد. هر کدام از ارزیابان به تنهایی نمره‌ای به عملکرد داوطلب اختصاص می‌دهند و سپس قبل از تصمیم‌گیری نهایی در مورد نمره نهایی در مورد نقاط ضعف و قوت عملکرد داوطلب توافق می‌کنند. سپس نمرات مربوط به شش مواجهه با هم جمع می‌شوند اما نمرات «مورد بالینی کامل» ضریب ۳ دارند. در مجموع، با جمع نمرات دو مورد بالینی کامل (۶=۳×۲) و سه مورد بالینی کوتاه، حداکثر نمره داوطلب ۷۰ (۱۰×۷) میشود و حدنصاب قبولی نمره ۴۰ است.

ساختار و ترتیب بیماران برای هر مرکز آزمون استاندارد شده است. هر بیمار در «مورد بالینی کامل» توسط دو داوطلب و هر بیمار در «مورد بالینی کوتاه» توسط چهار داوطلب ارزیابی می‌شود. هر دو نفر از ارزیابان، چهار داوطلب را در مواجهه با یک «مورد بالینی کوتاه» و دو داوطلب را در مواجهه با یک «مورد بالینی کامل» ارزیابی می‌کنند. این ساختار بستر مناسبی برای تحلیل ضریب تعمیم‌پذیری فراهم می‌کند زیرا هر ارزیاب چندین داوطلب را در مواجهه با چندین مورد بررسی قرار می‌دهد و تعداد بالای بیمارستان‌ها (۴۰ تا ۵۰ بیمارستان برگزارکننده آزمون در سال)، بیماران، داوطلبان و ارزیابان نیز باعث غنای تحقیق در این زمینه می‌شود.

1. Vrij
2. Screening
3. Royal Australasian College of Physicians (RACP)
4. Patient Management Problem (PMP)

به منظور یکسان‌سازی ارزیابان، هر داوطلب توسط دو ارزیاب مورد قضاوت قرار می‌گیرد که یکی از آنها فردی واجد شرایط و عضو پانل ملی امتحانات است. به منظور تأمین ثبات در قضاوت ارزیابان، این افراد در مراکز مختلف آزمون در گردش هستند و در طول شش سال هر سال در یک مرکز حضور دارند. ارزیاب دیگر از بیمارستان محل آزمون انتخاب می‌شود و قبلاً حداقل یک بار یک آزمون کامل را مورد مشاهده قرار داده است. همه ارزیابان به طور فعال تحت آموزش قرار می‌گیرند و باید در جلسه توجیهی و یکسان سازی قبل از آزمون شرکت کنند. در این جلسه به افراد فیلمی ضبط شده از یک آزمون ساختگی نشان داده می‌شود که در آن از آزمون‌دهنده‌های داوطلب و ارزیابان پانل ملی استفاده شده است. ارزیابان حاضر در جلسه پس از مشاهده فیلم، عملکرد داوطلب در تعامل با بیمار را با استفاده از فرم‌هایی که توضیح داده شد نمره‌دهی می‌کنند. سپس نمرات با هم مقایسه شده و در مورد آن با توجه مخصوص به نمرات دو سر طیف بحث می‌شود. سپس نمرات ارزیابان پانل ملی ارائه می‌شود و دلیل اعطای این نمرات بحث و بررسی می‌شود

نمونه استفاده شده در بحرین

در بحرین از فرم اصلاح شده‌ای از «مورد بالینی کامل» به نام DOCEE در امتحان تراکمی دوره پزشکی عمومی استفاده می‌شود. هر دانشجوی پزشکی با ۴ مورد بالینی کامل روبه رو می‌شود. موارد بیماری شامل یک کودک، یک خانم باردار با مشکل بالینی (مانند دیابت یا افزایش فشارخون)، دو بیمار بالغ (یکی با مشکل داخلی و دیگری با مشکل جراحی) است. بیماران بر اساس بلوپرینت از قبل تهیه شده بر اساس مشکلات شایع سلامتی انتخاب می‌شوند. هر دانشجو توسط دو پانل از ارزیابان مورد ارزیابی قرار می‌گیرد که هر پانل عملکرد دانشجو را در مقابل دو بیمار ارزیابی می‌کنند. هر پانل از دو ارزیاب از تخصص‌های مختلف (اطفال، زنان، داخلی و جراحی) تشکیل شده است. داوطلبان و ارزیابان در مورد روش ارزیابی، مسائل اجرائی و پشتیبانی و سیستم نمره‌دهی و فرم‌ها توجیه شده‌اند. مواجهه دانشجو و بیمار مورد مشاهده قرار می‌گیرد . هر سه دانشجو با بیماران و ارزیابان مشابه برخورد دارند. هر مورد بالینی کامل ۴۵ دقیقه طول می کشد و ۱۵ دقیقه نیز به بحث بین ارزیاب و داوطلب اختصاص دارد. ارزیابی که تخصص مربوط به مورد بیماری (متخصص) را دارد، بحث را هدایت می‌کند و ارزیابان دیگر (غیرمتخصص) نیز از دانشجو سؤال می‌کنند. سپس هر ارزیاب به صورت مستقل به داوطلب نمره می‌دهد. نمره‌دهی بر اساس یک چک‌لیست ساختارمند با چهار حیطه مهارتی مشتمل بر ۱۱ سؤال است. حیطه‌های مورد ارزیابی در این آزمون عبارت هستند از:

- مهارت‌های جمع‌آوری اطلاعات (مهارت‌های ارتباطی، اخذ شرح‌حال، معاینه فیزیکی، سازمان‌دهی اطلاعات و ارائه)
- مهارت‌های تحلیل و استدلال (تعیین مشکل، تولید فرضیه، طرح تشخیص)
- مهارت‌های تصمیم‌گیری (تفسیر تست‌های آزمایشگاهی و رادیولوژیکی، طرح مدیریت بیماری و تصمیمات درمانی و پیشگیرانه)
- نگرش حرفه‌ای (احترام به بیمار، تواناییهای ارتباطی با بیمار و ارزیاب)

هر سؤال به عنوان رضایت‌بخش و عدم رضایت، علامت‌گذاری می‌شود. نمره نهایی ۰۰۱ و نمره قبولی ۰۶ در نظر گرفته شده است. نمره نهایی داوطلب میانگین نمرات ارائه شده از هشت ارزیاب در مواجهه با چهار مورد بیماری است. زمان کل آزمون سه ساعت است. (حمدی و همکاران ۳۰۰۲).

نمونه استفاده شده در انگلیس

با معرفی OSCE برخی از دانشکده‌های انگلیس، آزمون «مورد کامل بالینی» را از آزمون‌های نهایی دوره پزشکی عمومی حذف کردند اما بورد امتحانات دانشکده پزشکی GKT تصمیم گرفت یک مورد بالینی در امتحانات تراکمی دانشکده داشته باشد. به این ترتیب در آزمون پایان دوره پزشکی عمومی سه روزه این دانشکده، داوطلبان در آزمون‌های نوشتاری و بالینی شرکت می‌کنند. آزمون نوشتاری شامل سؤالات چندگزینه‌ای، کوته‌پاسخ و سؤالات تشریحی است. آزمون بالینی شامل OSCE دارای ۲۰ ایستگاه و یک مورد بالینی مصاحبه با بیمار است. در «مورد بالینی کامل» این دانشکده دو تغییر وجود دارد:

- حذف معاینه بالینی از آزمون به دلیل ارزیابی در OSCE
- مشاهده فرایند مصاحبه با بیمار

در آزمون «مورد بالینی کامل» برگزار شده در دانشکده پزشکی GKT، توانمندی فراگیران در شرح‌حال گیری از بیمار با چک‌لیست و ارزیابی کلی مورد ارزیابی قرار می‌گیرد. چک‌لیست شامل ۲۰ سؤال است که موارد درج شده در آن در جدول زیر آمده است. ارزیابی کلی با مقیاس لیکرت پنج‌تایی در خصوص سابقه بیماری قبلی، سابقه روانی-اجتماعی، مهارت‌های مصاحبه و تسلط کلی در مهارت صورت می‌گیرد.

موضوع	سؤال	حداکثر نمره
مشخصات بیمار	معرفی/سن/شغل	۱
بیماری فعلی	نشانه ها: نوع، توصیف، مدت، عوامل تسکین دهنده، عوامل تشدید کننده، عوامل مستعد کننده، مصرف دارو، علائم روانی	۷
سابقه بیماری قبلی	بیماری، مصرف دارو، بستری، آلرژی	۳
مرور سیستم‌ها	سیستم ادراری، اشتها، وزن، عملکرد روده، قاعدگی در صورت لزوم	۴
سابقه روانی-اجتماعی	خانواده، همسر، الکل، سیگار، تغذیه	۵
نمره کلی		۲۰

گام‌های طراحی و اجرای آزمون «مورد بالینی کامل» مطلوب

در آزمون «مورد بالینی کامل» سه متغیر وجود دارد: بیمار، ارزیاب و داوطلب. همان‌طور که قبلاً نیز اشاره شد در آزمون‌های مبتنی بر محل کار در حالت ایده‌ال باید تنها متغیر داوطلب باشد و اثر ارزیاب و بیمار حذف شود. رعایت موارد زیر به طراحی یک آزمون مورد بالینی کامل مناسب کمک می‌کند (جدول ۲-۲۰):

جدول ۲-۲۰: خلاصه مراحل طراحی و اجرای «مورد بالینی کامل»

ردیف	عنوان مرحله	توضیح
۱	تهیه بلوپرینت آزمون	ضروری است بلوپرینت آزمون بر اساس پوشش برنامه درسی پایه و با در نظر گرفتن تنوع موارد تهیه شود.
۲	تعیین مهارتهای مورد سنجش	در آزمون «مورد بالینی کامل» می‌توان طیف وسیعی از مهارتهای بالینی شامل شرح‌حال‌گیری، معاینه بالینی، مهارتهای ارتباطی، سنتز یافته‌ها، مهارتهای تشخیص، مدیریت بیماری و آموزش بیمار را مورد ارزیابی قرار داد. قبل از آزمون باید به توافق برسیم که چه طیفی از مهارت‌های بالینی را می‌خواهیم اندازه‌گیری کنیم.
۳	تعیین مهارتهای مورد مشاهده	مشاهده مورد بالینی کامل روایی آن را افزایش می‌دهد. ممکن است صرفاً مهارت‌های خاصی مانند شرح‌حال‌گیری یا مهارتهای ارتباطی مورد مشاهده قرار گیرد.
۴	طراحی فرم‌های ارزیابی	تهیه فرم‌های نمره‌دهی کلی یا چک‌لیست در آزمون مورد بالینی کامل به منظور غلبه بر مشکل ذهنی بودن نوع سنتی آزمون صورت می‌گیرد. فرم‌های ارزیابی نمره‌دهی می‌توانند به صورت چک‌لیست (آزمون DOCEE) یا به صورت نمره‌دهی گلوبال (سیستم RACP و آزمون OSLER) طراحی شده باشند.
۵	تصمیم گیری در مورد تعداد و ویژگی ارزیابان	به منظور افزایش پایایی آزمون در اغلب موارد از دو ارزیاب استفاده می‌شود. این موضوع به ویژه در آزمون‌های حساس و مهم صادق است. تعداد بیشتر از سه نفر تأثیری بر پایایی آزمون ندارد. در صورتی که تعداد ارزیابان محدود است، افزایش تعداد موارد بر افزایش تعداد ارزیابان برای هر داوطلب ارجحیت دارد. معمولاً ارزیابان در مورد بالینی کامل از استادان بخش که دانش و مهارت تخصصی لازم را دارند، انتخاب می‌شوند. در مواردی که از بیش از یک ارزیاب استفاده می‌شود، می‌توان از پزشکان عمومی یا استادان دارای تخصص مرتبط اما از بیمارستان غیر از محل برگزاری آزمون استفاده کرد.
۶	توجیه و آموزش ارزیابان و اطلاع‌رسانی	قبل از آزمون لازم است ارزیابان با فرم‌های ارزیابی و شیوه نمره‌دهی آشنا شوند و به منظور بهبود مهارت‌های مشاهده و قضاوت در دوره‌های آموزشی شرکت کنند. ارزیابان باید توجیه شوند که درجه دشواری مورد بیماری را در زمان قضاوت در مورد عملکرد فراگیر و نمره‌دهی آن در نظر بگیرند. ضروری است سیلابوس دوره، بلوپرینت آزمون، فرم‌های نمره‌دهی و راهنماهای برگزاری آزمون در اختیار داوطلبان و ارزیابان قرار گیرد.
۷	اجرای آزمون	لازم است بلافاصله قبل از شروع آزمون ارزیابان با بیماران ملاقات کنند. این ملاقات به منظور آشنایی با بیمار، تطبیق فرم‌های ارزیابی، تخمین دشواری موارد بیماری و از همه مهم‌تر هماهنگی و ثبات قضاوت ارزیابان (در صورت استفاده از بیش از یک ارزیاب) انجام می‌شود. ضروری است روند شرح‌حال‌گیری و معاینه فیزیکی داوطلب مورد مشاهده قرار گیرد. لازم است بر اساس معیارهای مشخص عملکرد داوطلب ارزیابی و سپس به صورت ساختارمند نمره‌دهی شود.
۸	نظرخواهی از فراگیران و ارزیابان	پس از اجرای آزمون یکی از اقداماتی که می‌توان به منظور اطمینان از کیفیت آزمون انجام داد نظرخواهی از داوطلبان و ارزیابان در مورد کیفیت آزمون برگزار شده و بررسی میزان رضایت آنهاست.

تهیه بلوپرینت آزمون

قبل از آزمون باید موارد بیماری که در «مورد بالینی کامل» مورد استفاده قرار می‌گیرد مشخص شود. موارد بر اساس برنامه درسی پایه و با استفاده از بلوپرینت آزمون انتخاب می‌شود. به عنوان مثال آزمونی که در انتهای بخش جراحی دوره کارورزی برگزار می‌شود، می‌تواند شامل بیماران با مشکلات عروق محیطی، کوله سیستیت، پانکراتیت، سرطان پستان و مشکلات ریوی باشد. در صورتی که مورد بالینی کامل بخشی از یک سیستم ارزیابی است لازم است از تکراری بودن موارد بیماری پرهیز شود. استفاده از بلوپرینت آزمون روایی محتوایی آن را افزایش می‌دهد. از سوی دیگر افزایش تعداد موارد منجر به افزایش روایی و پایایی آزمون می‌شود. بنابراین، لازم است با توجه به منابع و زمان در دسترس در مورد تعداد بیماران تصمیم‌گیری شود.

تعیین مهارت‌های مورد سنجش

در آزمون «مورد بالینی کامل» می‌توان طیف وسیعی از مهارت‌های بالینی شامل شرح‌حال‌گیری، معاینه بالینی، مهارت‌های ارتباطی، سنتز یافته‌ها، مهارت‌های تشخیص، مدیریت بیماری و آموزش بیمار را مورد ارزیابی قرار داد. قبل از آزمون باید به توافق برسیم که چه طیفی از مهارت‌های بالینی را می‌خواهیم اندازه‌گیری کنیم. مهارت‌های مورد ارزیابی نیز باید بر اساس بلوپرینت آزمون انتخاب شوند. در آزمون OSLER چهار حیطه شرح‌حال‌گیری، معاینه بالینی، مهارت‌های تشخیص و مدیریت بیماری مورد ارزیابی قرار می‌گیرد.

تعیین مهارت‌های مورد مشاهده

همان‌طور که پیشتر نیز اشاره شد مشاهده «مورد بالینی کامل» روایی این آزمون را افزایش می‌دهد. با توجه به اینکه مشاهده کل عملکرد داوطلب زمان‌بر و نیازمند منابع است، ممکن است صرفاً مهارت‌های خاصی مانند شرح‌حال‌گیری یا مهارت‌های ارتباطی مورد مشاهده قرار گیرد.

طراحی فرم‌های ارزیابی

تهیه فرم‌های نمره‌دهی کلی یا چک‌لیست در آزمون «مورد بالینی کامل» به منظور غلبه بر مشکل ذهنی بودن آزمون سنتی صورت می‌گیرد. در تدوین فرم‌ها قابلیت اجرای آن و مسأله زمان را باید در نظر داشته باشیم و در نتیجه از تدوین فرم‌های طولانی پرهیز کنیم. هم اکنون فرم‌های متنوعی از این آزمون در سیستم‌های آموزشی استفاده می‌شود که از آن جمله می‌توان به فرم‌های استفاده شده توسط کالج سلطنتی پزشکان استرالیا (RACP)، فرم ده سؤالی مربوط به آزمون OSLER و فرم آزمون DOCEE اشاره کرد.

فرم‌های ارزیابی نمره‌دهی می‌توانند به صورت چک‌لیست (آزمون DOCEE) یا به صورت نمره‌دهی گلوبال (سیستم RACP و آزمون OSLER) طراحی شده باشند. در آزمون RACP، عملکرد داوطلب در پنج حیطه و با مقیاس لیکرت هفت‌تایی مورد ارزیابی قرار می‌گیرد (شکل ۱–۲۰). فرم آزمون OSLER که یکی از اولین تلاش‌ها به منظور ساختارمند کردن مورد بالینی کامل بود، در اینجا به صورت تفصیلی مورد بحث قرار می‌گیرد (شکل ۲–۲۰). در این فرم چهار سؤال مربوط به مهارت شرح‌حال‌گیری، سه سؤال مربوط به معاینه بالینی و سه سؤال باقی‌مانده مربوط به مهارت‌های تشخیص و مدیریت بیماری است:

- سؤال‌های مربوط به شرح‌حال‌گیری شامل سرعت و وضوح ارائه، فرایند برقراری ارتباط، رویکرد سیستماتیک و تعیین مشکلات بیمار است. سرعت ارائه، سرعت تکلم با مکث‌های مناسب را می‌سنجد. صحبت کردن خیلی سریع و غیرقابل‌فهم یا صحبت کردن آهسته و زمان‌بر، نمره مناسب را دریافت نمی‌کند. وضوح ارائه به این معنی است که ارزیاب بتواند داستانی را که در پس شرح‌حال قرار دارد درک کند. در مورد مهارت‌های ارتباطی ارزیاب می‌تواند

روی یک قسمت خاص از بیماری متمرکز شود و فرایند ارتباط داوطلب با بیمار را ارزیابی کند. به عنوان مثال ارزیاب می‌تواند از داوطلب درخواست کند تا در عرض سه دقیقه یک شرح‌حال مربوط به سیستم قلبی عروقی یا شرح‌حال مربوط به سابقه خانوادگی از بیمار بگیرد و سپس ارزیاب بر اساس مشاهده این فرایند و گوش دادن به بقیه شرح‌حال در مورد مهارت‌های ارتباطی داوطلب قضاوت می‌کند. مثال دیگر ارزیابی مهارت‌های ارتباطی در بخش مدیریت بیماری است که ارزیاب از داوطلب می‌خواهد که نحوه مصرف و همچنین عوارض جانبی یک دارو مانند ضد انعقادها را به بیمار آموزش دهد.

- سؤالات مربوط به مهارت‌های معاینه فیزیکی، هم تکنیک معاینه و هم رویکرد سیستماتیک به معاینه را مورد ارزیابی قرار می‌دهند.
- در بخش تشخیص بیماری، ارزیاب توانایی داوطلب در رسیدن به تشخیص مناسب با تناوب منطقی و در زمان محدود را بررسی می‌کند. مدیریت بیماری از این نظر اهمیت دارد که ممکن است داوطلبی در مراحل اولیه خوب عمل کند اما نتواند در نهایت بیمار را به طور مناسب مدیریت کند. شم بالینی توانایی کلی داوطلب در کنار هم‌گذاشتن بخش‌های مختلف بیماری، تعیین مشکل بیمار و حل این مشکلات است.

از آنجا که یکی از اشکالاتی که به نوع سنتی آزمون «مورد بالینی کامل» وارد بود، نمره‌دهی ذهنی آن بود، یکی از اصلاحاتی که در مورد OSLER صورت گرفته است، سیستمی است که برای نمره‌دهی آن پیشنهاد شده است (شکل ۳-۲۰). سیستم نمره‌دهی در این آزمون پنج‌تایی است و بالاترین نمره ۸۰ و پایین‌ترین نمره ۳۵ است. نمره‌دهی عملکرد در سه بخش انجام می‌شود:

- p^+ (خیلی خوب/عالی): ۶۰ تا ۸۰
- P (قبول/مرزی): ۵۰ و ۵۵
- P^- (رد): ۳۵ تا ۴۵

با توجه به این که در این آزمون دو ارزیاب عملکرد داوطلب را نمره‌دهی می‌کنند، دو بخش برای نمره‌دهی در نظر گرفته شده است که اولی مربوط به نمره‌ای است که ارزیاب به صورت مستقل به داوطلب می‌دهد و دومی مربوط به نمره‌ای است که پس از بحث و رسیدن به توافق بین دو ارزیاب لحاظ می‌شود (شکل ۲-۲۰).

استفاده از سیستم نمره‌دهی ساختارمند فرایند نمره‌دهی را عینی می‌کند و از خطاهایی مانند سهل‌گیری یا سخت‌گیری در نمره دادن می‌کاهد. حمدی و همکاران (۲۰۰۳) نمرات آزمون DOCEE را با نمرات آزمون‌های MCQ[1]، SAQ[2]، PMP و OSCE برگزار شده برای همان داوطلبان مقایسه کردند. میانگین، انحراف معیار و دامنه نمرات برای آزمون MCQ به ترتیب ۵۸، ۱۰ و ۴۴ و برای آزمون DOCEE به ترتیب ۷۰، ۹ و ۴۱ بود. پژوهشگران نتیجه گرفتند که ساختارمند کردن سیستم نمره‌دهی در «مورد بالینی کامل» موجب کاهش خطاهای مربوط به نمره‌دهی این آزمون‌ها می‌شود.

از موارد دیگری که لازم است در زمان اجرای آزمون «مورد بالینی کامل» مد نظر قرار داد، دشواری موارد بیماری است. در آزمون OSLER که توسط گلیسون برگزار شد با تعیین دشواری موارد توسط ارزیابان تلاش شد بر مشکل ویژگی مورد غلبه شود تا پایایی و روایی آزمون تضمین شود. در این آزمون، موارد بالینی از نظر دشواری به سه دسته تقسیم شدند:

- موارد بالینی استاندارد که شامل یک مشکل بالینی هستند.
- موارد بالینی دشوار که حداکثر سه مشکل بالینی دارند.
- موارد بالینی خیلی دشوار که بیش از سه مشکل بالینی دارند.

البته گاهی بیماری که صرفاً یک مشکل بالینی دارد، از سطح دشواری بالایی برخوردار است در نتیجه دشواری باید

1. Multiple Choice Question
2. Short Answer Question

در ارتباط با بیمار مورد سؤال ارزیابی شود. در فرم‌های ارزیابی RACP، دشواری و پیچیدگی مورد بیماری توسط ارزیابان از کم تا زیاد درجه‌بندی می‌شود (شکل ۱-۲۰).

در مجموع در فرم‌های ارزیابی باید مواردی از قبیل حیطه‌های مورد ارزیابی، نحوه قضاوت در مورد عملکرد داوطلب در هر حیطه، نحوه نمره‌دهی، نمره هر سؤال و وزن نمره هر سؤال در نمره نهایی مشخص باشد. همچنین نحوه قضاوت نهایی و معیار رد و قبولی داوطلب باید مشخص باشد. دستورالعمل‌ها و راهنمایی‌های لازم در این مورد باید قبل از آزمون به ارزیابان داده شود. تصمیم‌گیری در مورد نمره رد و قبولی نیز یکی از مواردی است که قبل از آزمون باید صورت پذیرد. در سیستم ارزیابی RACP نمره قبولی ۴ (از نمره کل ۷)، در آزمون DOCEE، ۶۰ (از نمره کل ۱۰۰) و در آزمون OSLER، ۵۰ (از نمره کل ۸۰) است.

نام و نام خانوادگی داوطلب:

نیمسال : اول ☐ دوم ☐ — تاریخ ارزیابی:

سال تحصیلی: ۱ ☐ ۲ ☐ ۳ ☐ ۴ و بالاتر ☐ — شماره مورد بالینی در ترم جاری: ۱ ☐ ۲ ☐ ۳ ☐ ۴ ☐

نام ارزیاب اول: . — نام ارزیاب دوم:

سمت ارزیاب اول: — سمت ارزیاب دوم: .

ایمیل ارزیاب اول: — ایمیل ارزیاب دوم:

بیمارستان محل برگزاری آزمون:

توصیف مورد بیماری:

سن بیمار: — جنس بیمار مرد ☐ زن ☐

دشواری بیمار: کم ☐ متوسط ☐ زیاد ☐ — نوع ارزیابی: تکوینی ☐ رسمی ☐

لطفا عملکرد داوطلب را با توجه به عملکرد مورد انتظار در سال آن تحصیلی نمره‌دهی کنید

	۱	۲	۳	۴	۵	۶	۷	مشاهده نشد
۱. مهارت معاینه	☐	☐	☐	☐	☐	☐	☐	☐
۲. مهارت مصاحبه	☐	☐	☐	☐	☐	☐	☐	☐
۳. یافته‌های بالینی و تفسیر	☐	☐	☐	☐	☐	☐	☐	☐
۴. طرح کوتاه مدت مدیریت بیمار	☐	☐	☐	☐	☐	☐	☐	☐
۵. تأثیر بیماری بر بیمار و خانواده	☐	☐	☐	☐	☐	☐	☐	☐
۶. طرح بلند مدت مدیریت بیمار	☐	☐	☐	☐	☐	☐	☐	☐
۷. عملکرد کلی	☐	☐	☐	☐	☐	☐	☐	

موارد مثبت عملکرد	موارد پیشنهادی برای ارتقای عملکرد

میزان رضایت ارزیاب از روش ارزیابی: کم ۱ ۲ ۳ ۴ ۵ ۶ ۷ ۸ ۹ زیاد

میزان رضایت داوطلب از روش ارزیابی: کم ۱ ۲ ۳ ۴ ۵ ۶ ۷ ۸ ۹ زیاد

امضای اولین ارزیاب — امضای فراگیر

امضای دومین ارزیاب

شکل ۱-۲۰: فرم نمره‌دهی آزمون «مورد بالینی کامل» مربوط به RACP (https://www.racp.edu.au/)

نام و نام خانوادگی داوطلب نام ارزیاب	شماره آزمون نام ارزیاب همکار	تاریخ برگزاری

برای تکمیل فرم به راهنمای تفصیلی نمره‌دهی مراجعه شود.
لازم است ارزیابان برای عملکرد داوطلب در هر یک از سؤالات زیر قبل از بررسی با همکار خود نمره‌ای درنظر بگیرند و سپس نمره به توافق رسیده با همکار خود را ثبت کنند
نمره قبولی ۵۰ است. لطفا در مقیاس ۵ تایی (مانند ۸۰، ۷۵، ۷۰، ۶۵، ۶۰ و ...) نمره دهید. از اعداد مابین برای نمره استفاده نکنید.
ارزیابان نباید در زمان قضاوت از دادن نمرات بالا یا پائین امتناع ورزند.

درجه بندی	نمره‌دهی
P^+ = خیلی خوب/عالی	۶۰-۸۰+
P = قبول/مرزی	۵۰-۵۵
P^- = رد	۳۵-۴۵

ارائه شرح حال	نمره		نمره مورد توافق
سرعت/وضوح			
مهارت‌های ارتباطی			
ارائه سیستماتیک			
جمع‌آوری اطلاعات صحیح			
معاینه بالینی			
سیستماتیک			
تکنیک			
جمع‌آوری یافته‌های صحیح			
اداره بیمار			
تشخیص مناسب			
مدیریت مناسب			
شم بالینی			

توصیه‌های بیشتر:

لطفاً دشواری مورد را با علامت (×) مشخص نمائید

	نظر ارزیاب	نظر مورد توافق	نظر ارزیاب		نظر مورد توافق دو ارزیاب	
استاندارد			امتیاز کلی	نمره	امتیاز مورد توافق	نمره مورد توافق
دشوار						
خیلی دشوار						

شکل ۲-۲۰: فرم مربوط به OSLER (گلیسون ۱۹۹۷)

طرح تفصیلی نمره‌دهی		
p^+	۸۰	**فوق‌العاده:** ارائه صحیح و واضح شرح‌حال بیمار، نمایش علائم بالینی و سازمان‌دهی مدیریت بیماری. داوطلب واضحا مهارتهای ارتباطی و شم بالینی عالی دارد. (داوطلبان درجه ۱)
	۷۵	**عالی:** عملکرد عالی در ارائه مورد، مهارتهای ارتباطی، تکنیک‌های معاینه و نشان دادن علائم بالینی بیمار به طور صحیح. داوطلب در برخی از معیارهای ارزیابی و نه همه آنها عملکرد فوق‌العاده دارد. (داوطلبان درجه ۱)
	۷۰	**عالی در برخی جنبه‌ها:** در برخی از جنبه‌های ارائه مورد، مهارتهای ارتباطی، تکنیک‌های معاینه و نشان دادن علائم بالینی بیمار به طور صحیح عملکرد عالی دارد؛ همچنین مهارتهای ارتباطی، تشخیص و سازمان‌دهی مدیریت بیماری و شم بالینی عملکرد عالی دارد. (داوطلبان درجه ۱)
	٦٥	**بسیار خوب:** در ارائه مورد همه جنبه‌های اصلی را پوشش می‌دهد؛ موارد کمی از حذف و اولویت‌بندی مناسب. در مهارتهای ارتباطی و شم بالینی به طور وضوح بالاتر از حد متوسط است. (داوطلبان درجه ۲، طبقه ۱)
	٦٠	**بسیار خوب در برخی جنبه‌ها:** در برخی از جنبه‌های ارائه و مهارتهای ارتباطی و نه همه جنبه‌ها عملکرد بسیار خوب دارد. هر چند، در اکثر جنبه‌ها عملکرد خوب با یک شم بالینی خوب شکل گرفته دارد. (داوطلبان درجه ۲، طبقه ۲)
P	۵۵	**خوب:** در ارائه و مهارتهای ارتباطی عملکرد خوبی دارد بدون آنکه مورد غیر طبیعی دیده شود. استانداردهای کافی را در مهارتهای معاینه دارد. مشکل بیمار را تشخیص می‌دهد و چارچوب قابل قبولی برای درمان بیمار ارئه می‌دهد.
	۵۰	**متوسط:** توانایی ارائه و مهارتهای ارتباطی مناسب دارد. مهارتهای معاینه، تشخیص، مدیریت بیمار و شم بالینی را در حد قابل قبول ارئه می‌دهد. داوطلب مرزی مطمئنی است که استانداردهای قبولی را دارد.
P^-	٤٥	**متوسط:** در ارائه، مهارتهای ارتباطی و نمایش علائم بالینی عملکرد ضعیفی دارد. در پیدا کردن مشکلات بیمار تلاش بدون ثمری دارد. ممکن است داوطلب در برخی جنبه‌ها عملکرد مناسبی داشته باشد ولی در مجموع استانداردهای قبولی را ندارد.
	*	نمره ۴۰ استفاده نمی‌شود.
	۳۵	**مردود:** در ارائه، مهارتهای ارتباطی و نمایش علائم بالینی عملکرد بسیار ضعیفی دارد. به طور وضوح، این داوطلب نیازمند دوره‌های آموزشی بیشتر است.

شکل ۳-۲۰: طرح تفصیلی نمره‌دهی در OSLER (گلیسون ۱۹۹۷)

تصمیم‌گیری در مورد تعداد و ویژگی ارزیابان

به منظور افزایش پایایی آزمون در اغلب موارد از دو ارزیاب استفاده می‌شود. این موضوع به ویژه در آزمون‌های حساس و مهم صادق است. لازم به ذکر است تعداد بیشتر از سه نفر تأثیری بر پایایی آزمون ندارد. در صورتی که تعداد ارزیابان محدود است، افزایش تعداد موارد بر افزایش تعداد ارزیابان برای هر داوطلب ارجحیت دارد درنتیجه در این صورت توصیه می‌شود از یک ارزیاب برای هر داوطلب استفاده شود.

معمولاً ارزیابان در مورد بالینی کامل از استادان بخش که دانش و مهارت تخصصی لازم را دارند، انتخاب می‌شوند. در مواردی که از بیش از یک ارزیاب استفاده می‌شود، می‌توان از پزشکان عمومی یا استادان دارای تخصص مرتبط اما از بیمارستان غیر از محل برگزاری آزمون استفاده کرد. در آزمون «مورد بالینی کامل» برگزار شده در دانشگاه بازل[1] کشور سوئیس دو ارزیاب، یک نفر از استادان بخش داخلی و یک پزشک عمومی، عملکرد داوطلب را مورد ارزیابی قرار می‌دهند. زلر و همکاران[2] (۲۰۰۳) نتایج این آزمون‌ها را بررسی کردند و سنجش ارزیابان، بیماران و دانشجویان از عملکرد دانشجویان را با هم مقایسه کردند. نتایج مطالعه حاکی از ارتباط بالای قضاوت استادان بخش و پزشکان عمومی (r =۰/۸۳) و ارتباط نسبتاً پایین قضاوت دانشجویان و ارزیابان (r =۰/۴۶) بود. بیماران به طور معنی‌داری نمرات بالاتری در مقایسه با ارزیابان به

1. Basel
2. Zeller et al.

دانشجویان دادند و استادان نیز به طور معنی‌داری در مقایسه با خودارزیابی دانشجویان نمرات بالاتری به دانشجویان دادند.

در مجموع قبل از برگزاری آزمون ضروری است در مورد تعداد ارزیابان برای هر مواجهه داوطلب با بیمار و معیارهای انتخاب آنها تصمیم‌گیری شود. به عنوان مثال کالج سلطنتی پزشکان تایلند در آزمون بورد رشته داخلی معیارهای انتخاب ارزیاب را به صورت زیر تعریف کرده است:

- دو ارزیاب برای هر داوطلب، یک ارزیاب از استادان بخش داخلی و ارزیاب دیگر از بیمارستان دیگری که دارای بخش داخلی است.
- سن ارزیابان بین ۳۵ تا ۶۵ سال باشد.
- دارای مدرک بورد بیماری‌های داخلی یا رشته‌های مرتبط باشند.
- سابقه آموزشی حداقل ۱۰ سال به عنوان متخصص داخلی داشته باشند.
-

توجیه و آموزش ارزیابان و اطلاع‌رسانی

قبل از آزمون لازم است ارزیابان با فرم‌های ارزیابی و شیوه نمره‌دهی آشنا شوند. برای رسیدن به این هدف می‌توان راهنماهایی را برای ارزیابان تدوین نمود. در صورتی که بیش از یک ارزیاب برای هر داوطلب استفاده می‌شود باید در مورد نمره‌دهی مستقل از هم توجیه شوند. یکی از اشکالاتی که به آزمون مورد بالینی سنتی وارد است، متفاوت بودن موارد بیماری و در نتیجه تفاوت در دشواری آنها برای داوطلبان مختلف است. ارزیابان باید توجیه شوند که درجه دشواری مورد بیماری را در زمان قضاوت در مورد عملکرد فراگیر و نمره‌دهی آن در نظر بگیرند. همچنین ضروری است مانند دیگر روش‌های ارزیابی مبتنی بر محل کار برنامه‌های مدونی به منظور توانمندسازی اعضای هیأت علمی در نظر گرفته شود. این برنامه‌ها به منظور بهبود مهارت‌های مشاهده و قضاوت اجرا می‌شود و می‌تواند منجر به بهبود ویژگی‌های روان‌سنجی آزمون از جمله پایایی و روایی آن شود. در نهایت، به منظور آشناسازی داوطلبان با اهداف، محتوا و نحوه برگزاری آزمون باید سیلابوس دوره، بلوپرینت آزمون، فرم‌های نمره‌دهی و راهنماهای برگزاری آزمون در اختیار داوطلبان و ارزیابان قرار گیرد.

اجرای آزمون

لازم است بلافاصله قبل از شروع آزمون ارزیابان با بیماران ملاقات کنند. این ملاقات به منظور آشنایی با بیمار، تطبیق فرم‌های ارزیابی، تخمین دشواری موارد بیماری و از همه مهم‌تر هماهنگی و ثبات قضاوت ارزیابان (در صورت استفاده از بیش از یک ارزیاب) انجام می‌شود. زمان مناسب این ملاقات معمولاً ۱۵ تا ۲۰ دقیقه است. در برخی از مطالعات این زمان ۴۵ دقیقه نیز گزارش شده است.

همان‌طور که در ابتدای فصل اشاره شد، در آزمون «مورد بالینی کامل» سنتی، داوطلب پس از اخذ شرح‌حال و معاینه بیمار، موارد مربوط را در مقابل ارزیاب یا ارزیابان ارائه می‌دهد و فرایند مواجهه داوطلب با بیمار مورد مشاهده قرار نمی‌گیرد. عدم مشاهده مهارت‌های اخذ شرح‌حال و معاینه بیمار ممکن است مانع از تشخیص مشکلات و نقص‌های داوطلبان در این مهارت‌ها شود. این نقص‌ها می‌تواند شامل موارد زیر باشد:

- گوش ندادن به صحبت‌های بیمار
- عدم اجازه به بیمار برای تمام کردن صحبت‌هایش
- استفاده از سؤالات بسته‌پاسخ با پاسخ‌های «بله» و «خیر»
- استفاده از سؤالات نامناسب
- عدم دقت و تأمل به پیشنهادهای بیمار
- حذف برخی از قسمت‌ها مانند سابقه خانوادگی مشکل بیمار

- مدیریت نامناسب زمان
- عدم رعایت ترتیب منطقی پرسش‌ها
- رویکرد غیرسیستماتیک به معاینه بیمار (رویکرد سیستماتیک شامل مشاهده، لمس، دق و سمع)
- حذف بخش مشاهده در معاینه فیزیکی
- معاینه ناکامل برخی از سیستم‌ها شامل سیستم تنفسی، عضلانی اسکلتی و نورولوژیک
- معاینه ناکامل قدرت عضلانی و رفلکس‌های وتری در معاینه سیستم عصبی

این موارد ممکن است به طور شایعی در ارائه آزمون «مورد بالینی کامل» رخ دهد، بدون آنکه تشخیص داده شود. در واقع، یک ارائه و بحث بالینی خوب لزوماً تضمین کننده معاینه بالینی خوب نیست. یکی از بهترین راهکارها برای غلبه بر محدودیت ذکر شده در بالا، مشاهده روند شرح‌حال‌گیری و معاینه فیزیکی در آزمون است. مشاهده می‌تواند مستقیم یا مرور فیلم‌های ضبط شده در نوارهای ویدیوئی باشد. در مطالعه پاولاکیس و لورنت[1] (۲۰۰۱) نشان داده شد مشاهده «مورد بالینی کامل» می‌تواند ضعف‌های تکنیکی و سازمان‌دهی داوطلبان در اخذ شرح‌حال، معاینه فیزیکی و همچنین مدیریت زمان را مشخص کند. در نتیجه می‌توان با ارائه بازخورد به موقع و سازنده به داوطلبانی که عملکرد ضعیف داشته‌اند، تأثیر آموزشی این آزمون را افزایش داد. همچنین نتایج این مطالعه نشان داد بسیاری از داوطلبان با وجود مشکلاتی که در مهارت‌های شرح‌حال‌گیری و معاینه فیزیکی دارند نمره قبولی کسب نمودند. این نتایج تأیید می‌کند در بسیاری از موارد آزمون «مورد بالینی کامل» سنتی نمی‌تواند ضعف عملکرد داوطلبان را در مهارت‌های مذکور مشخص کند.

در مرحله بعد نوبت به ارزیابی عملکرد داوطلب می‌رسد که لازم است بر اساس معیارهای مشخص عملکرد داوطلب ارزیابی و سپس به صورت ساختارمند نمره‌دهی شود. نمونه‌ای از سؤالات و معیارهایی که می‌توان در «مورد بالینی کامل» به کار گرفت، در بخش فرم‌های ارزیابی توضیح داده شد. افزایش جنبه‌های مورد ارزیابی پایایی آزمون را افزایش می‌دهد. البته افزایش تعداد جنبه‌های مورد ارزیابی بیش از ۱۰ مورد تأثیری بر پایایی آزمون ندارد و علاوه بر آن زمان آزمون را افزایش می‌دهد.

در انتها به فراگیران در مورد عملکردشان بازخورد ارائه می‌شود. بازخورد مهم‌ترین و در واقع نقطه عطف روش‌های ارزیابی مبتنی بر محل کار است. لازم است فرصت کافی برای آن در نظر گرفته شود و ارزیابان آموزش کافی در مورد نحوه ارائه بازخورد مناسب دریافت کنند.

نظرخواهی از فراگیران و ارزیابان

پس از اجرای آزمون‌های مبتنی بر محل کار، اقداماتی به منظور اطمینان از کیفیت آزمون برگزارشده انجام می‌شود که در تقریباً تمامی روش‌های ارزیابی این خانواده مشترک است. یکی از این اقدامات، نظرخواهی از داوطلبان و ارزیابان در مورد کیفیت آزمون برگزار شده و بررسی میزان رضایت آنهاست. در این بخش تنها به درج نمونه‌ای از این فرم‌ها در مورد کیفیت آزمون «مورد بالینی کامل» بسنده می‌کنیم. جدول شماره ۳-۲۰، فرم نظرخواهی از دانشجویان و جدول شماره ۴-۲۰، فرم نظرخواهی از ارزیابان را نشان می‌دهد. در «مورد بالینی کامل» برگزار شده در سال چهارم دوره پزشکی عمومی دانشگاه نیوکاسل[2]، برای هر مواجهه، دانشجو و دو ارزیاب فرم مربوط را تکمیل می‌کنند.

1. Pavlakis & Laurent
2. Newcastle

جدول ۳-۲۰ : فرم نظرخواهی از داوطلبان (السون و همکاران ۲۰۰۰)

ردیف	سؤالات	کاملاً موافقم	تا حدودی موافقم	نه موافقم و نه مخالفم	تا حدودی مخالفم	کاملاً مخالفم
۱	به من فرصت داده شد تا دانش خود را در همه حیطه‌ها (مثل استدلال بالینی، تولید هیپوتز و تشخیص افتراقی) نشان دهم					
۲	سؤالات زیادی در مورد یک حیطه خاص از من پرسیده شد					
۳	به نظر می‌رسید ارزیابان با مشکل بیمار آشنا بودند					
۴	احساس کردم ارزیابان از استاندارد مورد نیاز در سطح آموزشی من آگاه هستند					
۵	آزمون در مجموع عادلانه بود					
۶	مواردی وجود داشت که علی‌رغم انتظار من پرسیده نشد					
۷	به نظر می‌رسید ارزیابان از آمادگی لازم به منظور برگزاری آزمون برخوردارند. (برای مثال می‌دانند چه می‌پرسند)					
۸	در مجموع عملکرد من در آزمون توانایی من را منعکس کرد. (برای مثال نمره‌ای را که مستحق آن هستم دریافت می‌کنم)					

جدول ٤-۲۰ : فرم نظرخواهی از ارزیابان (السون و همکاران ۲۰۰۰)

ردیف	سؤالات	کاملاً موافقم	تا حدودی موافقم	نه موافقم و نه مخالفم	تا حدودی مخالفم	کاملاً مخالفم
۱	به دانشجویان فرصت داده شد تا دانش خود را در همه حیطه‌ها (مثل استدلال بالینی، تولید هیپوتز و تشخیص افتراقی) نشان دهند					
۲	سؤالات زیادی در مورد یک حیطه خاص پرسیده نشد					
۳	من با جزئیات شرح‌حال و معاینه بیمار آشنا بودم					
۴	من از استاندارد مورد نیاز در سطح آموزشی دانشجویان آگاهی داشتم					
۵	از قبل می‌دانستم چه سؤالاتی بپرسم					
۶	در انتها نتوانستم تمام مواردی که مد نظر داشتم را بپرسم					
۷	می‌توانستم در مورد رد یا قبولی دانشجو تصمیم بگیرم					
۸	فکر می‌کنم در مجموع قضاوت عادلانه‌ای داشتم					

سودمندی آزمون «مورد بالینی کامل»

پایایی آزمون «مورد بالینی کامل»

یکی از مهم‌ترین محدودیت‌های «مورد بالینی کامل» سنتی که استفاده از آن را محدود کرده است، پایایی پایین و عدم تعمیم‌پذیری نتایج آن است. در طول سی سال گذشته پژوهش‌های زیادی بر عدم تکرارپذیری نتایج این آزمون تأکید کرده‌اند. به عنوان نمونه:

- امتحان شفاهی بورد داخلی آمریکا برای بیماری‌های قلبی و عروقی در اوایل دهه ۱۹۷۰ شامل دو «مورد بالینی کامل» با دو ارزیاب بود. ضریب تعمیم‌پذیری نتایج این آزمون ۰/۳۹ بود؛ به این معنی که ۳۹ درصد تغییرپذیری نمرات مربوط به تفاوت‌های موجود در توانایی فراگیران و ۶۱ درصد آن ناشی از خطای اندازه‌گیری بود. کاهش موارد به یک مواجهه این ضریب را به ۰/۲۴ کاهش داد (نورسینی ۲۰۰۲).

در گذشته دانشکده‌های پزشکی تنها یک مورد را ارزیابی می‌کردند. احتمالاً پیش‌فرض آنها این بود که عملکرد یک دانشجو در یک مورد می‌تواند پیش‌گویی‌کننده عملکرد او در همه موارد باشد اما زمانی که اهمیت ویژگی مورد مشخص شد جهت‌گیری ارزیابی بالینی به سمت روش‌های ارزیابی ساختارمند چندایستگاهی مانند OSCE پیش رفت. مطالعات نشان داده‌اند در «مورد بالینی کامل» نیز افزایش موارد، پایایی آزمون را افزایش می‌دهد:

- مطالعه وس و همکاران (۲۰۰۱) نشان داد با استفاده از هشت تا ده مورد بالینی و ارزیابی هر مورد توسط ارزیاب متفاوت، پایایی به بالاتر از ۰/۸ رسید. البته زمان هر مواجهه در این مطالعه ۲۰ دقیقه بود. در این مطالعه همچنین نشان داده شد که در صورت افزایش تعداد ارزیابان به دو ارزیاب برای هر مواجهه، میزان افزایش پایایی محدود است؛ به طوری که پایایی آزمون با ۱۰ مورد از ۰/۸۵ به ۰/۸۸ رسید. در این مطالعه پایایی بین ارزیابان در صورت مشاهده تعامل دانشجو با بیمار افزایش یافت.
- مطالعه ویلکینسون و همکاران (۲۰۰۸) نشان داد با استفاده از دو ارزیاب برای هر مواجهه، ضریب پایایی برای یک مورد بالینی کامل ۰/۴۳ و برای دو مورد ۰/۶۰ بود و در صورت استفاده از شش مورد بالینی کامل این ضریب به ۰/۸۲ رسید.
- در آزمون DOCEE با چهار مورد مواجهه و دو ارزیاب برای هر مواجهه، ضریب آلفای کرونباخ ۰/۸۵ و ضریب تعمیم‌پذیری ۰/۸۴ گزارش شد که رقم بسیار بالایی است و احتمالاً به دلیل اصلاحاتی است که در این آزمون به منظور غلبه بر مشکلات «مورد بالینی کامل» سنتی انجام شده است. واضح است که فراهم نمودن این تعداد مورد بیماری برای آزمون نیاز به صرف وقت زیادی دارد که عملاً قابلیت اجرای آزمون را کاهش می‌دهد (حمدی و همکاران ۲۰۰۳).

البته علی‌رغم اینکه منابع بسیاری از جمله موارد مذکور تأکید کرده‌اند عملکرد خوب فراگیر در یک مورد بالینی کامل نمی‌تواند نتایج خوب در موارد دیگر را تضمین کند:

- السون در سال ۱۹۹۹ در مطالعه خود با مقایسه نمرات آزمون «مورد بالینی کامل» در چهار رشته طب داخلی، اطفال، زنان و زایمان و جراحی در طول شش سال نشان داد به استثنای دانشجویان مرزی، یک مورد بالینی کامل می‌تواند پیشگویی‌کننده عملکرد دانشجو در یک رشته باشد. او همچنان نشان داد برای دانشجویان مرزی دو «مورد بالینی کامل» برای پیشگویی پیامد در این چهار رشته کافی است.
- همچنین در مطالعه ویلکینسون و همکاران (۲۰۰۸) نیز واریانس نمرات مربوط به تفاوت موارد بیماری بسیار ناچیز بود و ویژگی مورد و تغییرپذیری مربوط به تفاوت بیماران آنقدر که انتظار می‌رفت، در واریانس نمرات نقش نداشت.
- این گونه مطالعات که از توانایی تعمیم‌پذیری مورد بالینی کامل حمایت می‌کند ناچیز هستند و باید در مقابل شواهد فراوان مخالف تفسیر شوند.

در مجموع به نظر می‌رسد با توجه به اینکه افزایش تعداد مواجهه نیازمند صرف وقت زیادی برای آزمون است، همان‌طور که در فصل اول نیز در مورد تعداد مواجهه مورد نیاز در آزمون‌های مبتنی بر محل کار بیان شد، لزومی به اختصاص زمان چهار تا

پنج ساعت برای همه داوطلبان نیست. بلکه این زمان می‌تواند برای فراگیرانی که مشخصاً رد یا قبول هستند، دو ساعت و برای دانشجویان مرزی بیشتر باشد (سنجش متوالی). همچنین با توجه به مطالعات مختلف از جمله مطالعه وس و حمدی به نظر می‌رسد اختصاص زمان سه ساعت با چهار «مورد بالینی کامل»، ضریب تعمیم‌پذیری بالای ۰/۸ را فراهم می‌کند.

در مورد پایایی بین ارزیابان، استفاده از راه‌کارهای افزایش تعداد ارزیابان، آموزش ارزیابان، توافق ارزیابان قبل از آزمون (در مورد جنبه‌های مورد ارزیابی)، استفاده از فرم‌های ساختارمند برای نمره‌دهی و در نهایت به کار بردن روش‌های آماری به منظور حذف تفاوت بین ارزیابان با هم افزایش قابل قبولی در تکرارپذیری نتایج ایجاد می‌کنند. در این رابطه نتیجه چند مطالعه به این صورت است:

- در مطالعه ویلکینسون و همکاران، واریانس بین ارزیابان در حد بسیار ناچیز بود که پژوهشگران علت احتمالی آن را استفاده از سیستم نمره‌دهی ساختارمند و آموزش ارزیابان دانسته‌اند اما علت دیگر آن می‌تواند توافق ارزیابان قبل از ثبت نهایی نمره داوطلب باشد.
- در مطالعه حمدی، رابطه بین ارزیابان متخصص و غیرمتخصص (ارزیابان با تخصص غیرمرتبط با مورد بیماری) در آزمون DOCEE با استفاده از ضریب همبستگی پیرسون محاسبه شد که بین ۰/۸۲ تا ۰/۹۳ به دست آمد. همچنین در این آزمون توافق بین نمرات ارزیابان متخصص برای هر مواجهه نیز، خوب و مشابه توافق بین نمرات ارزیابان یک ایستگاه در OSCE گزارش شد.
- در مطالعه چیراکول و همکاران[1] (۲۰۱۰) نیز با استفاده از دو ارزیاب برای هر داوطلب ضریب پایایی بین ارزیابان خوب (بیش از ۰/۷۵) گزارش شد.

همچنین مطالعات نشان می‌دهند که افزایش تعداد ارزیابان در مقایسه با افزایش تعداد موارد، به میزان کمتری پایایی آزمون را بهبود می‌بخشد. در مطالعه وس و همکاران، افزایش ارزیابان به دو نفر پایایی را از ۰/۸۵ به ۰/۸۷ برای ده «مورد بالینی کامل» افزایش داد. تفاوت در پایایی این آزمون با یک و دو ارزیاب و با تعداد متفاوت مورد بالینی کامل در جدول ۵-۲۰ آمده است (وس و همکاران ۲۰۰۱).

جدول ۵-۲۰ : مقایسه پایایی آزمون «مورد بالینی کامل» و OSCE در زمان‌های برابر با استفاده از ضریب تعمیم‌پذیری (وس و همکاران ۲۰۰۱)

	مطالعه اول: ۷۹ داوطلب			مطالعه دوم: ۲۱۴ داوطلب	
زمان (دقیقه)	تعداد موارد بالینی کامل (تعداد ایستگاه‌های OSCE)	مورد بالینی کامل یک ارزیاب برای هر مورد	مورد بالینی کامل دو ارزیاب برای هر مورد	مورد بالینی کامل یک ارزیاب برای هر مورد	OSCE یک ارزیاب برای هر ایستگاه
±۲۰	۱ (۳)	۰/۳۶	۰/۱۴	۰/۳۴	۰/۲۱
±۴۰	۲ (۶)	۰/۵۳	۰/۵۸	۰/۵۰	۰/۳۴
±۶۰	۳ (۹)	۰/۶۳	۰/۶۸	۰/۶۰	۰/۴۴
±۱۰۰	۵ (۵۱)	۰/۷۴	۰/۷۸	۰/۷۲	۰/۵۷
±۱۲۰	۶ (۱۸)	۰/۷۷	۰/۸۱	۰/۷۵	۰/۶۱
±۱۴۰	۷ (۱۲)	۰/۸۰	۰/۸۳	۰/۷۸	۰/۶۵
±۱۶۰	۸ (۴۲)	۰/۸۲	۰/۸۵	۰/۸۰	۰/۶۸
±۱۸۰	۹ (۷۲)	۰/۸۳	۰/۸۶	۰/۸۲	۰/۷۰
±۲۰۰	۱۰ (۳۰)	۰/۸۵	۰/۸۷	۰/۴۸	۰/۷۲

1. Chierakul et al.

مسأله دیگری که در راستای ساختارمند کردن «مورد بالینی کامل» مطرح است، این است که آیا افزودن بیمار استاندارد به مورد بالینی کامل پایایی آن را افزایش می‌دهد یا خیر. مطالعات نشان می‌دهند افزودن بیمار استانداردشده و عینیت بخشیدن به مواجهه، دستاورد کمی در مقایسه با راهبردهای غلبه بر ویژگی مورد دارد. پژوهش‌های بیشتری مورد نیاز است تا مشخص شود روش‌های ترکیبی مانند استفاده از بیمار واقعی در OSCE به منظور کاهش محدودیت‌های هر دو روش چه تأثیری بر ویژگی‌های روان‌سنجی این روشها دارد.

روایی آزمون «مورد بالینی کامل»

از آنجا که «مورد بالینی کامل»، تعامل بالینی بین پزشک و بیمار واقعی را ارزیابی می‌کند، به محیط طبابت واقعی پزشک بسیار نزدیک است و دارای روایی صوری بالایی است. استفاده از بلوپرینت برای انتخاب موارد بیماری و افزایش موارد بیماری، روایی محتوایی آزمون را افزایش می‌دهد:

- هاردی و همکاران[1] (۱۹۹۸) پس از برگزاری آزمون «مورد بالینی کامل» و «مورد بالینی کوتاه» در بخش جراحی، میزان پوشش اهداف دوره جراحی در این دو آزمون را از طریق پرسشنامه مورد بررسی قرار دادند. ۹۰ درصد از ارزیابان این آزمون را در پوشش اهداف دوره موفق دانستند. به جز ارزیابی موارد حاد جراحی که ۴۵ درصد دستیابی به این هدف را توسط آزمون مورد بالینی کامل و مورد بالینی کوتاه ذکر کردند.

در مورد روایی معیار آزمون «مورد بالینی کامل» هم تحقیقاتی انجام شده است:

- هاردی و همکاران نشان دادند بین نمرات آزمون «مورد بالینی کامل» در بخش جراحی و معدل پایان دوره پزشکی عمومی ارتباط ضعیفی وجود داشت. در صورتی که آزمون «مورد بالینی کوتاه» ارتباط خوبی با معدل دوره داشت. پژوهشگران نتیجه گرفتند احتمالاً «مورد بالینی کامل» جنبه متفاوتی از مهارت‌های بالینی را ارزیابی می‌کند.
- حمدی و همکاران نشان دادند که DOCEE با آزمون‌های PMP، SAQ، MCQ و OSCE که به صورت هم‌زمان برگزار شده بودند، ضریب همبستگی بین ۰/۶۷ تا ۰/۷۹ (MCQ=۰/۶۷ ، SAQ=۰/۷۹، PMP=۰/۷۰ و OSCE= ۰/۷۵) داشت. نویسندگان اعلام کردند با در نظر گرفتن این پیش‌فرض که انتظار داریم آزمون DOCEE توانمندی‌های بالینی غیر از آزمون‌های مذکور را بسنجد، این ضرایب همبستگی مناسب و قابل‌قبول است.
- چیراکول و همکاران (۲۰۱۰) ارتباط بسیار ضعیف (r= ۰/۲۵) اما معنی‌داری بین نمرات «مورد بالینی کامل» به عنوان آزمون میان‌ترم و آزمون پایان‌ترم یافتند. البته این احتمال وجود دارد که انتظار بالای ارزیابان از داوطلبان در آزمون پایان ترم موجب نمره‌دهی پایین آن‌ها شده باشد.
- دز مطالعه السون و همکاران، نمرات آزمون «مورد بالینی کامل» با نمرات آزمون کتبی انتهای سال ارتباط بسیار ضعیفی (r=۰/۲۷) داشت (السون و همکاران ۲۰۰۰).

تأثیر آموزشی آزمون «مورد بالینی کامل»

اگر در نظر بگیریم دانشجویان به یادگیری آن چیزی می‌پردازند که در آزمون آنها گنجانده می‌شود، «مورد بالینی کامل» باعث می‌شود فراگیران به یادگیری از بیمار در محیط واقعی ترغیب شوند. به عنوان مثال در دانشکده‌های پزشکی استرالیا، جایی که «مورد بالینی کامل» بخشی از ارزیابی سالانه دستیاران در محیط واقعی را تشکیل می‌داد، دیده شد که با حذف این روش از ارزیابی دستیاران سال چهارم در یکی از دانشکده‌ها، دستیاران کمتر بیمار می‌دیدند و بیشتر وقت خود را در کتابخانه به منظور آمادگی در آزمون کتبی می‌گذراندند. همچنین تجربه دانشکده‌های پزشکی استرالیا نشان می‌دهد با اضافه شدن آزمون «مورد بالینی کامل» در امتحانات کالج سلطنتی پزشکان استرالیا، فراگیران زمان بیشتری را صرف

1. Hardy et al.

معاینه بیمار می‌کنند و یافته‌هایشان را با فراگیران سطوح بالاتر کنترل می‌کنند.

در کل به نظر می‌رسد «مورد بالینی کامل» موقعیتی را برای یادگیری ادغام‌یافته فراهم می‌آورد تا دانشجویان به بیمار به عنوان یک کل بنگرند و نه مجموعه‌ای از سیستم‌های حیاتی. با وجود این، در مورد اثر آزمون بر میزان یادگیری مطالعات اندکی انجام شده است، به ویژه پژوهش‌هایی که به مقایسه تأثیر آموزشی «مورد بالینی کامل» با OSCE بپردازد.

یکی از مواردی که تأثیر آموزشی نوع غیرسنتی این آزمون را افزایش می‌دهد مشاهده فرایند اخذ شرح‌حال و معاینه فیزیکی است. از طریق مشاهده، مشکلات و ضعف‌های عملکرد فراگیران در این مهارت‌ها مشخص شده و با ارائه بازخورد موجب ارتقای آن می‌شود. همچنین اشکالات موجود در تدریس و برنامه درسی نیز با این روش آشکار می‌شود. به عنوان مثال پاولاکیس و لورنت (۲۰۰۱) با مشاهده «مورد بالینی کامل» در دستیاران نشان دادند علی‌رغم تصور شایع، دستیاران از مهارت‌های اخذ شرح‌حال در حد استاندارد و مورد انتظار برخوردار نیستند و بر لزوم آموزش این مهارت‌ها در برنامه‌های دستیاری تأکید کردند.

قابلیت اجرا و هزینه آزمون مورد بالینی کامل

آزمون «مورد بالینی کامل» یک آزمون زمان‌بر است. البته زمان زیادی از آزمون به تعامل داوطلب با بیمار اختصاص دارد که در این مدت، داوطلب مورد مشاهده قرار نمی‌گیرد. وس و جالی در این که این زمان به عنوان زمان آزمون در نظر گرفته شود، تردید دارند. مطالعات نشان داده‌اند که مدل IDOCEE نسبت به مدل سنتی آن از کارایی بالاتری برخوردار است (آبونا[1] و حمدی ۱۹۹۹).

افزایش موارد و ارزیابان به منظور افزایش پایایی قابلیت اجرای آزمون را با چالش مواجه می‌کند. اما با این وجود، این روش در مقایسه با OSCE نیاز به منابع کمتری دارد. به عنوان مثال برای اجرای ۱۰ «مورد بالینی کامل»، نیازمند ۱۰ ارزیاب متفاوت هستیم در حالی که برای اجرای آزمون OSCE با زمان برابر و پایایی قابل مقایسه، ۳۰ ایستگاه لازم است که به نوبه خود نیازمند ۳۰ ارزیاب مختلف است. در زمینه هزینه–اثربخشی آزمون «مورد بالینی کامل» مطالعه‌ای یافت نشد. به نظر می‌رسد مواردی که به منظور بهبود و اصلاح مورد بالینی کامل سنتی پیشنهاد شده است، هزینه این آزمون را افزایش می‌دهد.

مقبولیت آزمون «مورد بالینی کامل»

همان‌گونه که در قسمت پایایی عنوان شد، تفاوت بین ارزیابان باعث می‌شود سؤالاتی که از داوطلبان مختلف در آزمون «مورد بالینی کامل» پرسیده می‌شود متفاوت باشد. این موضوع علاوه بر تأثیر بر پایایی آزمون، عادلانه بودن آن را نیز زیر سؤال می‌برد. در زیر به نتایج چندین مطالعه در خصوص مقبولیت «مورد بالینی» کامل اشاره می‌شود. البته مطالعات این مقوله انگشت شمار هستند و در صورت انجام نیز تنها به ارزشیابی در سطح اول هرم کرک پاتریک اکتفا نموده‌اند.

- در جلسه گروه متمرکزی[2] که در دانشکده پزشکی دانشگاه نیوکاسل با هدف بررسی تجربیات داوطلبان در مورد آزمون OSLER انجام شد، داوطلبان به یکسان نبودن سؤالات مطرح شده توسط ارزیابان و وابستگی این سؤالات با تخصص ایشان اشاره داشتند. پژوهشگران ادعا کردند که ساختارمند کردن سیستم نمره‌دهی باعث پذیرش بیشتر آزمون توسط ارزیابان و داوطلبان نسبت به مورد بالینی کامل سنتی شد، موجب افزایش عینیت آزمون در حیطه‌های مورد ارزیابی، سؤالاتی که پرسیده می‌شوند و نمره‌دهی آزمون شد اما در درک دانشجویان از ناعادلانه بودن آزمون و توانایی ارزیابان در نمره‌دهی دانشجویان تغییری ایجاد نکرد (السون و همکاران ۲۰۰۰).
- در مطالعه زلر و همکاران (۲۰۰۳)، ارزیابان، بیمار و داوطلبان آزمون «مورد بالینی کامل» را عادلانه ارزیابی کردند.

1. Abouna
2. Focus group

همچنین اکثر دانشجویان (۸۰ درصد) و اعضای هیأت علمی (بیش از ۶۰ درصد) معتقد بودند که این آزمون روش مناسبی برای ارزیابی مهارت‌های بالینی و ارتباط پزشک و بیمار است.

- آبونا و حمدی (۱۹۹۹) نیز در مطالعه خود نشان دادند که بیش از ۹۰ درصد ارزیابان و داوطلبان از ساختار و سازمان‌دهی آزمون IDOCEE و همچنین توانایی این آزمون در سنجش مهارت‌های بالینی رضایت داشتند.
- در مطالعه هاردی و همکاران، اکثر داوطلبان معتقد بودند مورد بالینی کامل و کوتاه ارزیابی مناسبی از مهارت‌های آن‌ها فراهم می‌کند. در این مطالعه برای داوطلبان مقبولیت آزمون «مورد بالینی کامل» بسیار بیشتر از «مورد بالینی کوتاه» بود. میزان پذیرش ارزیابان نیز بالا و برای هر دو آزمون مشابه بود (بیش از ۹۰ درصد).

از یک سو با توجه به مزایای غیرقابل چشم‌پوشی «مورد بالینی کامل» نمی‌توان به آسانی این روش ارزیابی را کنار گذاشت و از سوی دیگر بدون پشتوانه پژوهشی قوی نمی‌توان آن را مبنای تصمیم‌گیری عملکرد داوطلبان در موارد حساس و مهم قرار داد. در اینجا مواردی که در خصوص این آزمون نیاز به بررسی بیشتر دارند پیشنهاد شده است:

- **تعداد موارد:** چه تعداد «مورد بالینی کامل» لازم است تا یک ارزیابی پایا از توانمندی بالینی فراهم کند؟
- **تعداد ارزیابان:** ارزیابان چقدر در پایایی آزمون نقش دارند و در صورتی که تعداد مناسب مورد بالینی فراهم باشد، چه تعداد ارزیاب مورد نیاز است؟
- **روایی سازه:** اثر مورد بالینی واقعی چیست؟ ارزش افزوده استفاده از بیمار واقعی در مقابل بیمار استانداردشده چیست؟ چگونه «مورد بالینی کامل» می‌تواند با روش‌های دیگر ترکیب شود؟
- **تأثیر آموزشی:** یک «مورد بالینی کامل» پایا چه تأثیرات آموزشی در مقایسه با OSCE دارد؟
- **کاربرد:** با توجه به این که OSCE روی اجزای توانمندی بالینی متمرکز است و ما نیازمند روشی هستیم که تعامل دانشجو با بیمار را به صورت جامع بررسی کند، بالاخره مورد بالینی کامل در صحنه رقابت باقی بماند یا نه؟

در مجموع می‌توان گفت اخیراً تلاش‌های زیادی در جهت اصلاح آزمون «مورد بالینی کامل» شامل افزایش تعداد موارد، مشاهده آن، ساختارمند کردن و استانداردسازی آزمون انجام شده است. همچنین اخیراً نظام‌های ارزیابی به منظور کاهش معایب «مورد بالینی کامل» این روش ارزیابی را به همراه نتایج آزمون‌های دیگر مبنای قضاوت در مورد عملکرد فراگیر قرار می‌دهند. به هر حال این اقدامات جدید هستند و هنوز اطلاعات کافی و محکمی در مورد ویژگی‌های روان‌سنجی آن‌ها منتشر نشده است. در نتیجه ما به مستندات و شواهد بیشتری در مورد کارآیی این آزمونها نیاز داریم قبل از آنکه با هدف ارزیابی تراکمی به کار روند.

توصیه‌های عملی

در مجموع توصیه می شود آزمون «مورد بالینی کامل»:

- برای آزمون‌هایی که حساسیت کمتر دارند و تکوینی هستند، استفاده شود.
- در صورت استفاده در آزمون‌های مهم و تراکمی، تنها معیار تصمیم‌گیری عملکرد فراگیر نباشد.
- به منظور افزایش روایی و پایایی«مشاهده» بخشی از فرایند مصاحبه و معاینه توسط ارزیاب باشد.
- به منظور افزایش روایی و پایایی، از فرم نمره‌دهی ساختارمند مانند آزمون OSLER یا فرم‌های استفاده شده در سیستم RACP استفاده شود.
- از ارزیابان آموزش‌دیده استفاده شود تا بدین وسیله تفاوت در قضاوت و سخت‌گیری کاهش یابد.

سؤالات رایج در مورد آزمون «مورد بالینی کامل»

مشاهده مستقیم چه ارزش افزوده‌ای برای آزمون «مورد بالینی کامل» به همراه دارد؟

علی‌رغم پیشرفت‌های اخیر در فن‌آوری، مطالعات نشان می‌دهند که مهارت‌های مصاحبه و معاینه بخش بسیار مهمی از مدیریت بیماری توسط پزشکان را تشکیل می‌دهند. همچنین ارتباط بیمار و پزشک رابطه مستقیمی با پیامدهای بیمار دارد. در نتیجه توانایی اعضای هیأت علمی در مشاهده دقیق فراگیران در زمان انجام این مهارت‌ها و ارائه بازخورد مؤثر یکی از مهم‌ترین جنبه‌های آموزش پزشکی است. هر چند بیماران استانداردشده می‌توانند به عنوان مکمل در ارزیابی و ارائه بازخورد به کار روند، اما نمی‌تواند جایگزین نقش محوری مشاهده‌گری اعضای هیأت علمی شوند.

هامپتون و همکاران[1] در مطالعه خود در سال ۱۹۷۵ نشان دادند که یک شرح‌حال پزشکی خوب منجر به تشخیص پزشکی مناسب در ۸۲ درصد از موارد شد و تست‌های آزمایشگاهی تنها در یک مورد از ۸۰ مورد توانستند اطلاعاتی بیش از معاینه و مصاحبه فراهم آورند. پترسون و همکاران[2] نیز در سال ۱۹۹۲ نشان دادند مصاحبه با بیمار در ۷۶ درصد موارد منجر به تشخیص نهایی صحیح در ۸۰ بیمار مراجعه‌کننده به درمانگاه (برای بار اول) شد.

با توجه به اهمیت مهارت‌های مصاحبه و معاینه در تشخیص بیماری‌ها، این سؤال مطرح است که مشاهده روند این مهارت‌ها در آزمون «مورد بالینی کامل» چه تأثیری بر کیفیت آن دارد؟

- مطالعات اولیه‌ای که در آزمون «مورد بالینی کامل» از مشاهده استفاده کردند، مشاهده را در این آزمون توصیه نمودند اما در مورد ویژگی‌های روان‌سنجی آزمون اطلاعاتی ارائه نکردند (گلیسون ۱۹۹۴، نیوبل[3] ۱۹۹۴، پرایس[4] ۱۹۹۴).
- وس و جالی در مطالعه خود به این نتیجه رسیدند که مشاهده فرآیند مصاحبه با بیمار در آزمون «مورد بالینی کامل» بخش مجزایی از توانمندی بالینی را می‌سنجد که در نوع سنتی آزمون مورد ارزیابی قرار نمی‌گیرد و در نتیجه روایی «مورد بالینی کامل» را افزایش می‌دهد. در این مطالعه، پایایی بین ارزیابان در موارد مشاهده‌شده بالاتر بود اما میانگین نمرات دانشجویان در موارد مشاهده‌شده کمی پایین‌تر از ارائه مدل سنتی بود البته دامنه و توزیع نمرات برای هر دو گروه تفاوت مشخصی نداشت.
- پاولاکیس و لورنت در مطالعه خود با مشاهده روند آزمون «مورد بالینی کامل»، اشتباهاتی را که در بخش‌های مختلف اخذ شرح‌حال، معاینه فیزیکی، ارائه و بحث توسط دستیاران روی داد، نشان دادند. پژوهشگران نتیجه گرفتند مشاهده «مورد بالینی کامل» باعث افزایش روایی و تأثیر آموزشی آزمون می‌شود که هم بر داوطلبان و هم بر برنامه آموزشی است.
- زلر و همکاران نیز به مدت ۱۰ دقیقه در ابتدای «مورد بالینی کامل»، مهارت شرح‌حال‌گیری داوطلبان را مورد ارزیابی قرار دادند. پژوهشگران نتیجه گرفتند مشاهده، نمرات این آزمون را بهبود می‌بخشد و در این صورت مورد بالینی کامل یک ابزار مفید و معتبر در ارزیابی مهارت‌های بالینی فراگیران پزشکی است.
- بر خلاف مطالعات پیشین، مطالعه ویلکینسون و همکاران از تأثیر مشاهده بر پایایی «آزمون مورد بالینی» کامل حمایت نمی‌کند. این محققان معتقد هستند تأثیر مشاهده تنها بر کشف نقاط قوت و ضعف عملکرد فراگیر و در نتیجه بر کیفیت بازخورد ارائه‌شده و افزایش تأثیر آموزشی آزمون است.

1. Hompton et al.
2. Peterson et al.
3. Newble
4. Price

وس و جالی ۲۰۰۱

این مطالعه به این سؤال پاسخ می‌دهد که آیا مشاهده فرایند مصاحبه در مورد بالینی کامل ارزش افزوده‌ای برای این آزمون به ارمغان می‌آورد؟ در این مطالعه ۱۵۵ دانشجوی پزشکی در آزمون «مورد بالینی کامل» با حذف بخش معاینه بالینی شرکت کردند. در یک بخش، توانمندی شرح‌حال‌گیری فراگیران مورد مشاهده قرار گرفت و در بخش دیگر داوطلبان طبق روش سنتی به ارائه شرح‌حال پرداختند. عملکرد دو گروه با چک‌لیست و فرم نمره‌دهی کلی توسط ارزیابان متفاوت مورد ارزیابی قرار گرفت.
نتایج مطالعه نشان داد که نمرات موارد بالینی مشاهده‌شده با موارد مشاهده نشده ارتباط ضعیفی داشتند. به طوری‌که ضریب همبستگی نمرات چک‌لیست مشاهده شده و مشاهده نشده ۰/۳۸ بود و این ضریب برای نمرات فرم‌های نمره‌دهی کلی ۰/۳۳ بود. درحالی‌که نمرات چک‌لیست و فرم نمره‌دهی کلی موارد مشاهده شده با هم ۰/۶۴ و نمرات چک‌لیست و فرم نمره‌دهی کلی موارد مشاهده نشده ۰/۶۱ بود که ارتباط بالایی را نشان می‌دهد. ضریب پایایی بین ارزیابان برای موارد مشاهده بالاتر (چک‌لیست ۰/۷۲ و فرم نمره‌دهی کلی ۰/۷۱) بود. این ضریب برای موارد مشاهده نشده کمتر (چک‌لیست ۰/۳۸ و فرم نمره‌دهی کلی ۰/۶) بود. در این مطالعه ارتباط بین آزمون مورد بالینی و OSCE که این دانشجویان در آن شرکت کرده بودند نیز بررسی شد. ارتباط نمرات فرم نمره‌دهی کلی موارد بالینی مشاهده نشده با OSCE، ۰/۳۶ بود و افزودن نمرات فرم نمره‌دهی کلی موارد بالینی مشاهده به آن واریانس را به ۰/۵۰ رساند. به نظر می‌رسد بر اساس نتایج این مطالعه:

- نمرات مصاحبه مشاهده شده ارتباط ضعیفی با نمرات مصاحبه مشاهده نشده داشت.
- نمرات مستقل از هم دو ارزیاب، توافق بالایی با یکدیگر -به خصوص برای مصاحبه‌های مشاهده شده- داشتند. پس احتمال دارد که یک ارزیاب برای این موارد کافی باشد.
- نمرات مصاحبه مشاهده شده و مشاهده نشده به طور مشخص و مجزا با توانمندی‌های بالینی ارتباط داشت (بر اساس قضاوت با نمرات OSCE).
- مشاهده شرح‌حال‌گیری در مورد بالینی کامل بخش مشخصی بیش از ارائه به تنهایی به روایی آزمون می‌افزاید.
- مورد بالینی مشاهده شده و مورد بالینی محدود به ارائه و عدم مشاهده، هر کدام پارامترهای متفاوتی از توانمندی بالینی را می‌سنجند.
- نتایج به دست آمده توسط دانشجویان (میانگین نمرات) در «مورد بالینی کامل» مشاهده شده کمی بدتر از ارائه به تنهایی بود.
- دامنه و توزیع نمرات برای هر دو گروه تفاوت مشخصی نداشت.

کدام روش ارزیابی ارجح است؟ OSCE یا «مورد بالینی کامل»؟

مهم‌ترین نقطه قوت مورد بالینی کامل این است که به صورت جامع به معاینه بیمار می‌پردازد در حالی که OSCE با یک رویکرد تقلیل‌گرایی، مهارت‌های بالینی را به اجزاء آن تقسیم می‌کند و آنها را به صورت مجزا می‌سنجد. اما سؤال این است که آیا مجموع این اجزا می‌تواند نشان‌دهنده کل، یعنی برخورد داوطلب با یک مورد بالینی باشد؟ می‌توان مثال دانشجوی رشته موسیقی را برای مقایسه در این‌جا مطرح کرد. آیا ارزیابی توانایی دانشجو در نواختن تک‌تک نت‌ها می‌تواند جایگزین ارزیابی نواختن یک قطعه موسیقی به طور کامل شود؟

در مقابل، OSCE به دلیل دارا بودن ایستگاه‌های متعدد از پایایی و تعمیم‌پذیری بالاتری در مقابل «مورد بالینی کامل» برخوردار است. در واقع، گرایش دانشکده‌های پزشکی به استفاده از OSCE در ارزیابی فراگیران، به دلیل پایایی پایین «مورد بالینی کامل» به وجود آمد. استفاده از ایستگاه‌های متعدد در OSCE، آزمونی با پایایی بالاتر تولید کرد اما این مسأله در عوض قربانی کردن روایی آزمون در مقابل پایایی آن تمام شد. به بیان دیگر ساده کردن توانمندی‌های مورد سنجش به بخش‌های کوچک‌تر در OSCE به منزله از بین رفتن عمق ارزیابی به منظور کسب وسعت آن بود.

چالش بین روایی و پایایی در این دو آزمون این سؤال را مطرح می‌کند که در صورت مهیا نمودن بیمار کافی و اختصاص زمان مساوی، پایایی کدام یک بالاتر است؟ «مورد بالینی کامل» یا OSCE؟ برای پاسخ به این سؤال مطالعات کافی وجود ندارد. مطالعات معدودی که در این زمینه انجام شده است نشان می‌دهد اگر زمان یکسان به هر دو روش اختصاص داده شود، پایایی آنها قابل مقایسه است.

- وس و همکاران نشان دادند بدون در نظر گرفتن دیگر عوامل مؤثر بر پایایی، با احتساب زمان مساوی برای هر دو آزمون، پایایی «مورد بالینی کامل» کمی بالاتر از OSCE است. به عنوان مثال با اختصاص دادن زمان ۲۰۰ دقیقه برای اجرای ۱۰ مورد بالینی کامل ۲۰ دقیقه‌ای، پایایی ۰/۸۴ به دست خواهد آمد و اجرای آزمون OSCE با ۳۰ ایستگاه، پایایی ۰/۷۲ خواهد داشت. نتایج مربوط به پایایی دو آزمون در زمان‌های مختلف در جدول ۸ مقایسه شده است (وس و همکاران ۲۰۰۱).
- در مطالعه حمدی و همکاران، با اختصاص زمان ۱۸۰ دقیقه به آزمون DOCEE، ضریب پایایی ۰/۸۵ به دست آمد و

با اختصاص زمان ۱۳۰ دقیقه به OSCE، پایایی ۰/۷۸.

- نورمن[1] (۲۰۰۲) نشان داد اگر فرایند «مورد بالینی کامل»، مشاهده شود و تعداد موارد آن افزایش یابد، پایایی آن کمی بالاتر از OSCE است.

ترکیب «مورد بالینی کامل» و «مورد بالینی کوتاه» چه تأثیری بر پایایی این آزمون‌ها می‌گذارد؟

در برخی از سیستم‌های ارزیابی به منظور افزایش تعداد مواجهه فراگیران با موارد بیماری، از «مورد بالینی کوتاه» به همراه «مورد بالینی کامل» استفاده می‌شود. به عنوان مثال در سیستم ارزیابی دستیاری در کشور استرالیا و نیوزلند از دو «مورد بالینی کامل» و چهار «مورد بالینی کوتاه» برای ارزیابی سالانه دستیاران استفاده می‌شود. توضیح این سیستم با جزئیات در قسمت‌های پیشین این فصل آمده است.

ویلکینسون و همکاران در مطالعه‌ای که روی نتایج این آزمون‌ها انجام دادند دریافتند که ترکیب دو مورد بالینی کامل و چهار مورد بالینی کوتاه با ارزیابی هر مواجهه توسط دو ارزیاب، پایایی معادل ۰/۷۱ به دست می‌دهد. در صورتی‌که تعداد ارزیابان به سه نفر افزایش یابد پایایی با افزایشی جزئی به ۰/۷۳ می‌رسد. پژوهشگران در این مطالعه نشان دادند که افزایش تعداد موارد باعث افزایش پایایی «مورد بالینی کامل» می‌شود اما افزودن «مورد بالینی کوتاه» به «مورد بالینی کامل» برتری بر افزایش تعداد موارد «مورد بالینی کامل» ندارد. به این دلیل که با یکسان در نظر گرفتن عامل زمان برای هر دو، پایایی هر دو روش تقریباً یکسان است. محققان این مطالعه نتیجه‌گیری کردند مهمترین عامل در افزایش پایایی آزمون «مورد بالینی کامل»، زمان آزمون است، به طوری که افزایش موارد، افزودن «مورد بالینی کوتاه» و افزودن چند ایستگاه OSCE به آزمون «مورد بالینی کامل»، بدون افزایش زمان آزمون، تغییری در پایایی آزمون ایجاد نمی‌کند.

آیا می‌توان «مورد بالینی کامل» را به عنوان ارزیابی تراکمی استفاده کرد؟

استفاده از «مورد بالینی کامل» سنتی به عنوان ارزیابی تراکمی به دلیل پایایی پائین آن محدود است. از لحاظ نظری به نظر می‌رسد پایایی مناسب برای یک آزمون «مورد بالینی کامل» حساس و مهم و تراکمی با فراهم نمودن تعداد مناسب بیمار و ارزیاب، تخصیص زمان کافی و همراه نمودن نتایج دیگر آزمون‌های مبتنی بر محل کار فراهم می‌شود.

ویلکینسون و همکاران ۲۰۰۸

در این مطالعه نتایج آزمون‌های «مورد بالینی کامل» و «مورد بالینی کوتاه» به عنوان بخشی از ارزیابی سالانه دستیاران طب اطفال و داخلی در کشور استرالیا و نیوزلند در دو سال متوالی (۲۰۰۵ و ۲۰۰۶) به صورت جامع تحلیل شد. در مجموع ۸۱۶۰ نمره بررسی شد. تعداد ارزیابان ۵۱۸ و تعداد موارد بالینی کامل ۷۷۳ مورد بود. با استفاده از نظریه تعمیم‌پذیری تحلیل مؤلفه‌های مؤثر بر واریانس نمرات (شامل توانایی داوطلب، دشواری موارد و تفاوت بین ارزیابان) انجام شد. بر اساس نتایج این مطالعه:

- سهم واریانس مربوط به توانایی داوطلب در «مورد بالینی کامل»، ۳۳ و ۳۸ درصد (به ترتیب سال ۲۰۰۵ و ۲۰۰۶) بود. این مقدار برای «مورد بالینی کوتاه» به ترتیب ۹ و ۱۵ درصد بود.
- واریانس مربوط به موارد برای هر دو آزمون بسیار کم بود.
- واریانس مربوط به ارزیاب برای هر دو آزمون صفر بود.
- واریانس مربوط به «داوطلب × مورد» بیماری برای «مورد بالینی کامل»، ۲۹ و ۳۷ درصد (به ترتیب سال ۲۰۰۵ و ۲۰۰۶) و برای «مورد بالینی کوتاه» به ترتیب ۵۶ و ۵۸ درصد بود.
- واریانس مربوط به «داوطلب × ارزیاب» برای «مورد بالینی کامل»، ۲۶ و ۱۰ درصد (به ترتیب سال ۲۰۰۵ و ۲۰۰۶) و برای «مورد بالینی کوتاه» به ترتیب ۱۱ و ۱۰ درصد بود.
- ضریب پایایی یک «مورد بالینی کامل» ۰/۳۸ بود و بر اساس مطالعه تصمیم‌گیری انجام شده، چهار «مورد بالینی کامل» لازم بود تا ضریب پایایی به ۰/۷۵ برسد.

1. Norman

• ضریب پایایی یک «مورد بالینی کوتاه» کمتر از «مورد بالینی کامل» بود اما با احتساب زمان صرف‌شده، پایایی سه یا چهار «مورد بالینی کوتاه» با «مورد بالینی کامل» قابل مقایسه بود.
• ترکیب دو «مورد بالینی کامل» و چهار «مورد بالینی کوتاه» با ارزیابی هر مواجهه توسط دو ارزیاب پایایی ۰/۷۱ به دست داد. در صورتی‌که تعداد ارزیابان به سه نفر افزایش یابد، پایایی با افزایشی جزئی به ۰/۷۳ می‌رسد.
این نتایج نشان می‌دهد که واریانس مربوط به دشواری موارد بیماری بسیار کم است و برخی از داوطلبان از توانایی بیشتری در برخورد با برخی از حیطه‌ها یا سیستمها برخوردار هستند. واریانس ناچیز بین ارزیابان در حد ممکن است به دلیل استفاده از سیستم نمره‌دهی ساختارمند و آموزش ارزیابان یا توافق ارزیابان قبل از ثبت نهایی نمره داوطلب باشد.

منابع

1. Abouna GM, Hamdy H. The integrated direct observation clinical encounter examination (IDOCEE) – an objective assessment of students' clinical competence in a problem-based learning curriculum. Med Teach. 1999;21(1):67–72.
2. Amin Z, Seng CY, Eng KH. Practical guide to medical student assessment. World Scientific Publishing; 2006.
3. Chierakul N, Danchaivijitr S, Kontee P, Naruman C. Reliability and Validity of Long Case and Short Case in Internal Medicine Board Certification Examination. J Med Assoc Thai. 2010;93(4):424-8
4. Daelmans HEM, van der Hem-Stokroos HH, Hoogenboom RJI, Scherpbier AJJA, Stehouwer CDA, van der Vleuten CPM. Feasibility and reliability of an in-training assessment programme in an undergraduate clerkship. Med Educ. 2004;38:1270–7.
5. Gleeson F. AMEE Medical Education Guide No. 9. Assessment of clinical competence using the objective structured long examination record (OSLER). Med Teach. 1997;19(1):7–14.
6. Hamdy H, Prasad K, Williams R, Salih FA. Reliability and validity of the direct observation clinical encounter examination (DOCEE). Med Educ. 2003;37(3):205–12.
7. Hardy KH, Demos LL, McNeil JJ. Undergraduate surgical examinations: an appraisal of the clinical orals. Med Educ. 1998;32(6):582–9.
8. Holmboe ES. Faculty and the Observation of Trainees' Clinical Skills: Problems and Opportunities. Acad Med. 2004;79(1):16–22.
9. Holmboe ES, Hawkins RE. Practical guide to the evaluation of clinical competence. Philadelphia: Mosby/Elsevier; 2008.
10. Norcini JJ. The death of the long case? BMJ. 2002;324(7334):408–9.
11. Norcini JJ. The validity of the long case. Med Educ. 2001;35(8):735–6.
12. Olson LG, Coughlan J, Rolfe I, Hensley MJ. The effect of a structured question grid on the validity and perceived fairness of a medical long case assessment. Med Educ. 2000;34(1):46–52.
13. Pavlakis N, Laurent R. Role of the observed long case in postgraduate medical training. Intern Med J. 2001;31(9):523–8.
14. Ponnamperuma GG, Karunathilake IM, McAleer S, Davis MH. The Long Case and Its Modifications: A Literature Review. Med Educ. 2009;43(10):936–41.
15. Swanwick T. Uderstanding Medical Education: Evidence, Theory and Practice. West Sussex: John Wiely & Sons; 2010.
16. Teoh NC, Bowden FJ. The case for resurrecting the long case. BMJ. 2008;336(7655):1250.
17. van der Vleuten C. Making the best of the long case. Lancet. 1996;347(9003):704–5.
18. Wass V, Jolly B. Does observation add to the validity of the long case? Med Educ. 2001;35(8):729–34.

19. Wass V, Jones R, Van D, V. Standardized or real patients to test clinical competence? The long case revisited. Med Educ. 2001;35(4):321–5.
20. Wass V, van der Vleuten C. The long case. Med Educ. 2004;38(11):1176–80.
21. Wilkinson TJ, Campbell PJ, Judd SJ. Reliability of the long case. Med Educ. 2008;42(9):887–93.
22. Zeller A, Battegay M, Gyr N, Battegay E. Evaluation of unstructured medical school examinations: prospective observational study. Swiss Med Wkly. 2003;133(11-12):184–7.

فصل ۲۱

آزمون min–CEX

ساختار آزمون mini–CEX

همان‌طور که در فصل اول این بخش گفته شد، بورد طب داخلی آمریکا در سال ۱۹۷۲ و در پاسخ به محدودیت‌های آزمون‌های شفاهی، آزمون CEX[1] را برای ارزیابی دستیاران، به ویژه دستیاران سال اول، معرفی کرد. در آزمون CEX که در کنار بستر بیمار برگزار می‌شد و حداکثر مدت زمان آن دو ساعت بود، آزمون‌گر با مشاهده تعامل دستیار با یک بیمار، عملکرد وی را در حال اخذ شرح‌حال، معاینه، تشخیص و درمان ارزیابی میکرد. تقریباً مشابه آنچه در آزمون «مورد بالینی کامل» همراه با مشاهده صورت می‌گرفت. چیزی نگذشت که محدودیت‌های این آزمون نیز شناسایی شد و CEX به دلیل آنکه قضاوت در مورد عملکرد دستیاران را بر اساس نظر فقط یک آزمون‌گر و در تعامل با یک بیمار خاص انجام می‌داد، به عدم تعمیم‌پذیری به موقعیت‌های دیگر متهم شد و از روش‌های ارزیابی کنار گذاشته شد. به منظور پاسخ به این مشکلات، بورد داخلی آمریکا در سال ۱۹۹۵ روش ارزیابی mini-CEX[2] را به عنوان یکی از انواع ارزیابی‌های مبتنی بر محل کار معرفی نمود. این آزمون نیز از زمان معرفی تاکنون دستخوش تغییراتی شده است تا با ماهیت تخصص‌های بالینی مختلف و سطوح مختلف توانمندی مورد انتظار از فراگیران تطابق یابد.

آزمون CEX

CEX یا tCEX (traditional CEX) به طور روتین برای سال‌ها توسط بورد طب داخلی آمریکا برای ارزیابی عملکرد دستیاران این رشته انجام می‌شد و هنوز هم کماکان در برخی از رشته‌ها مانند روانپزشکی، طب اورژانس و دیگر رشته‌ها انجام می‌شود. این آزمون توسط یک پزشک باتجربه انجام می‌شود که تمام مراحل عملکرد دستیار را در برخورد با یک بیمار ناشناس برای دستیار (شامل مصاحبه با بیمار، انجام معاینه فیزیکی کامل، ارائه یافته‌ها و مدیریت بیمار) ارزیابی می‌کند. پس از مشاهده، ارزیاب به دستیار بازخورد می‌دهد و اطلاعات مربوط را در فرم از پیش تهیه‌شده توسط بورد تخصصی مربوط، وارد می‌نماید. بعداً، دستیار گزارش مکتوبی از پیگیری بیمار به ارزیاب تحویل می‌دهد. CEX به طور معمول دو ساعت به طول می‌انجامد. در بررسی انجام شده توسط بورد داخلی آمریکا تقریباً، ۸۲ درصد دستیاران داخلی یک CEX در طول سال اول دوره داشتند و تعداد بسیار کمتری بیش از دو مواجهه داشتند (نورسینی و همکاران ۱۹۹۵).

محدودیت‌های CEX شامل سه مورد زیر است:

- مشاهده مواجهه با یک بیمار و در نتیجه عدم تعمیم‌پذیری
- مشاهده عملکرد توسط یک ارزیاب
- تأکید بیش از حد بر کامل بودن مواجهه و مصنوعی شدن آن

در CEX تأکید بر کامل بودن مواجهه دستیار با بیمار، بدون محدودیت زمانی است. درحالی‌که در طبابت واقعی، دستیاران به صورت متمرکزتر با بیمار برخورد می‌کنند؛ با طیف وسیعی از بیماران، محیط و وظایف پیچیده روبرو هستند و توانایی دستیار در اولویت‌بندی تشخیص و درمان بیمار با درنظر گرفتن شرایط واقعی اهمیت بسیار دارد.

1. Clinical Evaluation Exercise (CEX)
2. Mini-Clinical Evaluation Exercise (Mini-CEX)

در mini-CEX یکی از اعضای هیأت علمی، عملکرد فراگیر را در مواجهه با بیمار مشاهده می‌کند و سپس با استفاده از مقیاس درجه‌بندی به هر کدام از توانمندیهای فراگیر در فرمی که به همین منظور تهیه شده است، نمره می‌دهد. در عین حال، هدف اصلی در این روش ارائه بازخورد بر اساس عملکرد مشاهده شده است. تفاوت این روش با CEX در این است که تنها قسمتی از مواجهه فراگیر با بیمار مشاهده و ارزیابی می‌شود نه لزوماً تمام مراحل شرح‌حال و معاینه تا تشخیص و درمان. بنابراین زمان mini-CEX کوتاه‌تر است به طوری که هر مواجهه با بیمار ۱۵ تا ۲۰ دقیقه طول می‌کشد و ۵ تا ۱۰ دقیقه نیز به ارائه بازخورد به فراگیر اختصاص می‌یابد. البته معمولاً، مدت زمان مواجهه برای فراگیران مقطع پزشکی عمومی طولانی‌تر است و ۳۰ تا ۴۵ دقیقه طول می‌کشد.

مواجهه‌ها می‌توانند در انواع مختلفی از موقعیت‌های بالینی شامل درمانگاه، اورژانس و بیماران بستری در بخش اتفاق افتند و طیف وسیعی از مشکلات بیماران را پوشش دهند. برای انتخاب بیمار می‌توان بیمارانی را که برای اولین بار مراجعه کرده‌اند، انتخاب نمود و یا از بیماران تحت پیگیری استفاده کرد.

توانمندی‌هایی که به طور معمول در mini-CEX مورد سنجش قرار می‌گیرند عبارت هستند از: مصاحبه، معاینه بالینی، تعهد حرفه‌ای، قضاوت بالینی، مشاوره، سازماندهی و کارآیی و توانمندی کلی. در این روش ارزیابی، می‌توان طیف وسیعی از مشکلات بالینی شامل ارائه شکایات (مانند درد قفسه سینه، تنگی نفس، درد شکم، سرفه و سرگیجه) و مشکلات بالینی (مانند آرتریت، بیماری‌های مزمن راه‌های هوایی، آنژین، افزایش فشارخون و دیابت) را مورد ارزیابی قرار داد. معمولاً فراگیر یک فعالیت بالینی متمرکز شامل اخذ شرح‌حال، معاینه فیزیکی و ... انجام می‌دهد و خلاصه‌ای را ارائه می‌دهد.

نکته اینجاست که ارزیابان نمی‌توانند در یک مواجهه واحد، تمام جنبه‌های مواجهه فراگیر با بیمار را ارزیابی کنند. این مسأله به این معنی است که در یک مواجهه ممکن است تمرکز بر مهارت‌های ارتباطی و مصاحبه باشد و در مواجهه دیگر، بر نحوه قضاوت بالینی و مراقبت از بیمار تمرکز شود. در نتیجه لازم است چندین مواجهه برای ارزیابی یک فرد در نظر گرفته شود. از فراگیر خواسته می‌شود به جای این‌که یک معاینه و شرح‌حال کامل انجام دهد، یک معاینه متمرکز انجام دهد؛ مثلاً تمایل بیمار برای خودکشی را ارزیابی کند. مهم این است که هر مواجهه برای پوشش برنامه درسی آن مقطع باشد و با برنامه‌ریزی انتخاب شود. در انتها پس از مشاهده عملکرد فراگیر با استفاده از مقیاس درجه‌بندی به هر کدام از توانمندیها نمره داده می‌شود.

در mini-CEX انتظار می‌رود عملکرد فراگیر در طول یک سال، در چند مواجهه و با استفاده از ارزیابان متفاوت ارزیابی شود. در واقع این روش امکان ارزیابی عملکرد فراگیر را در چندین مواجهه، در محیط‌های بالینی مختلف، توسط چندین ارزیاب و در برخورد با طیف وسیعی از مشکلات بالینی فراهم می‌کند. به عنوان مثال، در دوره پیش‌دستیاری در کشور انگلیس شش تا هشت مواجهه در طول یک سال برای ارزیابی استفاده می‌شود، هر مواجهه توسط یک ارزیاب متفاوت ارزیابی می‌شود و هر بار مهارتی که قبلاً مورد آزمون قرار نگرفته است، ارزیابی می‌شود.

آزمون mini-CEX در مقابل آزمون CEX

- متمرکز بر توانایی دستیاران برای حل مشکل بیمار
- ارزیابی در طیف وسیع‌تری از موقعیت‌های بالینی
- پایایی بالاتر نمرات و تعمیم‌پذیری بهتر
- فرصت‌های بیشتری برای مشاهده و ارائه بازخورد با بیش از یک عضو هیأت علمی و با بیش از یک بیمار
- احتمال اجرای مشکل‌تر به دلیل برنامه زمان‌بندی برای مواجهه‌های متعدد
- ممانعت از مشاهده مواجهه کامل دستیاران با یک بیمار با تأکید بیش از حد بر اجرای mini-CEX

انواع آزمون mini-CEX

به دلیل ساختار جالب آزمون mini-CEX انواعی از این آزمون به منظور سنجش توانمندی‌های خاص طراحی شده‌اند. در زیر به طور مختصر به دو مورد از انواع خاص آزمون mini-CEX اشاره می‌شود:

- Professionalism mini-CEX (P-MEX): این آزمون جهت ارزیابی توانمندی تعهد حرفه‌ای ابداع شد و در موقعیت‌های مختلف مواجهه با بیمار مانند بخش، درمانگاه، اورژانس، راند بالینی و جلسات گروه کوچک مورد استفاده قرار گرفت. حیطه‌های مورد ارزیابی شامل «مهارت‌های ارتباط پزشک و بیمار»، «مهارت‌های بازاندیشی»، «مدیریت زمان» و «مهارت‌های ارتباطی بین‌فردی» بودند. بررسی‌ها نشان داده است ۱۰ تا ۱۲ مواجهه برای رسیدن به ضریب تکرارپذیری ۰/۸ مورد نیاز است که با درنظر گرفتن هدف آزمون، هشت مواجهه برای آزمون‌های معیارمحور با خطای معیار اندازه‌گیری ۰/۱۱ کافی است (کروز و همکاران[1] ۲۰۰۶). همچنین این آزمون از روایی محتوا و سازه خوبی برخوردار است (کروز و همکاران ۲۰۰۶). به نظر می‌رسد آزمون P-MEX در ارتقای بازاندیشی فراگیران، تعیین و نشان دادن اهمیت مهارت‌های تعهد حرفه‌ای در طبابت مفید باشد.
- Palliative Care Clinical Evaluation Exercise: این آزمون برای اولین بار در دانشکده پزشکی دانشگاه پیتزبورگ[2] و با هدف ارزیابی مهارت‌های ارتباطی دستیاران داخلی در برخورد با بیماران در مراحل انتهای زندگی طراحی شد. در این آزمون، اعضای هیأت علمی مواجهه دستیاران با این دسته از بیماران و خانواده آنها را مورد مشاهده مستقیم قرار داده و با استفاده از فرم‌های ساختارمند عملکرد آنها را نمره‌دهی می‌کنند و در نهایت بازخورد ارائه می‌دهند. حیطه‌های مورد ارزیابی در این نوع از آزمون شامل «ارائه خبر بد» و «وضعیت نیاز به کد» است. با توجه به ماهیت توانمندی‌های مورد ارزیابی قبل از اجرای آزمون، اعضای هیأت علمی در خصوص شرایط، اهداف مواجهه و راهبردهای مورد نیاز مشاوره دریافت کردند. این آزمون از قابلیت اجرای نسبتاً خوب و تأثیر آموزشی مناسب برخوردار است. البته استفاده از آن منوط به فراهم بودن منابع از جمله هیأت علمی آگاه به محتوای مورد ارزیابی است. در واقع، Palliative CEX Care از این نظر که یک مواجهه را مورد ارزیابی قرار می‌دهد شبیه CEX و از این نظر که یک مواجهه کوتاه و متمرکز را مورد ارزیابی قرار می‌دهد، شبیه mini-CEX است (هن و همکاران[3] ۲۰۰۵).

نمونه‌های بالا به اندازه خود آزمون mini-CEX رایج نیستند و تاکنون در برنامه‌های آموزشی محدودی مورد استفاده قرار گرفته‌اند. به همین دلیل مطالعات بیشتری به منظور مشخص شدن سودمندی آنها قبل از به کارگیری وسیع لازم است. ذکر این موارد صرفاً به منظور نشان دادن قابلیت انعطاف‌پذیری آزمون mini-CEX و تطابق آن با شرایط و نیازهای مختلف است. این موضوع تا حدودی در مورد آزمون CEX نیز صدق می‌کند. به عنوان مثال بورد چشم‌پزشکی آمریکا در سال ۲۰۰۴ نوع خاصی از این آزمون تحت عنوان Ophtalmic mini-CEX (OCEX) را به منظور ارزیابی توانمندی‌های دستیاران چشم‌پزشکی طراحی کرد.

به جز موارد فوق، ابزارهای دیگری نیز وجود دارند که بر اساس مشاهده مستقیم عملکرد بالینی داوطلبان طراحی شده‌اند اما با mini-CEX تفاوت‌هایی دارند. در اینجا سه مورد از این روش‌ها را که نسبت به بقیه بیشتر مورد استفاده قرار گرفته‌اند، به اختصار توضیح می‌دهیم.

1. Cruess et al.
2. Pittsburgh
3. Han et al.

آزمون Clinical Work Sampling (CWS)

این روش ارزیابی در سال ۲۰۰۰ در کانادا و با هدف مشاهده مستقیم عملکرد بالینی ابداع شد. ایده اصلی آن از رویکرد نمونه‌گیری از کار در صنعت گرفته شده است که در آن مشاهده‌گران به صورت مرتب و در فواصل زمانی منظم عملکرد افراد را پایش و مستند می‌کنند. این روش نیازمند جمع‌آوری مجموعه‌ای از اطلاعات در ابعاد مختلف مواجهه با بیمار است. در این روش ارزیابی از چهار سری فرم با مقیاس نمره‌دهی لیکرت پنج‌تایی (از عملکرد نامطلوب تا عالی) استفاده می‌شود (جدول ۱-۲۱):

جدول ۱-۲۱: فرم‌های مورد استفاده در آزمون Clinical Work Sampling (ترن‌بال و همکاران ۲۰۰۰)

ردیف	نوع فرم	ارزیاب	زمان	تعداد موردنظر
۱	فرم نمره‌دهی پذیرش بیمار	عضو هیأت علمی	پذیرش بیمار	یک فرم در ازای پذیرش هر بیمار
۲	فرم نمره‌دهی بخش	عضو هیأت علمی	ترخیص بیمار	یک فرم در ازای هر بیمار تحت‌نظر
۳	فرم نمره‌دهی تیم چند رشته‌ای	عضو تیم پزشکی	ماهانه	یک فرم
۴	فرم نمره‌دهی بیمار	بیمار	ترخیص بیمار	یک فرم در ازای هر بیمار تحت‌نظر

- **فرم نمره‌دهی پذیرش بیمار[1]:** این فرم به منظور ارزیابی مهارت‌های هنگام پذیرش بیمار طراحی شده است که توسط عضو هیأت علمی مسؤول دستیار پس از راند بخش تکمیل می‌شود. از طریق این فرم، عملکرد فراگیران در چهار حوزه شامل مهارت‌های ارتباطی، معاینه فیزیکی، مهارت‌های تشخیصی و مدیریت بیماری و همچنین یک حوزه به صورت توانمندی کلی ارزیابی می‌شود و سپس بازخورد کتبی و شفاهی نیز به دستیار ارائه می‌شود.
- **فرم نمره‌دهی بخش[2]:** این فرم به منظور ارزیابی عملکرد دستیار در مدیریت بیمار بستری طراحی شده است. ابعاد مورد ارزیابی عبارت هستند از: مهارت‌های ارتباطی، معاینه فیزیکی، تشخیص، مشاوره، مدیریت بیماری، رفتار بین‌فردی، یادگیری مداوم، حمایت از سلامت جامعه[3] و توانمندی کلی. این فرم‌ها توسط اعضای هیأت علمی که عملکرد فراگیر را به طور مستقیم در زمان ترخیص بیمار مشاهده کرده‌اند، کامل می‌شود و همچنین بازخورد شفاهی و کتبی ارائه می‌شود. همه این مهارت‌ها تنها در یک موقعیت ارزیابی نمی‌شوند.
- **فرم نمره‌دهی تیم چندرشته‌ای[4]:** این فرم نیز بر اساس مشاهده مستقیم عملکرد کامل می‌شود و عملکرد فراگیر را در شش حیطه زیر مورد ارزیابی قرار می‌دهد: راهبردهای درمانی، مهارت‌های ارتباطی، مشاوره با پرستاران و دیگر متخصصان حرفه پزشکی، مدیریت منابع، برنامه‌ریزی ترخیص بیمار، ارتباط بین‌فردی و نیز توانمندی کلی. داده‌های مربوط به عملکرد دستیار توسط پرستار بخش به صورت روزانه در لاگ‌بوک پرستار ثبت می‌شود. سپس این ارزیابی در یک جلسه که در آن از حرفه‌های مختلف پزشکی شرکت دارند، توسط سرپرستار بخش نهایی می‌شود. در این جلسه افراد در مورد دستیارانی که به اندازه کافی با آن‌ها در تماس بوده‌اند، اظهار نظر می‌کنند. در این فرایند هیچ پزشکی مشارکت ندارد.
- **فرم نمره‌دهی بیمار:** این فرم که به منظور دریافت نظرات بیمار طراحی شده است، شامل هفت سؤال در چهار حیطه مهارت‌های ارتباطی، مهارت‌های همکاری، حمایت از سلامت جامعه، تعهد حرفه‌ای و نیز توانمندی کلی‌است.

1. Admission rating form
2. Ward Rating form
3. health advocacy
4. Multidisciplinary Team Rating Form

این فرم به روش مصاحبه با بیمار و توسط فردی از کارکنان بخش اداری یا پژوهشی بیمارستان تکمیل می‌شود.

در خصوص سودمندی آزمون Clinical Work Sampling باید گفت در صورتی که تعداد کافی موارد بیماری فراهم شود (تقریبا هفت مواجهه برای ضریب پایایی ۰/۷) این ابزار معتبر و پایا است (پلگریم و همکاران ۲۰۱۱). در مطالعه ترن‌بال و همکاران (۲۰۰۰) فرم‌های نمره‌دهی پذیرش بیمار و بخش از پایایی بسیار خوبی برخوردار بودند و چهار تا هشت مورد، پایایی بیش از ۰/۷ را فراهم می‌کرد. در همین مطالعه پایایی فرم نمره‌دهی تیم چندرشته‌ای بسیار پایین گزارش شد. البته این احتمال وجود دارد که این افراد بدون مشاهده عملکرد داوطلب فرم را تکمیل کرده باشند. ثبات درونی فرم‌ها بالا (آلفای کرونباخ ۰/۹) گزارش شد. فرم‌های مختلف با یکدیگر همبستگی داشتند که بیشترین همبستگی در مورد فرم نمره‌دهی پذیرش بیمار و بخش و کمترین همبستگی بین فرم نمره‌دهی بخش و فرم نمره‌دهی تیم چندرشته‌ای بود. روایی پیش‌بین این ابزار نیاز به بررسی دارد.

از نظر قابلیت اجرا در مطالعه ترن‌بال و همکاران (۲۰۰۰) تعداد فرم‌های تکمیل شده به ترتیب برای هریک از فرم‌های آزمون بر اساس ترتیب جدول بالا ۶۴ درصد، ۲۳ درصد، ۴۳ درصد و ۱۲ درصد بود. همان‌طور که نتایج نشان می‌دهد، میزان تکمیل فرم‌های نمره‌دهی بیمار بسیار پایین بود. پژوهشگران علت آن را ترخیص سریع بیمار بدون هماهنگی برای انجام مصاحبه و به خاطر نداشتن دستیار مربوطه بیمار ذکر کردند. فینلی و همکاران[1] (۲۰۰۶) این روش را در ارزیابی دستیاران رادیولوژی به کار بردند. مقبولیت از نظر میزان فرم‌های تکمیل شده کمتر از میزان مورد انتظار بود که با توجه به اختیاری بودن ورود به مطالعه طبیعی به نظر می‌رسد. محققان ذکر کردند این روش برای ارزیابی تراکمی مناسب نیست.

از نظر تاثیر آموزشی می‌توان گفت که چون در زمان ورود داده‌ها، در مورد عملکرد فرد بحث صورت می‌گیرد، در نتیجه تأثیر آموزشی مناسبی دارد.

آزمون Clinical Encounter Cards (CEC)

این ابزار در سال ۱۹۹۹ در دانشگاه مک مستر[2] کانادا ابداع شد و بسیار شبیه mini-CEX است. هدف اصلی آن ارزیابی عملکرد فراگیر بر اساس مشاهده مستقیم مواجهه با بیمار است. نمره‌دهی بر اساس عملکرد مشاهده شده در ابعاد شرح‌حال‌گیری، معاینه فیزیکی، رفتار حرفه‌ای، مهارت‌های تکنیکی، معرفی بیمار، تشخیص و حل مسأله (درمان) است. هر بعد از عملکرد بر اساس مقیاس لیکرت شش‌تایی (نامطلوب، پایین‌تر از مطلوب، در حد مطلوب، بالاتر از مطلوب، عالی و در حد فارغ‌التحصیل پزشکی) نمره‌دهی می‌شود.

علاوه بر ارزیابی کیفیت عملکرد، ارزیابان روی یک کارت چهار در شش اینچ بازخورد به عملکرد فراگیر را ثبت می‌کنند. مطالعات نشان داده‌اند در صورت تعداد کافی مواجهه، این روش در ارزیابی مهارت‌های بالینی معتبر، پایا و قابل اجرا است (تقریباً هشت مواجهه برای دستیابی به ضریب پایایی مساوی یا بیشتر از ۰/۸). علاوه بر این، معرفی این روش ارزیابی به دلیل ارائه بازخورد رضایت دانشجویان را به دنبال داشته است. این روش ارتباط مثبت معنی‌داری با آزمون بورد ملی ارزیابان پزشکی[3]، نمرات پایانی[4] و آزمون نمره‌دهی عملکرد بالینی[5] داشت. جالب آن‌که هیچ ارتباطی بین نتایج این آزمون با آزمون OSCE یافت نشد.

آزمون Blinded Patient Encounter (BPE)

اصول این روش نیز شبیه دیگر روش‌های مشاهده مستقیم است. وجه مشخصه آن این است که بخشی از جلسه

1. Finlay et al.
2. McMaster
3. National Board of Medical Examiners (NBME)
4. Final grades
5. Clinical performance ratings

آموزش بر بالین در دوره پزشکی عمومی را تشکیل می‌دهد. دانشجویان در گروه‌های چهار تا پنج نفره در جلسات آموزش بر بالین شرکت می‌کنند. ارزیابی با مشاهده مستقیم عملکرد یکی از دانشجویان در مورد مصاحبه یا معاینه بالینی متمرکز بیمار شروع می‌شود. این موارد توسط متخصص مسؤول جلسه، آموزش داده شده است. سپس انتظار می‌رود دانشجو تشخیص افتراقی‌ها را بر اساس یافته‌های بالینی مطرح کند. بیمار از قبل برای دانشجو شناخته‌شده نیست و به این دلیل اصطلاح «کور شده»[1] در عنوان این روش ارزیابی به کار می‌رود. بعد از معرفی بیمار، جلسه با تمرکز بر مهمترین نکات بالینی، روش‌های بررسی و تشخیصی مناسب و درمان‌های مناسب بیمار ادامه می‌یابد و در انتها در یک جلسه خصوصی به دانشجو در مورد عملکردش بازخورد داده می‌شود.

سیستم نمره‌دهی در این آزمون یک فرم با لیکرت نه‌تایی است که نمره ۱ تا ۳ برای عملکرد ضعیف، ۴ تا ۶ برای عملکرد مناسب و ۷ تا ۹ برای عملکرد خوب در نظر گرفته شده است. عملکرد دانشجو در مهارت‌های مصاحبه، معاینه فیزیکی و استدلال بالینی ارزیابی می‌شود و بازخورد هم داده می‌شود. همچنین در فرم فضایی برای نوشتن پیشنهادها در نظر گرفته شده است. دانشجویان برگه نمره را فقط به منظور بازخورد ارائه‌شده حفظ می‌کنند.

مزایا و محدودیت‌های آزمون mini-CEX

مزایای آزمون mini-CEX

- امکان مشاهده مستقیم عملکرد فراگیران
- امکان ارزیابی جامع از عملکرد فراگیران: مواجهه‌های متعدد، ارزیابان متعدد، موقعیت‌های بالینی متعدد، قضاوت‌های متعدد و ارزیابی در طول زمان
- مواجهه متمرکز با بیمار، شبیه شرایط واقعی
- استفاده آسان و قابلیت کاربرد مناسب
- امکان تطبیق آن با توجه به شرایط و نیازها
- تأثیر آموزشی بالا
- بهبود کمیت و کیفیت بازخورد

محدودیت‌های آزمون mini-CEX

- جدید بودن روش و ناآشنایی اعضای هیأت علمی با آن و در نتیجه لزوم آموزش اعضای هیأت علمی و ارزیابان به منظور بهبود ویژگی‌های روانسنجی آزمون
- مورد تردید بودن پایایی و روایی آن برای آزمون‌های حساس و مهم
- عدم امکان ارزیابی تمامی قسمتهای مختلف یک مهارت از طریق یک مواجهه بالینی منفرد

علاوه بر موارد مذکور، فراهم کردن فرصتی برای مشاهده متمرکز عملکرد در محیط بالینی و در زمان ارائه خدمات همیشه کار آسانی نیست. همچنین این احتمال وجود دارد که از منظر فراگیران نقش آموزشی اعضای هیأت علمی به واسطه نقش ارزیابی آنان مورد تهدید واقع شود. در برخی موارد نیز این احتمال وجود دارد که فراگیران از درخواست ارزیابی به دلیل مشخص شدن ضعف‌هایشان امتناع ورزند.

1. Blinded

خلاصه‌ای از ویژگی‌های آزمون mini-CEX

- بیماران واقعی
- شرایط بالینی واقعی
- موقعیت‌های طبابت واقعی
- وظایف بالینی واقعی
- محدودیت‌های محیط طبابت واقعی
- مواجهه‌های متعدد
- ارزیابان متعدد
- موقعیت‌های بالینی متعدد
- قضاوت‌های متعدد
- ارزیابی در زمان‌های مختلف
- مشاهده و ارزیابی عملکرد توسط اعضای هیأت علمی بالینی
- تعامل متمرکز و کوتاه بیمار و پزشک
- ارائه بازخورد سازنده

کاربرد آزمون mini-CEX

آزمون mini-CEX برای اولین بار در آمریکا طراحی شد و هم اکنون در بسیاری از مؤسسات آموزشی سراسر دنیا با اهداف تکوینی و تراکمی استفاده می‌شود. همانند دیگر روش‌های ارزیابی مبتنی بر محل کار، هدف اولیه و ذاتی این آزمون ارائه بازخورد و در نتیجه افزایش یادگیری است. در واقع، mini-CEX هم ابزاری آموزشی و هم ابزاری برای ارزیابی است. در اکثر موارد، در مرحله اجرای آزمایشی یا قبل از اجرای وسیع، ابتدا با هدف تکوینی به کار گرفته می‌شود تا در صورت اجرای موفقیت‌آمیز و حصول اطمینان از روایی و پایایی مناسب، برای مقاصد تراکمی نیز استفاده شود. باید توجه داشته باشیم در صورت استفاده از mini-CEX با هدف تراکمی این روش صرفاً برای آزمون‌های معیار محور مناسب است. (هیل و همکاران ۲۰۰۹). دلیل این امر در پایان این فصل در قسمت سؤالات رایج تشریح خواهد شد.

این آزمون در ابتدا با هدف ارزیابی دستیاران طراحی شد و در برنامه‌های دستیاری مختلف از جمله طب داخلی، جراحی، بیماری‌های زنان و زایمان، طب کودکان و طب اورژانس مورد استفاده قرار گرفت اما به تدریج در مقطع پزشکی عمومی نیز با موفقیت به کار گرفته شد. از سال ۲۰۰۲ به بعد این آزمون در ارزیابی کارآموزان در کانادا و آمریکا و به تدریج در سایر کشورها به کار گرفته شد. به عنوان مثال، دانشگاه ساتمپتون[1] در انگلیس از سال ۲۰۰۴ آزمون mini-CEX را به عنوان ارزیابی پایان بخش‌های بالینی در سال آخر دوره پزشکی عمومی[2] جایگزین مورد بالینی کامل کرده است.

همچنین اشاره شد که علاوه بر سنجش مهارت‌های بالینی معمول، انواعی از آن به صورت اختصاصی برای ارزیابی تعهدحرفه‌ای و مراقبت بیماران در مراحل انتهای زندگی، طراحی شده است و به کار می‌رود.

1. Southampton
2. Bachelor of Medicine (BM)

گام‌های طراحی و اجرای آزمون mini-CEX مطلوب

رعایت یک سری از موارد به طراحی یک آزمون mini-CEX مناسب کمک می‌کند. خلاصه این مراحل در جدول ۲-۲۱ نشان داده شده است.

جدول ۲-۲۱ : خلاصه مراحل طراحی و اجرای آزمون mini-CEX

ردیف	عنوان	توضیح
۱	تهیه بلوپرینت آزمون	ضروری است بلوپرینت آزمون با پوشش برنامه درسی اصلی و موارد بیماری ضروری و مهم، ارزیابی عملکرد داوطلبان در محیط‌های مختلف مانند بخش بیماران بستری، درمانگاه، اورژانس و ... و تعداد مناسب و متنوع موارد بیماری تدوین شود.
۲	تعیین مهارت‌های مورد سنجش	به طور روتین در آزمون mini-CEX شش حیطه اخذ شرح‌حال، معاینه بالینی، تعهد حرفه‌ای، قضاوت بالینی، مشاوره، سازمان‌دهی و یک آیتم هم به عنوان مهارت کلی مورد ارزیابی قرار می‌گیرد.
۳	طراحی فرم‌های ارزیابی	هرچند فرم‌های mini-CEX با توجه به نیاز هر تخصص، مقطع فراگیر و ترجیحات اعضای هیأت علمی با ساختارهای مختلفی طراحی شده‌اند، همه این فرم‌ها کم و بیش دارای سه بخش مشخص شامل بخش مربوط به اطلاعات داوطلب، ارزیاب و مواجهه، بخش مربوط به توانمندی‌های مود ارزیابی و نمره‌دهی عملکرد داوطلب و بخش مربوط به بازخورد و نظرات ارزیاب و داوطلب نسبت به مواجهه است.
۴	تعیین حداقل سطح قابل‌قبول عملکرد در هر حیطه	علاوه بر مهارت‌های مورد سنجش و تعریف آن، معیارهای نمره‌دهی در هر یک از سطوح و حداقل سطح مورد انتظار از داوطلب باید تعریف شوند.
۵	تصمیم‌گیری در مورد تعداد و ویژگی ارزیابان	انتخاب ارزیابان مناسب و تصمیم‌گیری در مورد تعداد و تنوع ارزیابان از دیگر مراحل طراحی آزمون mini-CEX است.
۶	اطلاع‌رسانی، آشناسازی و آموزش ارزیابان و داوطلبان	ضروری است ارزیابان توجیه شوند و آموزشهای لازم را دریافت کنند. همچنین داوطلبان با اهداف، محتوا و نحوه برگزاری آزمون آشنا شوند
۷	اجرای آزمون	لازم است برای اجرای آزمون هماهنگی‌های لازم انجام شود، مورد مناسب بیماری انتخاب شود و تعامل فراگیر با بیمار مورد مشاهده قرار گیرد، بر اساس معیارهای مشخص عملکرد داوطلب مورد ارزیابی قرار گیرد و در انتها به فراگیران در مورد عملکردشان بازخورد داده شود.
۸	بررسی کیفیت آزمون برگزار شده	ضروری است با استفاده از روش‌های آماری و کیفی کیفیت آزمون‌های برگزار شده مورد بررسی قرار گیرد.

تهیه بلوپرینت آزمون

به منظور افزایش کیفیت آزمون mini-CEX ضروری است در تهیه بلوپرینت آن موارد زیر لحاظ گردد:

- برنامه درسی اصلی و موارد بیماری ضروری و مهم، مبنای تدوین بلوپرینت قرار گیرد.
- عملکرد داوطلبان در محیط‌های مختلف مانند بخش بیماران بستری، درمانگاه، اورژانس و ... مورد ارزیابی قرار گیرد.
- تعداد مناسب و متنوع موارد بیماری به منظور پوشش اهداف دوره و همچنین دستیابی به پایایی قابل‌قبول در نظر گرفته شود و برای هر داوطلب، موارد بیماری غیرتکراری انتخاب شوند.

تعیین مهارت‌های مورد سنجش

به طور روتین در آزمون mini-CEX شش حیطه اخذ شرح‌حال، معاینه بالینی، تعهد حرفه‌ای، قضاوت بالینی، مشاوره، سازمان‌دهی (یا کارآیی[1]) و یک آیتم هم به عنوان مهارت کلی مورد ارزیابی قرار می‌گیرد. در برخی موارد از مهارت‌های ارتباطی به جای مشاوره استفاده می‌شود. تعریف هر یک از این مهارت‌ها باید مشخص باشد تا به یکسان‌سازی قضاوت ارزیابان کمک کند. این تعاریف می‌تواند در فرم‌های ارزیابی یا در راهنمای ارزیابان درج شود:

تعریف مهارت‌های مورد ارزیابی در فرم‌های mini-CEX (اقتباس از بورد بیماری‌های داخلی آمریکا)

- مصاحبه پزشکی (اخذ شرح‌حال): اجازه می‌دهد بیمار داستانش را بازگو کند، برای کسب اطلاعات دقیق و کافی از سؤالات/هدایت‌گرها به طور مؤثر استفاده می‌کند، به طور مناسبی به احساسات بیمار و سرنخ‌های غیرکلامی پاسخ می‌دهد.
- معاینه فیزیکی: از یک تناوب منطقی و کارآ پیروی می‌کند، بین اقدامات غربالگری و تشخیصی برای یافتن مشکل تعادل برقرار می‌کند، بیمار را آگاه می‌سازد و نسبت به راحتی بیمار حساس است.
- تعهد حرفه‌ای (مسائل انسانی): احترام، دلسوزی و همدلی نشان می‌دهد، اعتماد بیمار را جلب می‌کند، نسبت به نیازهای بیمار در ارتباط با احساس راحتی، حجب و حیا، رازداری و ارائه اطلاعات پاسخگو است.
- قضاوت بالینی: تست‌های تشخیصی مناسب را با در نظرگرفتن سود و زیان آن درخواست می‌کند.
- مشاوره: دلایل مربوط به انتخاب آزمایشات تشخیصی و درمان را توضیح می‌دهد، رضایت بیمار را کسب می‌کند و در ارتباط با مدیریت بیماری آموزش می‌دهد.
- سازمان‌دهی و کارآیی: اولویت‌بندی می‌کند، زمان را رعایت می‌کند و مفید و مختصر عمل می‌کند.
- مهارت بالینی کلی: قضاوت، سنتز، مراقبت، تأثیر و کارآیی از خود نشان می‌دهد.

در برخی از موارد با توجه به رشته تخصص مدنظر، مهارت‌های مورد سنجش اصلاح شده‌اند. موارد زیر نمونه‌ای از مهارت‌های مورد سنجش در بخش روان‌پزشکی هستند:

- مصاحبه پزشکی و مهارت‌های ارتباطی
- مهارت‌های معاینه
- مهارت‌های تصمیم‌گیری
- رفتارهای فردی و حرفه‌ای
- استفاده از زمان
- توانمندی بالینی کلی

طراحی فرم‌های ارزیابی

هرچند فرم‌های mini-CEX با توجه به نیاز هر تخصص، مقطع فراگیر و ترجیحات اعضای هیأت علمی با ساختارهای مختلفی طراحی شداند، همه این فرم‌ها کم و بیش دارای ویژگی‌های مشترکی هستند. اکثر فرم‌های mini-CEX دارای سه بخش مشخص هستند:

بخش ابتدای فرم به اطلاعات مربوط به داوطلب، ارزیاب و مواجهه اختصاص دارد. ارزیاب به طور معمول برای هر مواجهه بالینی در بخش ابتدایی فرم اطلاعات زیر را ثبت میکند:

- تاریخ انجام آزمون
- میزان دشواری مشکلات بیمار (کم، متوسط، زیاد) و جنس بیمار
- نوع ویزیت پزشک (جدید یا مراجعه مجدد)
- موقعیت (بخش، اورژانس، کلینیک یا بخش مراقبتهای ویژه)

1. Efficacy

- هدف از ویزیت (جمع‌آوری اطلاعات، تشخیص بیماری، درمان یا مشاوره)

بخش میانی فرم به توانمندی‌های مود ارزیابی و نمره‌دهی عملکرد داوطلب اختصاص دارد. همان‌طور که در ابتدای این فصل اشاره شد، در برخی از موارد علاوه بر حیطه‌های مورد ارزیابی، توصیفی از هر حیطه به عنوان مرجعی برای نمره‌دهی درج می‌شود. در اکثر فرم‌ها از نمره‌دهی گلوبال به صورت نمرات ۱ تا ۹ برای نمره‌دهی استفاده می‌شود که معمولاً نمرات ۱ تا ۳ «نامطلوب»، ۴ تا ۶ «مطلوب»، ۷ تا ۹ «عالی» درنظر گرفته می‌شوند. در برخی از فرم‌ها نمره ۴ به صورت ستونی مجزا و به عنوان نمره «مرزی» درنظر گرفته شده است و نمرات ۵ تا ۶ تحت عنوان «مطلوب» ذکر شده است. معمولاً گزینه‌ای نیز به عنوان «مشاهده نشد» برای هر یک از آیتم‌ها درنظر گرفته می‌شود. این سیستم نمره‌دهی که شرح داده شد، مربوط به فرم‌های بورد بیماری‌های داخلی کشور آمریکا است که در بسیاری از دیگر کشورهای جهان نیز استفاده می‌شود. در عین حال، برخی از دانشکده‌ها این فرم را به دلایلی مانند راحتی و پذیرش بیشتر ارزیابان تغییر داده‌اند. در کالج پزشکی مایوکلینیک[1] از فرم‌های اصلاح شده پنج‌تایی استفاده می‌شود (نمره ۱ «نیاز به بهبود»، نمرات ۲ تا ۴ «متوسط» و نمره ۵ به عنوان «۱۰ درصد بالا»). در برنامه پیش‌دستیاری کشور انگلیس نیز از فرم‌های شش‌تایی استفاده می‌شود که در آن نمرات ۱ و ۲ به عنوان «پایین‌تر از حدانتظار»، نمره ۳ «مرزی»، ۴ «درحد انتظار» و نمرات ۵ و ۶ به عنوان «بالاتر از حدانتظار» درنظر گرفته شده است. آیتم «بدون نمره» هم به مواردی که مشاهده نشده است اختصاص دارد.

بخش انتهایی فرم به بازخورد و نظرات ارزیاب و داوطلب نسبت به مواجهه اختصاص دارد و فضایی برای هر یک از موارد زیر تعبیه شده است:

- مدت زمان صرف شده برای مشاهده
- مدت زمان صرف شده برای ارائه بازخورد
- توصیه‌های مشروح و بازخورد بعد از مواجهه
- میزان رضایت ارزیاب از mini-CEX
- میزان رضایت داوطلب از mini-CEX
- دریافت آموزش توسط ارزیاب و نحوه آن

میزان رضایت به صورت مقیاس لیکرت از ۱ تا ۹ (از عدم رضایت کامل تا کاملاً رضایت‌بخش) ثبت می‌شوند. در برخی از موارد فضایی نیز برای بازاندیشی داوطلبان بر تجربه ارزیابی تعبیه شده است. نمونه‌ای از فرم مورد استفاده توسط بورد بیماری‌های داخلی آمریکا برای ارزیابی دستیاران این رشته در شکل ۱-۲۱ آمده است. همچنین فرم مورد استفاده توسط نظام ملی سلامت[2] انگلیس در برنامه پیش‌دستیاری در شکل ۲-۲۱ نشان داده شده است.

1. Mayoclinic
2. National Health Service (NHS)

آزمون mini-CEX

نام ارزیاب: تاریخ: نام فراگیر:

محیط: بخش ☐ سرپایی ☐ اورژانس ☐ سایر موارد ☐

بیمار: سن ☐ جنس: زن/ مرد جدید ☐ پیگیری ☐

پیچیدگی مشکل: کم ☐ متوسط ☐ زیاد ☐

هدف: جمع آوری اطلاعات ☐ تشخیص ☐ درمان ☐ مشاوره ☐

۱. **مهارت شرح‌حال گیری** (☐ مشاهده نشده است)

۱ ۲ ۳ ۴ ۵ ۶ ۷ ۸ ۹

عدم رضایت رضایت بخش عالی

۲. **مهارت معاینه فیزیکی** (☐ مشاهده نشده است)

۱ ۲ ۳ ۴ ۵ ۶ ۷ ۸ ۹

عدم رضایت رضایت بخش عالی

۳. **تعهد حرفه‌ای/ ویژگی های انسانی** (☐ مشاهده نشده است)

۱ ۲ ۳ ۴ ۵ ۶ ۷ ۸ ۹

عدم رضایت رضایت بخش عالی

۴. **قضاوت بالینی** (☐ مشاهده نشده است)

۱ ۲ ۳ ۴ ۵ ۶ ۷ ۸ ۹

عدم رضایت رضایت بخش عالی

۵. **مهارت‌های مشاوره** (☐ مشاهده نشده است)

۱ ۲ ۳ ۴ ۵ ۶ ۷ ۸ ۹

عدم رضایت رضایت بخش عالی

۶. **کارآمدی سازمانی** (☐ مشاهده نشده است)

۱ ۲ ۳ ۴ ۵ ۶ ۷ ۸ ۹

عدم رضایت رضایت بخش عالی

۷. **توانمندی بالینی کلی** (☐ مشاهده نشده است)

۱ ۲ ۳ ۴ ۵ ۶ ۷ ۸ ۹

عدم رضایت رضایت بخش عالی

مدت زمان آزمون: زمان مشاهده ..... دقیقه ارائه بازخورد ..... دقیقه

رضایت ارزیاب از اجرای تمرین: کم ۱ ۲ ۳ ۴ ۵ ۶ ۷ ۸ ۹ زیاد

رضایت فراگیر از اجرای تمرین: کم ۱ ۲ ۳ ۴ ۵ ۶ ۷ ۸ ۹ زیاد

پیشنهادها:

شکل ۱-۲۱: فرم mini-CEX برای ارزیابی دستیاران رشته داخلی، طراحی‌شده توسط بورد بیماری‌های داخلی آمریکا (وب سایت www.abim.org)

Mini-Clinical Evaluation Exercise (CEX) – F_1 Version

لطفا با گذاشتن علامت ضربدر مقابل سؤالات فرم را کامل کنید: ☒ لطفا از قلم مشکی برای تکمیل فرم استفاده کنید.

نام خانوادگی داوطلب									
نام داوطلب									
شماره GMC (دانشجویی)									
	ثبت شماره GMC الزامی است								
محیط بالینی	درمانگاه ☐	بستری ☐		پذیرش اورژانس ☐		پذیرش جراحی ☐			
مشکل بالینی	تنفسی ☐	گردش خون ☐	گوارشی ☐	نورولوژیک ☐	درد ☐	روانپزشکی ☐	دیگر موارد		
بیمار جدید یا پیگیری	جدید ☐	پیگیری ☐	تمرکز مواجهه	شرح‌حال	تشخیص	مدیریت	توضیح		
تعداد دفعات ویزیت قبلی بیمار توسط فراگیر	۰ ☐	۱ تا ۴ ☐	۵ تا ۹ ☐	۱۰ < ☐	پیچیدگی بیماری	کم ☐	متوسط ☐	زیاد ☐	
موقعیت ارزیاب					موارد دیگر				
تعداد دفعات انجام mini-CEX توسط ارزیاب با هر فراگیری	۰ ☐	۱ ☐	۲ ☐	۳ ☐	۴ ☐	۵-۹ ☐	۹< ☐		

لطفا حیطه‌های زیر را با استفاده از نمره‌دهی مقابل نمره‌گذاری کنید	زیر حد انتظار برای F_1	مرزی	در حد انتظار برای F_1	بالاتر از حد انتظار برای F_1	بدون نمره*
۱. مهارت مصاحبه	☐	☐	☐	☐	☐
۲. مهارت معاینه	☐	☐	☐	☐	☐
۳. مهارت ارتباطی	☐	☐	☐	☐	☐
۴. قضاوت بالینی	☐	☐	☐	☐	☐
۵. تعهد حرفه‌ای	☐	☐	☐	☐	☐
۶. سازمان‌دهی	☐	☐	☐	☐	☐
۷. مراقبت از بیمار در کل	☐	☐	☐	☐	☐

* بدون نمره: لطفا این مورد را در صورتی علامت بزنید که مهارت مورد نظر مشاهده نشده است و در نتیجه قادر به نمره‌گذاری نیستید.

موارد مثبت عملکرد	موارد پیشنهادی برای ارتقای عملکرد

اقدامات مورد توافق:

آیا در مورد این روش تاکنون آموزش دیده‌اید؟	چهره به چهره ☐	مطالعه دستورالعمل‌ها ☐	اینترنت/CD ☐
امضای ارزیاب	تاریخ	زمان صرف شده برای مشاهده (دقیقه)	زمان صرف شده برای بازخورد (دقیقه)

نام خانوادگی ارزیاب												
شماره پرسنلی ارزیاب												

شکل ۲-۲۱: فرم mini-CEX طراحی‌شده توسط نظام ملی سلامت انگلیس برای استفاده در برنامه پیش‌دستیاری (برای اطلاعات بیشتر در مورد این فرم و جزئیات توانمندی‌های مورد انتظار به سایت WWW.HCAT.NHS.UK مراجعه کنید).

تعیین حداقل سطح قابل‌قبول عملکرد در هر حیطه

علاوه بر مهارت‌های مورد سنجش و تعریف آن، معیارهای نمره‌دهی در هر یک از سطوح و حداقل سطح مورد انتظار از داوطلب باید تعریف شوند. به عنوان مثال، حداقل عملکرد مورد انتظار از داوطلبی که یکی از نمرات ۷ تا ۹ را دریافت می‌کند و در سطح مطلوب قرار می‌گیرد، چه بوده است؟ داوطلب مرزی در هر یک از مهارت‌های مورد ارزیابی چه عملکردی را نشان می‌دهد؟

مطالعات نشان داده‌اندکه تعاریف ارزیابان از هر یک از این سطوح عملکرد متفاوت است و این مسأله منجر به بروز خطا در نمره‌دهی می‌شود. در مطالعات کیفی انجام شده اعضای هیأت علمی اظهار داشتند که آنها با اطمینان می‌توانستند دستیاران با عملکرد ضعیف را تشخیص دهند اما اعلام ردی برای آنان مشکل بود. به این دلیل که نمی‌دانستند چه نوع رفتاری را باید مستند کنند تا قضاوت ردی دستیاران را حمایت کند. بنابراین تعریف معیارهای نمره‌دهی، آگاه‌سازی ارزیابان از این معیارها و آموزش آن‌ها در این رابطه در ارتقای کیفیت آزمون و بویژه بهبود پایایی بین ارزیابان تأثیر گذار است (الوس دلیما[1] ۲۰۱۰). هر چند، همان‌طور که در صفحات آتی در قسمت مربوط به نمره‌دهی اشاره خواهد شد، عوامل بسیاری بر موفقیت‌آمیز بودن مواجهه داوطلبان و بیمار و در واقع بر نمره آزمون تأثیرگذار است و درنتیجه تعریف دقیق معیارهای نمره‌دهی مشکل است.

نمونه‌ای از توصیف نمره‌بندی mini-CEX (ولر و همکاران ۲۰۰۹)

- عملکرد نامطلوب: استفاده از رویکردهای سؤال برانگیز و گاهی غیر قابل توجیه، راحت نبودن استادان با برخی از تصمیمات و تعاملات داوطلب، وجود دغدغه‌هایی در خصوص ایمنی بیمار، نقص بارز در دانش و مهارت.
- عملکرد مطلوب: قابل اعتماد اما نه چشمگیر، دارای رویکرد مناسب به بیمار و کارکنان، توانمند برای مقطع آموزشی مورد نظر.
- عملکرد عالی: عملکردی برجسته و چشمگیر، مطلع در همه حوزه‌ها، رویکردی عالی به بیمار و کارکنان، بالاتر از حد انتظار برای مقطع آموزشی مورد نظر.

تصمیم‌گیری در مورد تعداد و ویژگی ارزیابان

به نظر می‌رسد یکی از عواملی که بر نمرات mini-CEX تأثیرگذار است، انتخاب ارزیابان مناسب است. به طور معمول، در خصوص ارزیابی دستیاران، ارزیابان از میان اعضای هیأت علمی بالینی دارای تخصص مربوط به همان بخش بالینی انتخاب می‌شوند. در مورد دانشجویان پزشکی عمومی، ارزیابی توسط استادان یا دستیاران تخصصی و فوق‌تخصصی صورت می‌گیرد. مواردی نیز از ارزیابی و ارائه بازخورد توسط هم‌کلاسی‌ها[2] گزارش شده است که مسلماً در این موارد تأکید بر تأثیر آموزشی آزمون بیشتر بوده است. علاوه بر انتخاب ارزیاب، باید در مورد تعداد و تنوع ارزیابان نیز تصمیم‌گیری شود. معمولاً در شرایط واقعی استفاده از بیش از یک ارزیاب برای یک مواجهه امکان‌پذیر نیست اما نباید از ارزیابان یکسان برای مواجهه‌های متعدد یک داوطلب استفاده کرد. به عبارت دیگر توصیه می‌شود همه داوطلبان بین همه ارزیابان در طول یک دوره ارزیابی چرخش داشته باشند.

هرچند متون اطلاعات کمی در مورد تأثیر ویژگی‌های ارزیابان بر نتایج ارزیابی در اختیار می‌گذارند و هیچ یک از مطالعات به طور نظام‌مند به مقایسه عملکرد ارزیابان با تجربه و کم تجربه نپرداخته‌اند، نتایج متنوعی از نحوه ارزیابی و نمره‌دهی ارزیابان ارائه شده است. از جمله آنها می‌توان به تفاوت میزان سخت‌گیری اعضای هیأت علمی با یکدیگر در قضاوت، تفاوت در میزان سخت‌گیری اعضای هیأت علمی با دستیاران و همچنین ذهنی بودن قضاوت یک ارزیاب در خصوص یک داوطلب مشابه در موقعیت‌های گوناگون اشاره کرد. مسائل فوق اهمیت این مسأله را بیشتر نشان می‌دهد که به منظور کاهش خطاهای مربوط به نمره‌دهی ارزیابان، علاوه بر دقت در انتخاب ارزیابان و آموزش آن‌ها، باید ارزیابان با ویژگی‌های مختلف بین داوطلبان توزیع شوند. این موارد در بخش‌های آتی همین فصل با جزئیات بیشتر مود بحث قرار می‌گیرد.

1. Alves de Lima
2. Peer mini-CEX

اطلاع‌رسانی، آشناسازی و آموزش ارزیابان و داوطلبان

قبل از آزمون، ضروری است ارزیابان با فرم‌های ارزیابی و شیوه نمره‌دهی آن آشنا شوند. برای رسیدن به این هدف می‌توان راهنماهایی را برای ارزیابان تدوین نمود. در این راهنماها وظایف ارزیابان و داوطلبان و نحوه تکمیل فرم‌ها و نمره‌دهی آن مشخص شده است. راهنماها منجر به یکسان‌سازی قضاوت ارزیابان و در نتیجه بهبود پایایی بین ارزیابان می‌شود. علاوه بر این، می‌توان از جلسات توجیهی حضوری نیز استفاده کرد. نمونه‌ای از دستورالعمل آزمون mini-CEX تهیه شده برای ارزیابان برنامه پیش‌دستیاری انگلیس در زیر آمده است:

راهنمای ارزیاب در آزمون mini-CEX برنامه پیش‌دستیاری انگلیس (برگرفته از: WWW.HCAT.NHS.UK)

آزمون mini-CEX چیست؟
mini-CEX ارزیابی ساختارمند مواجهه بالینی مشاهده‌شده است و هدف از طراحی این ارزیابی مقطعی، ارائه بازخورد به فراگیر در مورد مهارت‌های ضروری برای مراقبت از بیمار است.
چه فردی به عنوان ارزیاب در این آزمون مناسب است؟
ضروری است ارزیاب در مورد روش ارزیابی و ارائه بازخورد آموزش‌های لازم را دریافت کرده باشد. همچنین خود از توانایی لازم در مواجهه با بیمار انتخاب‌شده برخوردار باشد. این فرد می‌تواند از بین اعضای هیأت علمی متخصص، دستیاران تخصصی و فوق تخصصی، پزشک عمومی و حتی پرستار با تجربه یا افراد مرتبط با حرفه‌های کاربردی بهداشتی باشد. در صورت امکان برای هر مواجهه باید از ارزیاب متفاوت استفاده شود و هر ارزیاب حداقل هر چهار ماه یک آزمون mini-CEX را داوری کند.
mini-CEX چگونه انجام می‌شود؟
فرایند آزمون معمولاً توسط فراگیر در برنامه پیش‌دستیاری هدایت می‌شود. وی مواجهه بالینی را انتخاب می‌کند که باید نمایان‌گر کار روزمره طبابت باشد. هر چند شما به عنوان ارزیاب می‌توانید یک ارزیابی از قبل هماهنگ نشده انجام دهید. فرایند مشاهده معمولاً ۲۰ دقیقه طول می‌کشد و بازخورد فوری در حدود ۵ دقیقه ارائه می‌شود. ممکن است گاهی زمان بیشتری مورد نیاز باشد.
mini-CEX چه توانمندی‌های بالینی را ارزیابی می‌کند؟
آزمون شامل سؤالاتی در هفت حیطه است و در فرم فضایی برای نظر شما تعبیه شده است تا نقاط قوت و نقاط قابل بهبود فراگیر را مشخص کنید. لزوماً همه حیطه‌های مورد سؤال در یک مواجهه ارزیابی نمی‌شود.
استاندارد مرجع برای ارزیابی چیست؟
داوطلب در هر مقطعی باید نسبت به استانداردهایی که به عنوان سطح رضایت بخش در برنامه درسی همان مقطع تعریف شده است، ارزیابی شود.
ارائه بازخورد چگونه باشد؟
به منظور به حداکثر رساندن اثر آموزشی آزمون، ارزیاب باید نقاط قوت و نقاط قابل بهبود فراگیر را مشخص کند و با توجه به آن‌ها برنامه عملیاتی مناسب طراحی شود. این مسأله باید با حساسیت و در محیطی کاملاً مناسب انجام شود.
چگونه می‌توانید به فرم آزمون دسترسی پیدا کنید؟
فراگیران برنامه پیش‌دستیاری می‌توانند برای شما یک پیام الکترونیکی به همراه فرم ارزیابی ارسال نمایند و از شما درخواست ارزیابی کنند یا این که شما می‌توانید در صورتی که سوپروایزر آن‌ها هستید، وارد کارپوشه الکترونیکی آن‌ها شوید و فرم را تهیه کنید. در صورت صلاحدید، فراگیران می‌توانند نتایج آزمون را از جانب شما وارد کارپوشه کنند که در نهایت ایمیلی برای اطلاع شما فرستاده می‌شود.
چگونه می‌توانید فرم را تکمیل کنید؟
آموزش: شما باید ثابت کنید که در مورد روش ارزیابی و بازخورد آموزش دیده‌اید.
محیط بالینی: مناسب‌ترین محیط بالینی را انتخاب کنید.
مشکل بالینی: انتخاب مسأله بالینی بر اساس حیطه‌هایی است که در برنامه درسی آمده است. مانند مشکلات گوارشی یا دستگاه تنفس.
انتخاب بیمار: می‌توانید از بیمار جدید یا تحت پیگیری استفاده کنید.
تمرکز مواجهه بالینی: در این ارتباط مناسب‌ترین را انتخاب کنید به عنوان مثال مهارت‌های تشخیصی باید حتماً جز مهارت‌های بالینی فراگیران دوره پیش‌دستیاری باشد.
پیچیدگی مورد بیماری: دشواری مورد را با توجه به مقطع فراگیر نمره‌دهی کنید.
نمره‌دهی سؤالات: شما باید طیف کامل نمره‌دهی را با توجه به توصیفاتی که برای هر یک ارائه شده است در نظر بگیرید. این که در شروع دوره نمرات پایین باشد چیز دور از انتظاری نیست.
فضای نوشتاری آزاد: برای به حداکثر رساندن استفاده از آزمون، فضایی در پایین سؤالات تعبیه شده است تا آزادانه در مورد نقاط ضعف و نقاط قابل بهبود فراگیر اظهار نظر کنید.
مرتبط ساختن ارزیابی با برنامه ارتقای فردی فراگیران: شما می‌توانید به آسانی mini-CEX را با این برنامه مرتبط سازید.
زمان: زمان صرف شده برای مشاهده و ارائه بازخورد را یادداشت کنید.
مشخصات ارزیاب: شماره نظام پزشکی خود را یادداشت نمایید. مسؤول برنامه آموزشی پیش‌دستیاری، ارزیابی را مرور می‌کند و به صورت تصادفی روایی تعدادی از فرم‌ها بررسی می‌شود.

همان‌گونه که در فصل اول بیان شد، برنامه‌های متنوعی جهت توانمندسازی اعضای هیأت علمی با هدف بهبود مهارت‌های مشاهده و قضاوت این افراد طراحی و اجرا شده است. درنظرگرفتن موارد زیر در آموزش ارزیابان ضروری است:

- تأکید بر ارتقای کیفیت مشاهده ارزیابان و مهارت ارائه بازخورد آنان
- تمرکز بر هر یک از موارد فوق با توجه به هدف آزمون mini-CEX (تکوینی یا تراکمی)
- تأمل در مورد تعریف حیطه‌های توانمندی، رفتارهای معرف هر حیطه که باید مورد مشاهده قرار گیرد و معیارهای نشان‌دهنده حداقل عملکرد مناسب برای هر توانمندی

علاوه بر ارزیابان، به منظور آشناسازی داوطلبان با اهداف، محتوا و نحوه برگزاری آزمون باید بلوپرینت آزمون، فرم‌های نمره‌دهی و راهنماهای برگزاری آزمون در اختیار فراگیران قرار داده شود.

اجرای آزمون

هماهنگی یا درخواست انجام آزمون می‌تواند از طرف داوطلب یا ارزیاب و یا فرد مسؤول امتحانات باشد. در اغلب موارد، برنامه کلی آزمون مشخص است و ارزیابی بر اساس شرایط بالینی پیش‌آمده و در صورت آماده بودن ارزیاب، داوطلب و شرایط محیطی انجام می‌شود.

ارزیابان باید مراقب باشند که مورد بیماری بر اساس بلوپرینت آزمون انتخاب شده باشد و برای هر داوطلب از مواجهه با موارد بیماری تکراری پرهیز کنند. ضروری است تعامل فراگیر با بیمار مورد مشاهده قرار گیرد. هر چند مشاهده بخشی جدانشدنی از آزمون mini-CEX است اما گزارش‌هایی وجود دارد مبنی بر اینکه اعضای هیأت علمی به دلیل تداخل با وظایف درمانی و کمبود وقت، بدون مشاهده عملکرد فراگیران با استناد به بیان شرح‌حال و تشخیص و درمان بیمار از سوی فراگیران، فرم‌های آزمون را تکمیل می‌کنند. یکی از راه‌کارهای پیشنهادی، تهیه فرم‌ها در اندازه کوچک و جیبی است تا به محض فراهم آمدن فرصت و بدون نیاز به برنامه زمان‌بندی شده از قبل، ارزیابی در حین فعالیت های روزمره صورت پذیرد.

مشاهده می‌تواند بسته به هدف آن از ۵ تا ۳۰ دقیقه به طول انجامد و دقت در اجرای آن توصیه می‌شود. همچنین لازم است فرصت کافی برای ارائه بازخورد در نظر گرفته شود و ارزیابان آموزش کافی در مورد نحوه ارائه بازخورد مناسب دریافت کنند. به طور معمول، فرم‌های mini-CEX در دو نسخه تهیه می‌شود. نسخه اصلی به عنوان بخشی از مستندات کارپوشه به داوطلب برگردانده می‌شود و نسخه کپی آن در مرکز آزمون نگهداری می‌شود. در صورت استفاده از سیستم برخط[1]، اطلاعات فرم mini-CEX توسط ارزیاب یا کارکنان اجرایی در سامانه وارد می‌شود.

بررسی کیفیت آزمون mini-CEX برگزار شده

پس از اجرای آزمون‌های مبتنی بر محل کار، اقداماتی به منظور اطمینان از کیفیت آزمون برگزارشده انجام می‌شود که در تقریباً تمام روش‌های ارزیابی این خانواده مشترک است و می‌تواند شامل طیفی از روش‌های آماری و کمی تا روش‌های کیفی باشد. برگزارکنندگان آزمون و مسؤولان دوره می‌توانند از نتایج این اقدامات برای اصلاح آزمون‌های در حال برگزاری استفاده کنند. به عنوان مثال در مطالعه‌ای که به شرح تجربه دانشگاه ساتمپتون پرداخته است، ذکر شده است که پس از بررسی آزمون‌های mini-CEX، فرم‌ها و نمره حدنصاب قبولی تغییر یافت (هیل و همکاران ۲۰۰۹). ایجاد یک پایگاه اطلاعاتی که اطلاعات و فرم‌های آزمون در آن جمع آوری می‌شود می‌تواند به اصلاح آزمون کمک کند.

1. Online

نمره‌دهی آزمون

همان‌طور که بیان شد، یکی از محاسن آزمون mini-CEX نمره‌دهی بر اساس فرم‌های ساختارمند است. با این وجود هنوز تغییرپذیری نمره‌دهی به عنوان یکی از عوامل تهدیدکننده پایایی و به دنبال آن روایی این آزمون مطرح است. مهمترین عوامل تاثیرگذار بر نمره‌دهی آزمون mini-CEX شامل سه دسته هستند: عوامل مربوط به ارزیابان، عوامل مربوط به فرم‌ها و عوامل مربوط به مواجهه بالینی که در این قسمت به صورت مفصل مورد بحث قرار می‌گیرند (گندم‌کار و جلیلی ۲۰۱۵).

عوامل مربوط به ارزیابان

یکی از مشکلات آزمون mini-CEX ذهنی بودن قضاوت ارزیابان است. به این معنی که نظر یک ارزیاب در خصوص داوطلبان مختلف با عملکرد یکسان و نیز در مورد یک داوطلب یکسان در موقعیت‌های مختلف می‌تواند متغیر باشد. به نظر می‌رسد عوامل زیر از طرف ارزیابان بر خطا در نمره‌دهی آزمون موثر باشد:

- **ویژگی‌های ارزیاب**: نتایج چندین مطالعه نشان داده است که استادان سختگیرانه‌تر از دستیاران نمره می‌دهند اما در خصوص دیگر ویژگی‌های آزمونگران پژوهش‌های بسیار کمی انجام شده است. کوگان و همکاران (۲۰۱۰) در مطالعه خود هیچ ارتباطی بین برخی از ویژگی‌های ارزیابان شامل اطلاعات دموگرافیک و نیز میزان تجربه آن‌ها (سال‌های تدریس) با نمره‌دهی فرم‌های mini-CEX پیدا نکردند. در حالی‌که بین مهارت‌های بالینی استادان و نمره‌دهی آن‌ها ارتباط منفی معنی‌داری وجود داشت، به این صورت که استادانی که در هر یک از حیطه‌های مهارت‌های بالینی از توانمندی بیشتری برخوردار بودند، به طور سخت‌گیرانه‌تری به حیطه مربوط نمره دادند. به نظر می‌رسد پزشکان بالینی طبابت و مهارت‌های خود را مبنای قضاوت عملکرد فراگیران قرار می‌دهند. این یافته‌ها اگر در پژوهش‌های دیگر تکرار شوند از این جهت اهمیت دارند که مهارت‌های بالینی برخلاف ویژگی‌های دموگرافیک قابل تغییر هستند و در نتیجه توانمندسازی ارزیابان می‌تواند علاوه بر مهارت‌های مشاهده و ارئه بازخورد شامل تقویت مهارت‌های بالینی استادان نیز باشد.
- **میزان سخت‌گیری ارزیابان در نمره‌دهی**: از جمله مواردی که از طرف ارزیابان بر نمره‌دهی این آزمون تأثیرگذار است تفاوت آن‌ها از نظر میزان سخت‌گیری در نمره‌دهی است. برخی از ارزیابان سهل‌گیر بوده و معمولاً انتهای بالای طیف نمرات را انتخاب می‌کنند (سهل‌گیرها یا کبوترها[1]) و برخی دیگر سخت‌گیر بوده و انتهای پایین طیف را انتخاب می‌کنند (سخت‌گیرها یا بازها[2]). در برخی از مطالعات، تا ۴۰ درصد واریانس نمرات آزمون mini-CEX مربوط به این موضوع بود (ولر و همکاران ۲۰۰۹). همان‌طور که ذکر شد این خصوصیت می‌تواند از سایر ویژگی‌های آزمونگران ناشی شده باشد.
- **ارفاق**: پژوهش‌ها نشان می‌دهند زمانی که نتایج آزمون برای دانشجو کاربردهای مهمی دارد و منجر به تصمیم‌گیری‌های سرنوشت‌ساز در مورد وی می‌شود، ارزیابان با ارفاق بیشتری عمل می‌کنند و نمرات بالاتری به آنها می‌دهند. به طوری‌که بررسی فرم‌های آزمون mini-CEX بورد داخلی آمریکا نشان داد اکثریت نمرات بین ۶ تا ۷ در فرم‌های نمره‌دهی ۹ تایی بودند (الوس دلیما و همکاران ۲۰۱۳). در حالی‌که اگر برگزاری آزمون mini-CEX بخشی از یک پروژه تحقیقاتی باشد یا زمانی‌که ارزیابان به مواجهه‌های ضبط‌شده قبلی نمره می‌دهند، نمرات پایین‌تر و عموماً نسبت به زمانی که این آزمون در محیط واقعی برگزار می‌شود دقیق‌تر هستند. این مسأله نشان می‌دهد که عوامل دیگری غیر از توانمندی فراگیر و خصوصیات و ویژگی‌های ارزیاب بر قضاوت ارزیابان تأثیرگذار است. در

1. Doves
2. Hawks

همین راستا ذکر شده است که یکی دیگر از علل بالا بودن طیف نمرات در آزمون mini-CEX این است که ارزیابان احتمالاً نمی‌توانند بین سطوح عملکردی نامطلوب و مطلوب تمایز قائل شوند یا از مردود کردن داوطلبان اجتناب می‌ورزند. درنتیجه بسیاری از موارد با عملکرد نامطلوب، در طیف فراگیرانِ با عملکرد مطلوب قرار می‌گیرند. علل احتمالی عبارتند از:

- به دلیل متفاوت بودن سطح توانمندی دستیاران در یک سال تحصیلی مشابه، تعیین دقیق سطح موفقیت دشوار است.
- هر فراگیر در یک حیطه توانمند و در حیطه دیگر ضعیف عمل می‌کند و درنتیجه جمع‌بندی و ارائه یک قضاوت کلی مشکل است.
- ارتباط فردی بین استادان و دستیاران از این جهت که آن‌ها از قبل دستیاران را می‌شناسند و یک دید قبلی در مورد توانمندی آن‌ها دارند موجب می‌شود اگر در آزمون هم عملکرد خوبی نداشته باشند بر اساس دید قبلی خود به آنها نمره قبولی بدهند.
- در مواردی که خود اعضای هیأت علمی دستیاری را برای دوره دستیاری پذیرفته‌اند، رد کردن وی برای عضو هیأت علمی مشکل است.
- اعضای هیأت علمی نمی‌دانند چه نوع رفتاری را باید مستند کنند تا از این قضاوت حمایت کند. در نتیجه صبر می‌کنند تا دستیاران توانمندی لازم جهت قبولی را کسب کنند و سپس عملکرد آن‌ها را مورد ارزیابی قرار می‌دهند تا مجبور نباشند آن‌ها را رد کنند. همچنین در مواردی که داوطلب مرزی است، به گونه‌ای رفتار می‌کنند که به نفع دانشجو باشد و در نهایت نمرات بیش از آن چیزی که باید باشد است (الوس دلیما و همکاران ۲۰۱۰).

با توجه به نقش مهم تفاوت سخت‌گیری ارزیابان در نمره‌دهی، منطقی به نظر می‌رسد که آموزش ارزیابان موجب ثبات در قضاوت آنان شود. این در حالی است که نتایج یک پژوهش نشان داد آموزش ارزیابان تأثیر معنی‌داری بر ثبات بین قضاوت ارزیابان نداشت. با حضور در برنامه‌های آموزشی، تنها اعتماد به نفس شرکت‌کنندگان در اجرای آزمون بهبود یافت، آن‌ها در مشاهده مستقیم عملکرد فراگیران احساس راحتی بیشتری داشتند و سخت‌گیرتر از همکاران گروه کنترل قضاوت کردند. همچنین برگزاری کارگاه آموزشی بر دانش ارزیابان در مورد mini-CEX و افزایش احتمال ارائه بازخورد به داوطلبان مؤثر بود (هولمبو و همکاران ۲۰۰۴).

عوامل مربوط به فرم‌های آزمون

مواردی که بیشترین مشکل را در این رابطه ایجاد می‌کنند، عبارت هستند از:

- **ارتباط بالای بین سؤالات (آیتم‌ها):** به نظر می‌رسد mini-CEX بیش از آنکه حیطه‌های جداگانه‌ای از توانمندی را ارزیابی کند، توانمندی بالینی کلی[1] را ارزیابی می‌کند. مطالعات مختلف از ثبات درونی بالای فرم‌های mini-CEX حکایت دارند (نورسینی ۱۹۹۵ و ۲۰۰۳، کوگان ۲۰۰۳، کوک و بکمن[2] ۲۰۰۹، مارگولیس[3] ۲۰۰۶، دارنینگ[4] ۲۰۰۲) اکثر مطالعات سطح بالایی از ثبات درونی را با محاسبه ضریب آلفای کرونباخ برای فرم‌های نمره‌دهی گزارش کردند. به عنوان مثال ولر و همکاران (۲۰۰۹) آلفای کرونباخ ۰/۹۵ را برای فرم‌های mini-CEX محاسبه کردند و نشان دادند همه سؤالات سهم نسبتاً مساوی در نمره یا ارزیابی کلی عملکرد دارند. علاوه بر این، دو مطالعه دیگر با استفاده از تحلیل عاملی فرم‌های آزمون mini-CEX را در مقطع دستیاری (کوک و همکاران ۲۰۱۰) و در مقطع پزشکی عمومی (هیل[5] و همکاران ۲۰۰۹) بررسی کردند و نتیجه گرفتند تنها یک عامل مسؤول واریانس نمرات mini-CEX است.

1. Overall clinical competence
2. Cook & Beckman
3. Margolis
4. Durning
5. Hill

ارتباط زیاد بین سؤالات و حیطه‌های توانمندی مطرح‌کننده این موضوع است که هرچند ارزیابی دانشجو در حیطه‌های جداگانه برای ارائه بازخورد موثر و اختصاصی به وی مفید است، ارزش افزوده‌ای بیشتر از نمره کلی (میانگین نمرات توانمندی‌ها) برای آزمون ندارد. به صورت کلی این مسأله می‌تواند ناشی از علل زیر باشد:

- ارزیابان نمی‌توانند بین ابعاد یا حیطه‌های مختلفی که باید ارزیابی شوند، تمایز قائل شوند.
- ممکن است خود این حیطه‌ها واقعاً با هم مربوط باشند. به عنوان مثال به نظر می‌رسد مهارت مصاحبه و مشاوره با هم ارتباط بسیار نزدیکی داشته باشند.[1]
- دلیل دیگر می‌تواند ناشی از «اثر هاله‌ای»[2] باشد، به این معنی که قضاوت در مورد داوطلب در یک توانمندی، بر قضاوت در مورد توانمندی‌های دیگر وی نیز تأثیر می‌گذارد. این مورد می‌تواند در همه فرم‌های نمره‌دهی کلی (گلوبال) رخ دهد.
- بخشی از مورد بیماری (به عنوان مثال، تشخیص بیماری، مدیریت بیماری و ...) که برای مواجهه متمرکز استفاده می‌شود، نمی‌تواند تمایزدهنده عملکرد فراگیر در حیطه‌های مورد ارزیابی باشد.
- این احتمال وجود دارد که اکثر فراگیران نقاط قوت و ضعف خیلی مشخصی در یک توانمندی خاص نشان نمی‌دهند.
- شاخص‌های نمره‌دهی مورد استفاده در فرم نمره‌دهی برای سؤالات مختلف تا حدودی مشابه است و با هم همپوشانی دارند. به عنوان مثال، توجه به راحتی بیمار هم در معاینه بالینی و هم در رفتار حرفه‌ای وجود دارد.

- **تعداد سؤالات فرم و تعداد گزینه‌های مقیاس نمره‌دهی:** در مورد تاثیر این موارد بر نمره‌دهی آزمون، تنها دو مطالعه در دسترس است (کوک و بکمن ۲۰۰۹، دوناتو و همکاران[3] ۲۰۰۸) که نتایج آنها نشان داد پایایی بین ارزیابان با تغییر تعداد توانمندی‌های مورد ارزیابی و همچنین تغییر تعداد گزینه‌های مقیاس تغییری نکرد. فقط مقیاس نمره‌دهی نه‌تایی نمرات دقیق‌تری فراهم می‌کرد به نحوی که با دقت بیشتری توانست دستیاران را به داوطلبان با عملکرد غیرمطلوب یا عالی تقسیم نماید. بنابراین با وجود این‌که ارتباط بین دو نوع فرم زیاد بود اما فرم نمره‌دهی نه‌تایی بهتر توانست دستیارانی را که به برنامه‌های جبرانی نیاز دارند، مشخص کند. با توجه به این‌که این نتایج تنها از دو مطالعه به دست آمده است که در برخی موارد نیز ضد و نقیض بوده‌اند، نمی‌توان آن را به عنوان عوامل تاثیرگذار بر نمره‌دهی آزمون mini-CEX مطرح نمود و مسلماً تحقیقات بیشتری مورد نیاز است.

عوامل مربوط به مواجهه بالینی

مهمترین جنبه‌های تأثیرگذار مربوط به مواجهه که در متون به آن اشاره شده است، عبارت هستند از:

- **پیچیدگی موارد:** بررسی فرم‌ها نشان داده است که مواجهه با بیماری دشوارتر موجب دریافت نمرات بالاتر می‌گردد. در واقع، ارتباط معنی‌دار، اما کوچکی بین نمره‌دهی ارزیاب و دشواری مورد بیماری مشاهده شد (۰/۱۵) که نشان‌دهنده این موضوع است که ارزیابان سعی می‌کنند تا حدی دشواری مورد بیماری را با ارفاق جبران کنند (هیل و همکاران ۲۰۰۹). در نتیجه آموزش و توجیه ارزیابان در مورد دشواری موارد می‌تواند در تعدیل این مورد تأثیرگذار باشد.
- **بخش تخصصی:** معمولاً نمره‌دهی ارزیابان در بخش‌های تخصصی مختلف متفاوت است. به عنوان مثال در برخی از مطالعات، بخش جراحی پایین‌ترین نمره و بخش بیماری‌های زنان و زایمان بالاترین نمرات را دارا بودند (هیل و همکاران ۲۰۰۹). این موضوع در ارزیابی در مقطع پزشکی عمومی که دانشجویان در طول یک سال در بخش‌های

---

۱. البته برخی از پژوهشگران با این موضوع موافق نیستند و معتقدند مربوط بودن مهارت‌های مختلف با هم همبستگی بسیار کوچکتری از آنچه در مورد حیطه‌های آزمون mini-CEX دیده می‌شود دارند. به عنوان مثال همبستگی بین مهارت مصاحبه و مشاوره ۰/۱۸ گزارش شده است که بسیار کوچکتر از همبستگی این دو مهارت در فرم‌های آزمون mini-CEX است (مارگولیس و همکاران ۲۰۰۶).

2. Halo effect
3. Donato et al.

مختلف چرخش دارند و نمره کل آزمون از طریق میانگین نمرات مواجهه‌های همه بخش‌ها در طول یک سال به دست می‌آید، اهمیت دارد. به منظور حذف اثر این عامل باید همه دانشجویان در همه بخش‌ها مورد ارزیابی قرار گیرند. در مبحث مربوط به پایایی در مورد تأثیر این عامل در واریانس نمرات خطای آزمون توضیح داده شده است.

- **تمرکز مورد:** این احتمال وجود دارد که نوع تمرکز مواجهه نیز بر نمره آزمون تأثیرگذار باشد. به عنوان مثال دیده شده است که مواجهات با تمرکز بر تشخیص کمترین نمره را به خود اختصاص داده‌اند (هیل و همکاران ۲۰۰۹).

علاوه بر سه دسته کلی که در خصوص عوامل موثر بر نمره‌دهی مورد بحث قرار گرفت، فاکتورهای پراکنده‌ای نیز در متون ذکر شده است. هر چند این موارد در مطالعات مربوط به آزمون mini-CEX کمابیش دیده می‌شوند اما به طور قطعی به اثبات نرسیداند. ذکر آنها در اینجا از این جهت اهمیت دارد که طراحان و اجراکنندگان آزمون این عوامل را به منظور بهبود روایی آزمون درنظر داشته باشند. به عنوان مثال، بین مدت زمان (طول) بازخورد و نمره داوطلب همبستگی وجود دارد. به این صورت که در نمره پایین‌تر، مدت زمان بازخورد بیشتر شده است. همچنین ارتباط مثبت اما ضعیفی بین نمرات استادان و رضایت آنان از آزمون وجود دارد. اما بین نمرات و سن بیمار و نیز بین طول مواجهه و نمره آزمون، ارتباطی مشاهده نشده است.

نکته دیگری که ذکر آن در اینجا لازم است، نحوه محاسبه نمره نهایی یک مواجهه mini-CEX است به نظر می‌رسد بهتر است با توجه به مسأله تک‌بعدی بودن مهارت‌های موجود در فرم، میانگین نمرات توانمندی‌های مختلف به عنوان نمره کل در نظر گرفته شود و عملکرد داوطلب در یک توانمندی به تنهایی ملاک قضاوت قرار نگیرد. همچنین نمره توانمندی کلی نیز به صورت جداگانه اعلام شود و از نمرات هر توانمندی به منظور ارائه بازخورد و نیز ارجاع داوطلب جهت اقدامات اصلاحی استفاده شود. این موضوع در صورتی که آزمون mini-CEX با اهداف تراکمی و به عنوان یک آزمون معیارمحور استفاده می‌شود، از اهمیت ویژه‌ای برخوردار است. به عنوان مثال، در دانشگاه ساتمپتون، معیار ارجاع دانشجویان سه نمره مرزی یا زیر حدانتظار در یک توانمندی خاص در سه آزمون mini-CEX یک چرخش بالینی[1]، یا پنج نمره مرزی یا زیر حدانتظار در هر یک از توانمندی‌ها در سه آزمون برگزارشده در یک چرخش بالینی است. این فراگیران سپس در آزمون تراکمی دوره (یکسال کارآموزی) در همان حیطه‌ای که ارجاع شده‌اند، مجدداً مورد آزمون قرار می‌گیرند.

با توجه به آن‌که یکی از منابع خطا در نمرات آزمون mini-CEX هر داوطلب، تفاوت بخش‌های تخصصی است (هیل و همکاران ۲۰۰۹)، توصیه می‌شود برای محاسبه میانگین نمرات آزمون، از نمرات مواجهه‌های مختلف در یک بخش تخصصی میانگین گرفته شود. این امر به تعیین دانشجویان با عملکرد ضعیف در یک حوزه تخصصی خاص نیز کمک می‌کند.

ارائه بازخورد

مدت زمان ۵ تا ۱۰ دقیقه برای ارائه بازخورد توصیه می‌شود. در مطالعه کوگان و همکاران (۲۰۰۳) میانگین مدت زمان مشاهده عملکرد و بازخورد به ترتیب ۲۱ و ۸ دقیقه و در بررسی هور و همکاران[2] (۲۰۰۰) این زمان‌ها به ترتیب ۳۰ و ۱۵ دقیقه بود.

از لحاظ تمرکز جلسات بازخورد، یافته‌های مختلف و در برخی موارد متناقضی گزارش شده است. به عنوان مثال مطالعات نشان می‌دهند بیشتر بازخوردها مربوط به حیطه مصاحبه، معاینه بالینی، مهارت‌های ارتباطی و مشاوره بودند و بازخورد در خصوص دانش پزشکی و تعهد حرفه‌ای ناشایع بودند. البته برخی از مطالعات گزارش کردند تعهد حرفه‌ای بیشترین حیطه‌ای بود که در مورد آن بازخورد ارائه شد (لیائو[3] و همکاران ۲۰۱۳). در کل، رویکرد اعضای هیأت علمی

1. Clinical rotation
2. Hauer et al.
3. Liao

در مورد تمرکز بازخورد به دو صورت استادمحور[1] و یادگیرنده‌محور[2] است. در رویکرد استاد محور، اعضای هیأت علمی به ارائه بازخورد در مورد آنچه در آن تخصص دارند یا مورد علاقه ایشان است (مانند معاینه فیزیکی و مشاوره) می‌پردازند. برخی دیگر بر مهارت‌های اختصاصی مواجهه مورد نظر و برخی دیگر بر مهارت‌های کلی‌تر (به عنوان مثال، مهارت‌هایی که معمولاً دستیاران در آن دچار مشکل هستند) تأکید می‌کنند. رویکرد دیگری که کمتر توسط ارزیابان استفاده می‌شود رویکرد یادگیرنده محور است که در آن تمرکز ارزیابان بر نقاط ضعفی است که داوطلب در ارزیابی از خود نشان می‌دهد (کوگان و همکاران ۲۰۱۲).

از لحاظ ترتیب محتوایی جلسات بازخورد نیز بین ارزیابان تفاوت‌هایی مشاهده می‌شود. به عنوان مثال، برخی از استادان جلسه بازخورد را منطبق با ترتیب زمانی وقایع مواجهه تنظیم می‌کنند. در حالی‌که دیگران بر اساس ترتیب توانمندی‌های موجود در فرم‌ها جلسه بازخورد را مدیریت می‌کنند (کوگان و همکاران ۲۰۱۲، لیائو و همکاران ۲۰۱۳).

فنون و سبک‌های ارائه بازخورد توسط اعضای هیأت علمی را می‌توان در دو دسته رهنمودی[3] (دستوری) و تفصیلی[4] جای داد. در نوع رهنمودی، ارزیاب بر مشاهده و قضاوت خود استناد می‌کند و هدف آن ارائه اطلاعات به صورت فهرستی بلند از موارد مشاهده شده است. در این رویکرد، ارزیابان در زمان ارائه بازخورد از داوطلبان سؤالی نمی‌پرسند و استنباط استادان از عملکرد فراگیران تشریح نمی‌شود و در نتیجه مورد اصلاح یا اعتباریابی واقع نمی‌شود. در نقطه مقابل، در رویکرد تفصیلی اعضای هیأت علمی به فراگیران کمک می‌کنند تا عملکرد خود را مورد ارزیابی قرار دهند و بر آن بازاندیشی کنند. ویژگی این رویکرد بحث و پرسش‌گری است. ارزیابان در این خصوص از چندین راهبرد پرسش‌گری استفاده می‌کنند. در اکثر موارد این نوع بازخورد با خودارزیابی شروع می‌شود؛ کدام جنبه‌های مواجهه خوب انجام شده است؟ کدام بخش‌ها به درستی انجام نشده است؟ در آینده آن را چگونه اصلاح خواهد کرد و چه احساسی نسبت به مواجهه دارد؟ شروع بازخورد با خودارزیابی، زمینه را برای ارائه بازخورد منفی مهیا می‌کند و فرایند ارائه بازخورد کمتر آسیب‌رساننده می‌شود. از فواید رویکرد تفصیلی این است که استنباط اولیه ارزیابان از عملکرد داوطلب بر اساس مشاهده روشن‌تر می‌شود و فهم آن‌ها از دانش، نگرش و مهارت داوطلبان عمیق‌تر می‌شود. برخی از استادان به سؤالات خود ارزیابی در ابتدای جلسه ارائه بازخورد بسنده می‌کنند اما دیگران ممکن است در طول جلسه سؤالاتی از فراگیران بپرسند. مشکلاتی که در بازخوردهای ارائه شده توسط استادان وجود دارد، این است که در برخی مواقع با وجود این‌که آن‌ها فراگیران را ترغیب به خودارزیابی می‌کنند، اما بازخوردی که ارائه می‌دهند ارتباطی با این خودارزیابی ندارد. همچنین زمانی‌که خودارزیابی دستیاران مبهم است، به جای پرسش‌های بیشتر و روشن کردن موضوع به سرعت به بازگو کردن قضاوت خود می‌پردازند. علاوه بر این، در اکثر موارد ارزیابان نمی‌توانند به داوطلبان در زمینه‌هایی که با مشکل روبه رو هستند کمک کنند.

هولومبو و همکاران ۲۰۰۴

پژوهشگران کمیت و کیفیت بازخورد را در ۱۰۷ مورد mini-CEX که ضبط صدا شده بود، مورد بررسی قرار دادند. در ۸۰ درصد موارد ارزیابان حداقل یک توصیه برای بهبود عملکرد داوطلب ارائه دادند و در ۶۱ درصد موارد به داوطلب اجازه داده شد تا نسبت به بازخورد ارائه شده واکنش نشان دهد. تنها ۳۴ درصد ارزیابان از داوطلبان خواستند تعامل خود با بیمار را ارزیابی کنند و در نهایت در ۸ درصد موارد داوطلبان و ارزیابان بعد از جلسه برنامه عملیاتی خود را طرح‌ریزی کردند. پژوهشگران نتیجه‌گیری کردند که در جلسات مواجهه، بازخورد و پیشنهادهایی برای بهبود ارائه شده است اما خود ارزیابی فراگیران و طراحی برنامه عملیاتی کمتر استفاده شده است.
این یافته‌ها توسط مطالعات دیگری که به بررسی بازخوردهای نوشتاری در فرم‌ها پرداختند نیز تأیید شد (پلگریم و همکاران ۲۰۱۲، فرناندو و همکاران ۲۰۰۸).

1. Faculty-centered
2. Learner-centered
3. Directive
4. Elaborative

دلایل این امر می‌تواند عدم احساس راحتی اعضای هیأت علمی یا نداشتن تجربه کافی در این زمینه باشد. به عنوان مثال، دیده شده است در مواردی که خودارزیابی داوطلب یا برنامه‌ریزی عملیاتی صورت گرفته است، ارزیابان رضایت‌مندی پایینی از آزمون داشته‌اند و برعکس، میزان رضایت ارزیابان از مواجهه، با تعداد نکات مثبت ذکر شده در بازخورد ارتباط مثبت معنی‌داری داشته است (پلگریم و همکاران ۲۰۱۲، فرناندو و همکاران ۲۰۰۸).

این احتمال وجود دارد که ارزیابان نگران هستند که با ورود به خودارزیابی داوطلب بحثی باز شود که مدیریت آن مشکل باشد. پیشنهاد این است که ارزیابان با پرسیدن سؤال «چه چیزهای خوبی اتفاق افتاد؟» از خودارزیابی برای شروع ارائه بازخورد مثبت استفاده کنند و پس از آن بازخورد اصلاحی ارائه دهند. در مورد نادر بودن برنامه‌ریزی عملیاتی این احتمال وجود دارد که برنامه‌ریزی، مسؤولیتی را برای اعضای هیأت علمی با کمبود وقت، در ارتباط با تعهد برای مداخله و پیگیری این برنامه ایجاد می‌کند و درنتیجه از آن اجتناب می‌ورزند. علل احتمالی دیگر می‌تواند این باشد که ارزیابان تصور می‌کنند داوطلبان خود مسؤول برنامه‌ریزی عملیاتی هستند. در هر صورت مطالعات بیشتری در ارتباط با خودداری اعضای هیأت علمی از ارائه بازخورد تعاملی مورد نیاز است. الوس دلیما و همکاران (۲۰۱۰) گزارش کردند در مطالعه آن‌ها ارزیابان با دادن بازخورد راحت نبودند از این جهت که آموزش کافی ندیده بودند و می‌ترسیدند موجب اثرات نامطلوب مانند ناامیدی، کاهش انگیزه و کاهش اعتماد به نفس شود. همینطور آنها نمی‌دانستند اگر دانشجویی ناراحت شود، چه واکنشی باید نشان دهند.

همان‌طور که دیده می‌شود، کیفیت بازخورد ارائه شده توسط ارزیابان متنوع است و احتمالاً به عوامل متعددی بستگی دارد. عوامل مؤثر بر بازخورد در آزمون mini-CEX هنوز به طور کامل شناسایی نشده‌اند اما به نظر می‌رسد مجموعه‌ای از عوامل فردی و محیطی در آن نقش داشته باشند. کوگان و همکاران (۲۰۱۲) این عوامل را به صورت زیر ارائه نمودند:

- به نظر می‌رسد درک ارزیابان از توانایی خود در زمینه ارائه بازخورد یا به عبارت دیگر خودکارآمدی ایشان بر کارآیی بازخورد در این امر مؤثر باشد. حیطه‌هایی که اعضای هیأت علمی از خودکارآمدی لازم در ارتباط با ارائه بازخورد برخوردار نیستند، عبارت هستند از: مهارت تخصصی مورد ارزیابی، بازخورد درخصوص مهارت‌های غیرعینی مانند تعهد حرفه‌ای و همدلی، تشخیص مشکلات دستیاران، طرح‌ریزی برنامه عملیاتی و عدم اطمینان از رویکرد مناسب در ارائه بازخورد. مورد آخر به این معنی است که عنوان مثال، مطمئن نیستند که چگونه می‌توانند در ارائه بازخورد مثبت و منفی تعادل ایجاد کنند، نمی‌دانند که آیا باید در زمان ارائه بازخورد یادداشت بردارند یا نهایتاً نمره آزمون را باید به داوطلب اعلام کنند یا خیر. همچنین در مورد کافی بودن میزان بازخورد ارائه شده یا میزان جدیت لازم در ارائه بازخورد مطمئن نیستند.
- هیجانات اعضای هیأت علمی بر بازخورد ارائه شده تأثیرگذار است. آن‌ها باید بتوانند خشم و عصبانیت خود را فروکش کنند تا بازخورد اصلاحی توهین‌آمیز نباشد. همچنین ارائه بازخورد منفی به نوبه خود نیازمند شجاعت است و اکثر ارزیابان در این مورد احساس خوبی ندارند. از نظر پزشکان بالینی، ارائه بازخورد منفی به فراگیران به سختی ارائه خبر تشخیص سرطان به بیمار است (کوگان و همکاران ۲۰۱۲).
- رویکرد ارزیابان به ارائه بازخورد تحت تأثیر هدف آن‌ها از ارائه بازخورد قرار می‌گیرد. هدف اصلی اکثر ارزیابان بهبود مهارت‌های فراگیران با تأکید بر نقاط مثبت و توصیف نقاط قابل بهبود است. این مسأله نیازمند تلاش در ایجاد تعادل در جنبه‌های مثبت و منفی (انتقادی) بازخورد است. از یک سو، استادان تمایل دارند با ارائه بازخورد مثبت، فراگیران را به ادامه مهارت صحیح ترغیب کنند و از سوی دیگر می‌خواهند با انتقاد سازنده آن را تعدیل کنند. هر چند به نظر می‌رسد در اغلب موارد در واقعیت تأکید بر نقاط مثبت بیشتر است.

فرناندو و همکاران ۲۰۰۸

ارائه بازخورد مثبت و برنامه‌ریزی عملیاتی به طور معنی‌داری با ویژگی‌های ارزیاب ارتباط داشت. به طوری‌که اعضای هیأت علمی که با آموزش و تدریس دانشجویان پزشکی سر و کار داشتند، نسبت به افرادی که مسؤولیت‌های بالینی را انجام می‌دادند، بیشتر به ارائه این موارد در بازخورد ارائه شده پرداختند.

طیف نمرات در شش توانمندی و توانمندی کلی در فرم ارزیابی با ارائه بازخورد به همراه توصیه‌هایی برای بهبود و برنامه‌ریزی عملیاتی ارتباط داشت. به این معنی که ارزیابانی که با دقت بیشتری توانمندی‌ها را نمره‌دهی می‌کردند و نمرات آن‌ها طیف وسیع‌تری داشت، وقت بیشتری را صرف ارائه بازخورد می‌کردند.

- علاوه بر این، ارائه بازخورد تحت تأثیر ادراک اعضای هیأت علمی از فراگیران شامل عملکرد فراگیران، بینش آن‌ها، پذیرش بازخورد از سوی فراگیران و پتاسیل آن‌ها برای بهبود قرار می‌گیرد. به عنوان مثال، برای اکثر قریب به اتفاق استادان، ارائه بازخورد به فراگیرانِ با عملکرد ضعیف چالش‌برانگیز است. به ویژه اگر تعداد اشتباهات زیاد باشد و در نتیجه ایجاد تعادل در ارائه بازخورد مثبت و منفی دشوار است. همچنین، ارزیابان در ارائه بازخورد به فراگیرانی که نسبت به عملکرد خود آگاهی دارند، اشتباهات خود را می‌پذیرند و می‌خواهند آن را ارتقاء دهند، احساس راحتی بیشتری می‌کنند.
- عوامل زمینه‌ای و ارتباطات نیز در ارائه بازخورد نقش دارند. برای بسیاری از اعضای هیأت علمی ارتباط ارزیاب با داوطلب به ویژه دستیاران مهمترین عامل تعیین کننده بازخورد ارائه شده است. اگر شناخت و ارتباط ارزیاب با دستیار به مدت طولانی باشد، ارزیابان بهتر می‌توانند اعتماد و پذیرش آن‌ها را برای ارائه بازخورد جلب کنند. آگاهی از توانمندی قبلی فراگیران نیز بر نحوه ارائه بازخورد تأثیرگذار است.

در مجموع، آزمون mini-CEX پتانسیل ارائه بازخورد با کیفیت بالا و تعاملی را دارد. اولین گام مهم، مشاهده مستقیم عملکرد فراگیر است و mini-CEX امکان مشاهده پایا و ساختارمند مهارت‌های بالینی را فراهم می‌کند. به منظور افزایش کارآیی، بازخورد باید یک فرایند دوطرفه باشد و علاوه بر توصیه‌های ارزیابان، بازاندیشی داوطلبان بر تجربه ارزیابی را نیز در نظر داشته باشیم. این فرایند پویا و پیچیده تحت‌تأثیر عوامل متعددی مانند ادراک ارزیابان از توانایی‌های خود در ارائه بازخورد، هیجانات ایشان، هدف آن‌ها از ارائه بازخورد، درک آنها از بینش، مهارت‌ها و پتانسیل‌های داوطلبان آزمون و همچنین ارتباط داوطلب و ارزیاب و عوامل زمینه‌ای قرار می‌گیرد. فهم این عوامل در یافتن رویکردهای جدید در توانمندسازی اعضای هیأت علمی در ارائه بازخورد مؤثر کمک کننده است.

سودمندی آزمون mini–CEX

آزمون mini-CEX امکان مشاهدات متعدد عملکرد فراگیر در طول زمان و توسط ارزیابان متعدد را فراهم می‌کند. این امر منجر به بهبود پایایی و روایی آزمون می‌شود. ماهیت طولی mini-CEX یکی از مهمترین نقاط قوت این روش است.

روایی آزمون mini–CEX

حجم رو به رشدی از مطالعات از روایی آزمون mini-CEX حمایت می‌کند.

- **روایی محتوا:** mini-CEX امکان نمونه‌گیری از طیف وسیعی از موقعیت‌های بالینی و موارد بیماری (بیماران با اولین مراجعه یا تحت نظر و بیماران با طیف وسیعی از شکایات و مشکلات بالینی) و در نتیجه پوشش وسیعی از محتوا را فراهم می‌کند بنابراین روایی محتوایی خوبی دارد.
- **روایی صوری:** با توجه به این‌که mini-CEX به ارزیابی مهارت اخذ شرح‌حال و معاینه متمرکز در محیط کار

می‌پردازد، می‌تواند بازنمایی از طبابت واقعی داوطلبان باشد. علاوه بر این، عملکرد داوطلب در زمان انجام مهارت‌ها مورد مشاهده قرار می‌گیرد و در نتیجه باور بر این است که از روایی صوری مناسبی برخوردار است. البته شواهد محکمی که این موضوع را اثبات کند وجود ندارد.

- **روایی معیار:** اکثر مطالعاتی که در زمینه روایی mini-CEX انجام شده‌اند، به بررسی ارتباط نمرات این آزمون با نمرات آزمون‌های دیگری که حیطه‌های بالینی مشابه را می‌سنجند پرداخته‌اند. در مجموع، نتایج این پژوهش‌ها بر ارتباط بالای بین نمرات mini-CEX با آزمون‌های مختلف و در نتیجه روایی همزمان خوب این آزمون دلالت دارد. در زیر نمونه‌ای از نتایج پژوهش‌های انجام شده آمده است:
  - نمرات هر یک از توانمندی‌های مورد ارزیابی در mini-CEX مانند اخذ شرح‌حال، معاینه فیزیکی و غیره با نمرات توانمندی نظیر آن در آزمون بورد داخلی آمریکا تحت عنوان فرم‌های ارزیابی ماهانه[1] و آزمون ضمن دوره برگزارشده توسط جامعه پزشکان داخلی آمریکا[2] (که معادل آزمون ارتقای دستیاری در کشور است) ارتباط زیادی داشت (دارنینگ و همکاران ۲۰۰۲).
  - نمره توانمندی کلی دستیاران طب داخلی در mini-CEX با سه بخش آزمون جامع بیماری‌های داخلی برگزارشده توسط کالج سلطنتی پزشکان و جراحان کانادا[3] شامل آزمون شفاهی ساختارمند (r=۰/۷۳)، آزمون بالینی در کنار بستر بیمار (r = ۰/۶۷) و آزمون کتبی r = ۰/۷۲) ارتباط قابل ملاحظه‌ای داشت[4] (هاتالا و همکاران[5] ۲۰۰۶).
  - در مقطع پزشکی عمومی، نمرات آزمون mini-CEX دانشجویان با نمرات آزمون‌های دیگر کارآموزی آن‌ها ارتباط ضعیف تا متوسطی داشت. این آزمون‌ها شامل امتحان موضوعی بورد ملی آزمون‌گران پزشکی (p = ۰/۰۰۴، r = ۰/۲۲)، آزمون پایانی دوره (فرم‌های نمره‌دهی کلی) (p = ۰/۰۱۴، r = ۰/۱۹)، ارزیابی درون بخش (p = ۰/۰۰۱، r = ۰/۴۳)، ارزیابی در درمانگاه (p = ۰/۰۰۱، r = ۰/۳۵) و شرح‌حال‌های بیمار[6] (p = ۰/۰۳۵، r = ۰/۱۷) بود (کوگان و همکاران ۲۰۰۳).
- **روایی سازه:** این روش ارزیابی دارای روایی سازه معقولی است زیرا می‌تواند بین سطوح متفاوت عملکرد فراگیران تمایز قائل شود. از جمله شواهد موجود در این زمینه می‌توان به موارد زیر اشاره کرد:
  - میانگین نمرات توانمندی کلی دستیاران طب داخلی سال دوم به طور معنی‌داری بالاتر از نمرات دستیاران سال اول بود (نورسینی و همکاران ۱۹۹۵).
  - نمرات دستیاران سال اول طب داخلی در طول سال در همه حیطه‌های توانمندی افزایش یافت (نورسینی و همکاران ۲۰۰۳).
  - ارزیابان توانستند بین دستیاران سال دوم طب داخلی با سطوح مختلف عملکرد (ضعیف، مرزی و عالی) تمایز قائل شوند. بر اساس نتایج این مطالعه، تفاوت اندازه اثر بین فراگیران سه سطح مذکور در مهارت‌های مصاحبه، معاینه فیزیکی و مشاوره بزرگ بود: این تفاوت از ۰/۹۰ بین دستیاران عالی و مرزی (در مهارت مصاحبه) تا ۴ بین دستیاران عالی و ضعیف (در مهارت معاینه بالینی) در مقیاس نه‌تایی متغیر بود. (هولمبو ۲۰۰۳، اَل انصاری[7] و همکاران ۲۰۱۳).
  - نمرات دانشجویان پزشکی در طول یک سال کارآموزی در طول چهار چرخش بالینی افزایش یافت (کوگان و همکاران ۲۰۰۳)

---

1. Monthly Evaluation Form (MEF)
2. American Society of Internal Medicine In Training Examination (ITE)
3. Royal College of Physicians and Surgeons of Canada Comprehensive Examination in Internal Medicine (RCPSC IM)

۴. آزمون RCPSC از سه بخش تشکیل شده است: آزمون کتبی شامل ۲۰۰ سؤال چند گزینه‌ای، آزمون شفاهی ساختارمند با هدف ارزیابی توانایی جمع‌بندی و سنتز اطلاعات، بحث، تفسیر و مدیریت موارد بیماری و یک آزمون بالینی در بخش شامل ارزیابی مهارت معاینه بالینی، مهارت‌های ارتباطی و اخلاقی است. این آزمون دو ساعت به طول می‌انجامد و در دو بخش یک ساعته صبح و بعدازظهر برگزار می‌شود.

5. Hatala et al.
6. Patient write-ups
7. Al Ansari

- میانگین نمرات دستیاران قلب در حیطه‌های معاینه بالینی، قضاوت بالینی و توانمندی در بین دستیاران سال‌های اول تا چهارم افزایش معنی‌داری نشان داد (الوس دلیما و همکاران ۲۰۰۷)

آل‌انصاری و همکاران ۲۰۱۳

این مطالعه یک متآنالیز با هدف بررسی روایی سازه و روایی معیار (روایی همزمان و پیش‌بین) آزمون mini-CEX بود. پژوهشگران ۱۱ مقاله مرتبط با هدف متآنالیز و منطبق با معیارهای ورود را مورد بررسی قرار دادند. آن‌ها مقالات را از این نظر به چهار دسته تقسیم کردند: مقالاتی که تغییر عملکرد دستیاران را در سال‌های تحصیلی مختلف (سال اول به دوم، دوم به سوم و غیره) گزارش کردند (گروه A)، مقالاتی که عملکرد فراگیران (هم‌کلاسی‌ها) مختلف (عالی/برتر، مرزی/بالاتر از نمره قبولی و ضعیف/قبول) را در یک گروه با هم مقایسه کردند (گروه B)، مقالاتی که به بررسی تفاوت نمره‌دهی دستیاران و استادان پرداختند (گروه C) و مقالاتی که ارتباط نمرات mini-CEX با آزمون‌های دیگر را مورد بررسی قرار دادند (گروه D). تفاوت اندازه اثر با ضریب کوهن (d) یا ضریب همبستگی پیرسون (r) گزارش شد.

- هفت مقاله روایی سازه را مورد بررسی قرار دادند که از این تعداد ۴ مقاله در گروه A، تفاوت اندازه اثر را از (۰/۰۴-۰/۴۶ CI ۹۵٪، d=۰/۲۵) در تعهد حرفه‌ای تا (۰/۳۱-۰/۷۰ CI ۹۵٪، d =۰/۵۰) در توانمندی کلی در بین دستیاران سال‌های مختلف گزارش کردند. در کل، ترکیب مطالعات این گروه برای توانمندی کلی یک تفاوت اندازه اثر «متوسط» (۰/۳۱-۰/۷۰ CI ۹۵٪، d=۰/۵۰) را گزارش کرد.
- دو مقاله در گروه B، تفاوت «متوسطی» را بین همکلاسی‌ها، در مهارت‌های بالینی (سه مورد) و توانمندی کلی گزارش کرد.
- در گروه C، نتایج دو مطالعه مورد بررسی قرار گرفت و نشان داد نمره‌دهی اعضای هیأت علمی به دانشجویان در هر ۷ حیطه به طور پایداری سخت‌گیرانه‌تر از ارزیابی دانشجویان توسط دستیاران است. این تفاوت اندازه اثر در توانمندی کلی به صورت (۰/۶۲- CI ۰/۱۵ ۹۵٪، d = ۰/۳۸) بود.
- پنج مطالعه در گروه D، تفاوت اندازه اثر برای هر یک از مهارت‌های بالینی و صلاحیت کلی را "متوسط" گزارش کرد. اما تفاوت اندازه اثر در نمره کلی آزمون mini-CEX و آزمون‌های دیگر "کوچک" (۰/۳۵-۰/۱۶ CI ۹۵٪، d = ۰/۲۶) بود.

یافته‌های این مطالعه در مورد روایی معیار و روایی سازه آزمون mini-CEX، یک اندازه اثر پایدار و متوسط در نمرات آیتم‌ها و نمره مهارت کلی نشان می‌دهد.

با توجه به مطالبی که در بالا بیان شد به نظر می‌رسد آزمون mini-CEX از روایی نسبتاً مناسبی برخوردار باشد. بررسی متون نشان می‌دهد تغییرپذیری نمره‌دهی آزمون mini-CEX به عنوان یکی از عوامل تهدیدکننده روایی این آزمون مطرح است. به این موارد در قسمت نمره‌دهی با جزئیات بیشتر پرداخته شده است.

پایایی آزمون mini-CEX

در مجموع می‌توان گفت در آزمون mini-CEX سعی شده است پایایی با افزایش تعداد مشاهدات، استفاده از ارزیابان مختلف و استفاده از فرم‌های ساختارمند با تعداد سؤالات مناسب بهبود داده شود:

- **تعداد مواجهه:** پژوهش‌های انجام شده در مقاطع پزشکی عمومی و دستیاری، ضرایب پایایی متفاوتی به ازای تعداد مواجهات مختلف گزارش کرده‌اند. به عنوان مثال، در مقطع دستیاری برای هشت و چهارده مواجهه به ترتیب ضرایب تعمیم‌پذیری ۰/۷۱ و ۰/۸۱ (نورسینی و همکاران ۱۹۹۵)، برای هشت مواجهه ۰/۷۷ (کوگان و همکاران ۲۰۰۳) و در ارزیابی دانش‌آموختگان پزشکی بین‌المللی در استرالیا برای هشت مواجهه ۰/۸۸ (نیر و همکاران[1] ۲۰۰۸) گزارش شده است. در مقطع پزشکی عمومی، هیل و همکاران (۲۰۰۹) برای پانزده مواجهه ضریب پایایی ۰/۷۳ گزارش کردند.

مجموع این گزارش‌ها این احتمال را مطرح کردند که آزمون mini-CEX در پزشکی عمومی پایایی کمتری از مقطع دستیاری دارد. اما باید در نظر داشت که اولاً فرم‌های مورد استفاده در این مطالعات متفاوت بوده‌اند. به عنوان مثال فرم‌های استفاده شده توسط بورد داخلی امریکا (نورسینی ۱۹۹۵) لیکرت نه‌تایی و فرم‌های مورد استفاده توسط سلامت ملی انگلیس (هیل و همکاران ۲۰۰۹) لیکرت شش‌تایی هستند. ثانیاً دلیل این تفاوت می‌تواند یکسان نبودن روش‌های آماری مورد استفاده (مدل رگرسیونی محدود یا گسترده[2]) در مطالعات مختلف باشد. به طوری‌که با استفاده از مدل رگرسیونی محدود مانند آن‌چه در مطالعات اولیه استفاده شده است (نورسینی و همکاران ۱۹۹۵)، پایایی بالاتر از زمانی که مدل رگرسیونی

1. Nair et al.
2. Extended or limited regression model

گسترده استفاده شده است، به دست آمده است و در صورتی که الگوی تحلیل استفاده‌شده یکسان باشد، نتایج در هر دو مقطع تحصیلی تقریباً یکسان بوده است (هیل و همکاران ۲۰۰۹). ثالثاً توجه به این نکته نیز لازم است که یکی دیگر از دلایل این تفاوت می‌تواند شرایط مطالعه باشد. در مواردی که نتایج آزمون در محیط واقعی مورد بررسی قرار گرفته است، تعداد مواجهه‌های مورد نیاز کمتر از مواردی بوده است که مطالعه در شرایط کنترل‌شده (مانند ارزیابی سناریوهای ضبط‌شده توسط استادان) صورت گرفته است. شاید به این دلیل که طیف عملکرد فراگیران در سناریوهای ضبط شده متنوع‌تر از عملکرد در محیط واقعی بوده است.

هر چند به صورت کلی تعداد بیش از ده مواجهه برای دستیابی به پایایی بیش از ۰/۸ ضروری است، اما همان‌طور که در فصول پیشین نیز بیان شد با استفاده از فاصله اطمینان و خطای معیار اندازه‌گیری می‌توان با تعداد کمتر مواجهه اطلاعات بیشتری کسب کرد. به نتایج ضریب تعمیم‌پذیری و خطای معیار اندازه‌گیری آزمون mini-CEX برای ۱ تا ۱۴ مواجهه که در جدول ۳-۲۱ نشان داده شده است، توجه کنید:

جدول ۳-۲۱: ضریب تعمیم‌پذیری و خطای معیار اندازه‌گیری آزمون mini-CEX برای ۱ تا ۱۴ مواجهه (نورسینی و همکاران ۱۹۹۵)

تعداد مواجهه	ضریب تعمیم‌پذیری	خطای معیار اندازه‌گیری
۱	۰/۲۳	۰/۷۱
۲	۰/۳۸	۰/۵۰
۴	۰/۵۵	۰/۳۵
۶	۰/۶۵	۰/۲۹
۸	۰/۷۱	۰/۲۵
۱۰	۰/۷۵	۰/۲۲
۱۲	۰/۷۸	۰/۲۰
۱۴	۰/۸۱	۰/۱۹

یک دستیار با عملکرد متوسط را در نظر بگیرید که در آزمون فوق شرکت کرده است و میانگین نمره ۶/۶ را در چهار مواجهه کسب کرده است. هر چند ضریب تعمیم‌پذیری با این تعداد مواجهه ۰/۵۵ است اما خطای معیار اندازه‌گیری ۰/۳۵ است. بنابراین نمره این دستیار در فاصله اطمینان ۹۵ درصد بین ۵/۹ تا ۷/۳ می‌شود (۰/۷ ± ۶/۶، ۰/۷ = ۲ × ۰/۳۵). این نمرات در مقیاس نمره‌دهی نه‌تایی در طیف نمرات مطلوب تا عالی قرار دارد و اگر از نظر کاربردی تمایز بین این نمرات اهمیت چندانی نداشته باشد، تعداد چهار مواجهه برای ارزیابی دستیار مورد نظر کفایت می‌کند.

اکنون دستیاری را درنظر بگیرید که در آزمون فوق با چهار مواجهه نمره مرزی چهار را دریافت کرده است. نمره وی در فاصله اطمینان ۹۵ درصد بین ۳/۳ تا ۴/۷ می‌شود (۰/۷ ± ۴ ۰/۷ = ۲ × ۰/۳۵) که در طیف نمرات غیرمطلوب تا مطلوب قرار دارد. برای برگزارکنندگان آزمون اهمیت دارد که دریابند دستیار مورد نظر نهایتاً در کدام دسته قرار می‌گیرد و این موضوع جهت تصمیم‌گیری در مورد عملکرد دستیار اهمیت دارد. پس تعداد بیشتر مواجهه لازم است تا تکلیف او به صورت دقیق مشخص شود.

در مجموع، استفاده از خطای معیار اندازه‌گیری و فاصله اطمینان، با کاهش تعداد مواجهه‌ها هزینه آزمون را کاهش می‌دهد. علاوه بر این، با افزایش تعداد مواجهه‌ها برای دستیار با عملکرد مرزی و افزایش احتمال ارائه بازخورد موجب

افزایش تأثیر آموزشی آزمون بر کسانی که به آن نیاز دارند می‌شود. بنابراین، mini-CEX ضریب تکرار پذیری عملی[1] دارد؛ به این معنی که نمرات به دست آمده از چهار مواجهه می‌تواند پیش‌بینی‌کننده این موضوع باشد که آیا مواجهه‌های بیشتری موردنیاز است. تعداد مواجهه مورد نیاز می‌تواند با توجه به هدف آزمون (تکوینی یا تراکمی)، حساسیت و اهمیت آن، ماهیت فراگیران و محل برگزاری آزمون تغییر کند.

- **تعداد ارزیاب:** ولر و همکاران (۲۰۰۹) در مطالعه خود با استفاده از نظریه تعمیم‌پذیری به تخمین منابع و میزان خطا در نمرات آزمون mini-CEX پرداختند. بر اساس نتایج این مطالعه، سه منبع اصلی تغییرپذیری نمرات آزمون عبارت بودند از: سخت‌گیری ارزیاب در نمره‌دهی (۴۰ درصد)، ویژگی موارد (۴۰ درصد) و ذهنی بودن قضاوت ارزیابان (۱۵درصد). در مطالعه هیل و همکاران (۲۰۰۹) مهمترین منابع خطا، سخت‌گیری ارزیاب (۲۹ درصد) و تفاوت در بخش‌های تخصصی فراگیران[2] (۱۳ درصد) بودند. مورد آخر به این معنی است که آنهایی که در برخی از بخش‌ها خوب عمل کرده‌اند در برخی دیگر خوب عمل نکرده‌اند. مطالعات دیگر نیز در این زمینه مهمترین منابع خطا در نمرات mini-CEX را سخت‌گیری ارزیاب یا تفاوت بین ارزیابان گزارش کردند.

مارگولیس و همکاران (۲۰۰۶) علاوه بر موضوع فوق به نتایج جالبی در مطالعه خود دست یافتند. آنها نشان دادند که سخت‌گیری ارزیاب، منبع مهمی از واریانس نمرات است به طوری‌که از واریانس مورد بیماری بیشتر است. پژوهشگران نتیجه گرفتند این احتمال وجود دارد که موضوع ویژگی موارد بیماری در مورد آزمون mini-CEX کمتر صدق کند. مطالعه الوس دلیما و همکاران (۲۰۱۳) نیز این یافته‌ها را تأیید کرد. بر اساس نتایج این دو مطالعه به نظر می‌رسد مهمترین منبع خطا، سخت‌گیری ارزیابان و در واقع پایایی بین ارزیابان است تا ویژگی موارد و این مسأله برعکس آزمون OSCE است که در آن واریانس موارد بالا و واریانس ارزیابان پایین است. یکی از دلایل موضوع می‌تواند این باشد که شاید در mini-CEX متخصصان چیزی را ارزیابی می‌کنند که ورای موارد بیماری است. علت احتمالی دیگر این است که آزمون mini-CEX مواردی از مهارت‌های بالینی را که ویژگی موارد در مورد آن‌ها وجود دارد (مانند جمع‌آوری اطلاعات و تصمیم‌گیری بالینی) ارزیابی نمی‌کند.

کاربرد عملی دو پژوهش مذکور این است که با افزایش تعداد مواجهات بدون تنوع در ارزیابان و استفاده از ارزیابان یکسان نمی‌توان پایایی آزمون را افزایش داد. به بیان دیگر هر داوطلب باید توسط تعداد متنوعی از ارزیابان مورد قضاوت قرار گیرد حتی زمانی‌که تعداد موارد بیماری بالا است. به طوری که در صورت استفاده از ارزیابان متفاوت برای مواجهه‌های مختلف، ۹ مواجهه پایایی مناسبی را به دست می‌دهد. در حالی‌که اگر ارزیاب برای همه مواجهه‌های یک داوطلب یکسان باشد، ۱۵ مواجهه برای رسیدن به سطح مطلوب پایایی مورد نیاز است. حال اگر دو ارزیاب برای هر مواجهه استفاده شود، که در واقعیت احتمال عملی شدن آن کم است، با نصف این تعداد پایایی مورد نظر به دست می‌آید.

با یک مثال این مطلب را بیشتر باز می‌کنیم. در صورتی‌که یک ارزیاب تمام ده مورد مواجهه mini-CEX یک آزمون‌شونده را ارزیابی کند، پایایی آزمون ۰/۳۹ است. در حالی‌که با همین تعداد مواجهه و با ارزیابان متناوب برای هر آزمون‌شونده در هر مواجهه، ضریب پایایی ۰/۸۳ خواهد بود (الوس دلیما و همکاران ۲۰۱۳). البته این یافته‌ها با نتایج مطالعه نیر و همکاران (۲۰۰۸) در مورد دانشجویان بین‌المللی متفاوت است مبنی بر این‌که یک ارزیاب یکسان برای هشت مواجهه ضریب پایایی خوبی (۰/۸۸) را به دنبال داشت.

- **ثبات درونی:** همان‌طور که در قسمت مربوط به نمره‌دهی عنوان شد، مطالعات مختلف با استفاده از محاسبه ضریب آلفای کرونباخ و تحلیل عاملی حیطه‌های توانمندی فرم mini-CEX، به ثبات درونی بالای آزمون (همبستگی معنی‌دار آماری بین توانمندی‌ها با هم و توانمندی‌ها با توانمندی کلی) اشاره دارند. هرچند، این یافته‌ها روایی نمرات آزمون را زیر

1. Pragmatic
2. Student attachment-specific aptitude

سؤال می‌برد به این معنی که mini-CEX نمی‌تواند بین سطوح اختصاصی توانمندی تمایز قائل شود. موارد مربوط به ثبات در نمره‌دهی و خطاهای مربوط به آن در قسمت مربوط به نمره‌دهی به تفصیل شرح داده شد.

تأثیر آموزشی آزمون mini-CEX

مانند دیگر روش‌های ارزیابی مبتنی بر محل کار، یکی از اهداف اصلی mini-CEX ارتقای یادگیری است و این مهم به میزان زیادی از طریق بازخورد ارائه شده به دست می‌آید. بیشتر مطالعاتی که در این زمینه انجام شده است، در سطح بررسی نظرات و دیدگاه‌های داوطلبان و گاهی اعضای هیأت علمی در مورد اثر آموزشی آزمون بوده است. هر چند تعداد این مطالعات زیاد نیست، اما اکثریت قریب به اتفاق بر تأثیر آموزشی خوب این روش ارزیابی دلالت دارند. از جمله اثرات آموزشی mini-CEX می‌توان به توجه بیشتر به مهارت‌های بالینی، صرف وقت بیشتر برای تمرین شرح‌حال و معاینه بالینی متمرکز، افزایش زمان مطالعه، افزایش انگیزه برای یادگیری، ارتقای مهارت‌های خودتنظیمی برای مطالعه، کاهش مطالعه و یادگیری در کتابخانه و تعیین دانشجویان مشکل‌دار در ابتدا و در طول چرخش بالینی اشاره کرد.

به نظر می‌رسد دلایل اصلی این تأثیر آموزشی مورد مشاهده واقع شدن و دریافت بازخورد از ارزیابان متعدد و در محیط‌های متعدد است (هیل و کندل[1] ۲۰۰۷). در مطالعه ولر و همکاران (۲۰۰۹) فراگیران و اعضای هیأت علمی نظر مثبتی نسبت به اثر مثبت آموزشی این روش ارزیابی داشتند. همچنین اظهار داشتند این ابزار کمیت و کیفیت بازخورد را بهبود می‌بخشد و نیمی از آن‌ها معتقد بودند که آزمون mini-CEX اعتماد به نفس آن‌ها را برای درخواست و یا ارائه بازخورد بهبود می‌بخشد.

نظرات دانشجویان سال آخر پزشکی در مورد آزمون mini-CEX (هیل و کندل ۲۰۰۷)

«به نظر من مهمترین ویژگی این آزمون این است که شما یک جلسه آموزشی دارید؛ شما مورد مشاهده قرار می‌گیرید، عملکرد شما نقد می‌شود که این بسیار مفید است و از همه مهمتر این‌که به ندرت این فرصت پیش می‌آید که یکی از استادان شما را در حین انجام معاینه بالینی مشاهده کند».

«آزمون باعث شده من در بخش فعال‌تر باشم-خیلی بیشتر- چون می‌دانم کسی هست که برخورد من با بیمار را مورد مشاهده قرار می‌دهد، پس من باید آن را به بهترین نحو انجام دهم.»

هزینه و قابلیت اجرای آزمون mini-CEX

در کل یکی از مزایای این آزمون، امکان‌پذیری نسبتاً خوب آن است و این در حالی است که اکثر آزمون‌های مبتنی بر محل کار خیلی راحت اجرا نمی‌شوند. هیل و کندل (۲۰۰۷) گزارش کردند هر چند قبل از اجرای این آزمون، نگرانی‌هایی در مورد پیاده‌سازی این آزمون وجود داشت اما تجربه نشان داد mini-CEX قابلیت اجرای خوبی دارد. هزینه و قابلیت اجرای آزمون را می‌توان از جنبه‌های زیر بررسی نمود:

- **زمان صرف‌شده برای آزمون:** انجام این آزمون آسان است و زمان‌بر نیست (کوگان ۲۰۰۲ و ولر ۲۰۰۹). زمان صرف شده برای آزمون وابسته به موقعیت بالینی، پیچیدگی موارد و نوبت مراجعه بیمار متفاوت است. زمان صرف شده برای دانشجویان بیشتر از دستیاران است و اگر دستیاران به عنوان ارزیاب برای دانشجویان پزشکی باشند زمان به طور معنی‌داری بیشتر از زمانی است که استادان به عنوان ارزیاب صرف می‌کنند (کوگان و همکاران ۲۰۰۳). معمولاً زمان صرف شده برای مشاهده ۲۵ دقیقه و زمان ارائه بازخورد تا یک سوم زمان صرف شده برای مشاهده است. یکی از مشکلات اجرایی آزمون mini-CEX هماهنگ کردن و یافتن زمانی برای اجرای آزمون است. با احتساب زمان صرف شده جهت هماهنگی، زمان کل صرف‌شده به یک ساعت می‌رسد.

1. Kendall

- **احتمال اجرای مواجهه‌های برنامه‌ریزی‌شده:** در این خصوص، نظرات ضد و نقیضی در متون مرتبط وجود دارد. برخی از پژوهش‌ها از اجرای موفقیت‌آمیز برنامه تعیین شده گزارش دادند به طوری‌که تا ۹۶ درصد تعداد مواردی که برنامه‌ریزی شده بود، اجرا شد. مطالعه دیگری اجرای متوسط هشت مواجهه در ۱۲ هفته چرخش بالینی دانشجویان پزشکی را گزارش کرد. اما در برخی موارد تنها تا ۱۴ درصد تعداد مواجهه‌های برنامه‌ریزی شده اجرا شد. به طوری‌که در طول ۲۰ ماه چهار مواجهه mini-CEX برای هر دستیار (۱۴ درصد گروه دستیاران) ثبت شد (یوسف[1] ۲۰۱۲).
- **میزان تکمیل فرم‌ها و آیتم‌ها:** رقم خوبی برای این مورد در آزمون mini-CEX گزارش شده است و برخی از مطالعات این رقم را تا ۹۶ درصد اعلام کردند. تور و همکاران[2] (۲۰۰۷) این آزمون را با قابلیت اجرایی بالا توصیف کردند و اجرای یک مواجهه در هر ماه با میزان تکمیل ۱۰۰ درصدی فرم‌ها را گزارش کردند.

به ندرت پژوهشی در زمینه بررسی هزینه آزمون mini-CEX انجام شده است. برزیل و همکاران[3] (۲۰۱۲) با محاسبه زمان صرف شده توسط اعضای هیأت علمی بالینی به تخمین هزینه آزمون پرداختند. علاوه بر زمان صرف شده توسط استادان، موارد دیگری که در محاسبه هزینه آزمون mini-CEX باید در نظر گرفته شود عبارتند از: زمان صرف‌شده توسط فراگیران و کارکنان اجرایی، هزینه مواد مصرفی (مانند فرم‌های پرینت‌شده)، هزینه‌های نرم‌افزاری (مانند مستند کردن اطلاعات)، هزینه‌های مربوط به تهیه بلوپرینت و حدنصاب قبولی.

علاوه بر این، مانند هر روش ارزیابی مبتنی بر محل کار آن‌چیزی که اهمیت بیشتری دارد هزینه–اثربخشی ابزار ارزیابی است. به عنوان مثال، یکی از مهمترین پیامدهای آزمون mini-CEX تشخیص مشکلات دستیاران و دانشجویان پزشکی در مهارت‌های مختلف بالینی است و در نتیجه هزینه‌های صرف شده باید در پرتوی پیامدهای به دست آمده تفسیر شود.

مقبولیت آزمون mini-CEX

نتایج مطالعات مختلف کمی و کیفی از پذیرش نسبتاً خوب این آزمون توسط داوطلبان و ارزیابان حکایت می‌کند (نورسینی ۱۹۹۵ و ۲۰۰۳، کوگان ۲۰۰۲، هیل و کندل ۲۰۰۷، هیل و همکاران ۲۰۰۹، سیدهو و همکاران[4] ۲۰۰۹). در یک نظرسنجی از فراگیران و استادان با مقیاس نمره‌دهی ده‌تایی، میانگین رضایت فراگیران و متخصصان به ترتیب ۷/۳ و ۷/۲ بود. متخصصان در مورد انجام این روش ارزیابی خنثی بودند اما ۴۵ درصد فراگیران برای انجام آن اشتیاق داشتند. اکثر آن‌ها برای درخواست آزمون از استاد خود راحت بودند. ارائه بازخورد نیز اکثراً در اتاق کار استادان اتفاق می‌افتاد که بیشتر فراگیران نسبت به این موضوع رضایت داشتند.

در یک مطالعه هم دانشجویان و هم استادان mini-CEX را به عنوان یک ارزیابی تراکمی، به «مورد بالینی کامل» که قبلاً برگزار می‌شد، ترجیح دادند (هیل و کندل ۲۰۰۷). همچنین دانشجویان این آزمون را عادلانه‌تر دانستند زیرا یک نمره ضعیف در یک مواجهه می‌توانست با نمرات دیگر مواجهه‌های آزمون mini-CEX در زمان دیگر جبران شود (هیل ۲۰۰۹). البته در مواردی نیز فراگیران به استرس‌زا بودن mini-CEX به عنوان یک روش ارزیابی اشاره داشته و معتقد بودند مشاهده عملکرد در طول آزمون بر نتایج آن تأثیر منفی دارد.

بررسی فرم‌های آزمون mini-CEX نشان داد بین رضایت ارزیابان با نمره mini-CEX، دشواری موارد بیماری و طول زمان مشاهده ارتباط مثبت وجود داشت که می‌تواند ناشی از اثر هاله‌ای یا دلایل دیگر باشد (نورسینی و همکاران ۲۰۰۳). پژوهشگران توجیه کردند احتمالاً موقعیتی که دستیاران خوب عمل می‌کنند یا با بیماران چالش‌برانگیز روبرو می‌شوند، برای ارزیابان رضایت‌بخش‌تر است. البته این امر می‌تواند منعکس‌کننده این موضوع نیز باشد که ارزیابان با دادن بازخورد

1. Yousuf
2. Torre et al.
3. Brazil et al.
4. Sidhu et al.

منفی مشکل دارند و در نتیجه از مواجهه‌هایی که در آن دستیاران خوب عمل نکرده‌اند، رضایت نداشتند. در مجموع، دادن نمره پایین توسط استادان به دانشجویانی که قبلاً با آن‌ها از نزدیک کار کرده‌اند مشکل است، خصوصاً به این دلیل که باید بازخورد فوری داده شود. در این رابطه نقل قول یکی از آزمونگران شنیدنی است: «افراد با این مسأله که کنار کسی بنشینند و بگویید عملکردت خوب و کافی نبود، مشکل دارند، خصوصاً اگر از قبل او را بشناسند» (هیل و همکاران ۲۰۰۹).

در مجموع آزمون mini-CEX روشی جدید در ارزیابی توانمندی‌های بالینی است و در نتیجه کاربرد وسیع آن مستلزم مطالعات بیشتر در مورد جنبه‌های مختلف آزمون است. از جمله مواردی که برای پژوهش بیشتر در این زمینه پیشنهاد می‌شوند عبارتند از:

- چه عوامل زمینه‌ای و عوامل مربوط به ارزیابان بر قضاوت آنان تأثیر گذار است؟ نحوه و میزان تأثیر این عوامل چگونه است؟ چگونه می توان این عوامل را به منظور ارتقای آزمون دستکاری کرد؟
- آیا آموزش ارزیابان میزان سخت‌گیری آن‌ها را یکسان می‌کند و بازخورد سازنده را ارتقاء می‌دهد؟ آیا آموزش ارزیابان ویژگی حیطه‌های صلاحیت[1] را کاهش می‌دهد؟

نورسینی و همکاران ۱۹۹۵ و نورسینی و همکاران ۲۰۰۳

یکی از اولین مطالعاتی که به بررسی ویژگی‌های روانسنجی و قابلیت اجرای آزمون mini-CEX پرداخت، مطالعه نورسینی و همکاران در سال ۱۹۹۵ همزمان با معرفی این روش ارزیابی بود. در این مطالعه ۳۸۸ مواجهه mini-CEX با ۸۸ دستیار و ۹۷ ارزیاب در پنج برنامه دستیاری طب داخلی در ایالت پنسیلوانیا مورد بررسی قرار گرفت. تعداد مواجهه برای هر دستیار بین دو تا ده (با میانگین ۴/۴ برای هر دستیار) بود. تعداد ارزیابی هر ارزیاب بین یک تا نه (با میانگین ۴ ارزیابی برای هر ارزیاب) بود.
نورسینی و همکاران در سال ۲۰۰۳ با تکمیل مطالعه قبلی خود ۱۲۲۸ مواجهه mini-CEX با ۴۲۱ دستیار و ۳۱۶ ارزیاب در ۲۱ برنامه دستیاری طب داخلی را مورد بررسی قرار دادند. نتایج این دو مطالعه نشان داد، mini-CEX:

- در موقعیت‌های مختلف بالینی متناسب با طبابت روزمره دستیاران به طور موفقیت‌آمیزی قابل اجرا است.
- قابل استفاده برای طیف وسیعی از مشکلات و شرایط بالینی است.
- یک ابزار آموزشی است، زیرا امکان تعامل دستیاران با چندین عضو هیأت علمی به عنوان الگو و همچنین امکان ارائه بازخورد را فراهم می‌آورد.

بر اساس یافته‌های این دو مطالعه:

- ارتباط آماری معنی‌داری بین اجزای مختلف توانمندی (مطالعه اول: ۰/۶۵ تا ۰/۸۱، مطالعه دوم: ۰/۶۱ تا ۰/۷۸، $P < ۰/۰۰۱$) و اجزای توانمندی با توانمندی کلی (مطالعه اول: ۰/۶۱ تا ۰/۶۸، مطالعه دوم: ۰/۷۳ تا ۰/۸۶، $P < ۰/۰۰۱$) وجود داشت.
- ارزیابان از ابزار ارزیابی رضایت داشتند (میانگین ۶ و۷ به ترتیب در مطالعه اول و دوم در مقیاس لیکرت نه‌تایی).
- رضایت ارزیابان با نمره توانمندی کلی دستیاران ارتباط مثبت داشت (۰/۲۱)، نشان‌دهنده این موضوع است که عملکرد دستیاران بر رضایت ارزیابان از ابزار ارزیابی تأثیرگذار است.
- رضایت ارزیابان با دشواری موارد بیماری ارتباط مثبت داشت (۰/۲۱).
- رضایت ارزیابان با طول زمان مواجهه ارتباط مثبت داشت (۰/۱۲).
- دستیاران از ابزار ارزیابی رضایت داشتند (میانگین ۶/۶ در مقیاس لیکرت نه‌تایی).
- رضایت دستیاران بیشتر از ارزیابان بود ($P< ۰/۰۵$)
- تفاوت معنی‌داری بین نمرات آزمون mini-CEX دستیاران طب داخلی سال اول و دوم مشاهده شد (نورسینی ۱۹۹۵).
- افزایشی در نمرات دستیاران سال اول در طول سال در همه حیطه‌های توانمندی دیده شد (نورسینی ۲۰۰۳).
- ارتباط معنی‌دار اما کوچکی بین نمره‌دهی ارزیاب و دشواری مورد بیماری مشاهده شد (۰/۱۵) که نشان‌دهنده این موضوع است که ارزیابان سعی می‌کنند تا حدی دشواری مورد بیماری را با ارفاق جبران کنند.

سؤالات رایج در مورد آزمون mini-CEX

آیا آزمون mini-CEX به عنوان ارزیابی با هدف تراکمی مناسب است؟

همان‌طور که پیشتر بیان شد، هدف اصلی این آزمون ارزیابی تکوینی و ارائه بازخورد است. با وجود این، بررسی‌ها در مقاطع پزشکی عمومی و دستیاری نشان‌دهنده کاربرد موفقیت آمیز آزمون با هدف تراکمی است. نمرات آزمون mini-CEX

1. Competency domain specificity

با تعداد مواجهه مناسب (به عنوان مثال بیش از ۱۵ مواجهه در طول یک سال در دوره پزشکی عمومی)، از دقت کافی در تمایز فراگیران با عملکرد ردی و قبولی برخوردار است. البته باید توجه داشت که این آزمون نمی‌تواند فراگیران با عملکرد متفاوت در طول طیف مقیاس نمره‌دهی را تمایز دهد بنابراین استفاده از این ابزار برای رتبه‌بندی دانشجویان مناسب نیست. آزمون mini-CEX حتی با تعداد مواجهه بالا تنها برای ارزیابی معیارمحور مناسب است یعنی جایی که لازم است دانشجویان حداقل استاندارد قابل قبول را کسب کنند (هیل و همکاران ۲۰۰۹).

در مجموع در صورتی‌که بخواهیم از این آزمون با هدف تراکمی استفاده کنیم باید نحوه تصمیم‌گیری در مورد افراد دقیقاً مشخص شود و دلایل متقن در مورد حدنصاب قبولی موجود باشد. از طرف دیگر حتی در صورت استفاده از آن به عنوان آزمون معیارمحور نیز باید تمام چالش‌ها و مشکلاتی را که در نمره‌دهی وجود دارد و در قسمت نمره‌دهی به آن اشاره شد، مد نظر داشت.

تفاوت در تعداد گزینه‌های مقیاس نمره‌دهی چه تأثیری بر ویژگی‌های روان‌سنجی آزمون mini-CEX (مخصوصاً پایایی و روایی) آن دارد؟

مطالعاتی که در خارج از حوزه آموزش پزشکی انجام شده است، این احتمال را مطرح می‌کنند که مقیاس‌های با گزینه‌های کمتر از پایایی کمتری برخوردارند. در عوض، افزایش تعداد گزینه‌ها به بیش از ظرفیت ارزیابان برای تمایز (۲ ± ۷) نیز موجب افزایش خطا در نمره‌دهی می‌شود. تأثیر تغییر تعداد گزینه‌ها بر دقت نمرات متفاوت بوده است، به طوری‌که در برخی موارد با افزایش تعداد گزینه‌ها هیچ تأثیری مشاهده نشده است و در موارد دیگر بهبود دقت نمرات گزارش شده است.

همچنین نشان داده شده است که مقیاس‌های با تعداد گزینه کمتر کارآتر هستند و استفاده از آنها آسان‌تر است. کوک و بکمن (۲۰۰۹) در مطالعه خود تفاوت دو فرم آزمون mini-CEX با مقیاس نمره‌دهی نه‌تایی و پنج‌تایی را بر پایایی بین ارزیابان و دقت نمرات در تمایز بین سطوح فراگیران مورد بررسی قرار دادند. هرچند پایایی بین ارزیابان برای هر دو نوع فرم یکی بود اما نمره‌دهی نه‌تایی نمرات دقیق‌تری فراهم کرد به نحوی که با دقت بیشتری توانست دستیاران را به داوطلبان با عملکرد غیرمطلوب یا عالی تقسیم نماید. بنابراین با وجود این‌که ارتباط بین دو نوع فرم زیاد بود اما فرم‌های نمره‌دهی نه‌تایی بهتر توانست دستیارانی را که به برنامه‌های جبرانی نیاز دارند، مشخص کند.

دوناتو و همکاران ۲۰۰۸

پژوهشگران به منظور کاهش محدودیت‌های فرم‌های معمول مورد استفاده در آزمون mini-CEX تغییراتی در این فرم‌ها به صورت زیر ایجاد کردند: کاهش مقیاس نمره‌دهی نه‌تایی به چهارتایی، افزودن توصیف و تعاریف مقیاس نمره‌دهی، کاهش توانمندی‌های مورد ارزیابی از هفت مورد به سه مورد (مهارت‌های برقراری ارتباط، دانش پزشکی و تعهد حرفه‌ای) و افزودن فضایی برای درج برنامه عملیاتی.

به منظور تسهیل ارزیابی و امکان به همراه داشتن فرم، آن را به صورت کارت کوچک قابل حمل در جیب طراحی کردند، به این صورت که فرم پشت و رو بود و با تا کردن آن از نیمه، کارت جیبی با چهار بخش درست می‌شد. سپس پژوهشگران به مقایسه فرم‌های مذکور با فرم‌های مورد استفاده توسط بورد بیماری‌های داخلی آمریکا پرداختند. فرض آن‌ها بر این بود که این تغییرات موجب افزایش دقت آزمون در تمایز دانشجویان با عملکرد نامطلوب و مطلوب و نیز ارتقای کمیت و کیفیت بازخورد ارائه شده می‌شود.

بر اساس نتایج مطالعه، فرم‌های جدید دقت کلی نمرات آزمون را افزایش داد (۸۵ درصد در مقابل ۷۳ درصد). این افزایش دقت برای فراگیران با عملکرد ضعیف بیشتر بود (۹۶ درصد در مقابل ۵۲ درصد). پایایی آزمون را در سطح ردی و قبولی و میزان مشاهدات را افزایش داد و نیز میزان توصیه‌های نوشتاری نیز افزایش یافت اما هیچ افزایشی در کمیت و کیفیت بازخورد ارائه شده مشاهده نشد. البته به دلیل تعدد مداخلاتی که در این مطالعه استفاده شده است، نمی‌توان به طور قابل اعتمادی نتیجه‌گیری کرد که کدام‌یک از تغییرات مسؤول افزایش دقت نمرات، بهبود پایایی و افزایش مشاهدات بوده‌اند.

آموزش ارزیابان چه تأثیری بر بهبود قضاوت ارزیابان و ویژگی‌های آزمون mini-CEX از جمله پایایی بین ارزیابان و روایی آن دارد؟

در این خصوص پژوهش‌ها نتایج بسیار متغیری را گزارش کرده‌اند. به عنوان مثال، اعضای هیأت علمی که در کارگاه آموزشی با هدف بهبود مهارت‌های مشاهده مستقیم و قضاوت عملکرد شرکت کردند، کارگاه را عالی ارزشیابی کردند و هشت ماه پس از آموزش در مشاهده مستقیم عملکرد فراگیران احساس راحتی بیشتری داشتند. آن‌ها سخت‌گیرتر از همکاران گروه کنترل قضاوت کردند اما پایایی بین ارزیابان نسبت به گروه کنترل که آموزشی دریافت نکردند، بهبودی نشان نداد (هولمبو و همکاران ۲۰۰۴).

کوک و همکاران (۲۰۰۸) کارآزمایی تصادفی را با هدف تأثیر آموزش ارزیابان بر بهبود پایایی بین ارزیابان و دقت نمره‌دهی انجام دادند. گروه مداخله در این پژوهش تحت کارگاه آموزشی با اقتباس از الگوی هولمبو (۲۰۰۴) قرار گرفت. پایایی بین ارزیابان و دقت نمره‌دهی در ابتدا و چهار هفته پس از کارگاه با استفاده از نوارهای ویدئویی مواجهه دستیار و بیمار ارزیابی شد. همچنین نمرات mini-CEX مربوط به ارزیابان شرکت کننده در کارگاه، ۱۲ ماه قبل و بعد از کارگاه در محیط واقعی با هم مقایسه شد. نتایج مطالعه نشان داد کارگاه تأثیر معنی‌داری بر پایایی بین‌ارزیابان نداشت. تنها اعتماد به نفس شرکت‌کنندگان در اجرای آزمون بهبود یافت. پژوهشگران موارد زیر را به عنوان دلایل عدم موفقیت ذکر کردند: کوتاه بودن مدت زمان کارگاه، عدم تأثیرپذیری ارزیابان از آموزش، ویژگی مورد آموزش‌های ارائه شده. به عنوان مثال، اگر آموزش ارزیابان در کارگاه با سناریویی در مورد مواجهه داوطلب با بیمار مبتلا به تنگی نفس صورت پذیرد، ممکن است در قضاوت آنها در ارزیابی مواجهه داوطلب با بیمار مبتلا به درد زانو اثری نداشته باشد. با توجه به نتایج مداخله پژوهشگران پیشنهاد کردند به این دلیل که هدف مهمتر آزمون mini-CEX تکوینی و ارائه بازخورد است، شاید بهتر این باشد که کارگاه‌ها بر آموزش نحوه ارائه بازخورد متمرکز شود. به عبارت دیگر باید ابتدا هدف mini-CEX را مشخص کنیم و با توجه به آن محتوا و تمرکز کارگاه را برنامه‌ریزی کنیم و در صورتی‌که هدف ارزیابی تکوینی است، بر جزء بازخورد تأکید بیشتری شود. همچنین مطالعه لیائو و همکاران (۲۰۱۳) نشان داد برگزاری کارگاه بر دانش ارزیابان در مورد mini-CEX و افزایش احتمال ارائه بازخورد به داوطلبان مؤثر است. با توجه به ناکافی بودن پژوهش‌های انجام شده در زمینه توانمندسازی ارزیابان و یکسان نبودن نتایج آن نمی‌توان استنباط قطعی در زمینه اثربخشی آن داشت.

منابع

1. Al Ansari, Ali SK, Donnon T. The Construct and Criterion Validity of the Mini-CEX: A Meta-Analysis of the Published Research. Acad Med. 2013;88(3):413-20.
2. Alves de Lima A, Barrero C, Baratta S, Castillo Costa Y, Bortman G, Carabajales J, *et al*. Validity, reliability, feasibility and satisfaction of the Mini-Clinical Evaluation Exercise (Mini-CEX) for cardiology residency training. Med Teach. 2007;29(8):785–90.
3. Alves de Lima A. Conde D, Aldunate L, Van der Vleuten CPM. Teachers' experiences of the role and function of the mini clinical evaluation exercise in postgraduate training .Int J Med Educ. 2010;1:68-73.
4. Alves de Lima A, Conde D, Costabel J, Corso J, Van der Vleuten CPM. A laboratory study on the reliability estimations of the mini-CEX. Adv Health Sci Educ. 2013;18:5–13.
5. Alves de Lima A, Henquin R, Thierer J, Paulin J, Lamari S´N, Belcastro F, *et al*. A qualitative study of the impact on learning of the mini clinical evaluation exercise in postgraduate training. Med Teach. 2005;27(1):46–52.
6. Alves de Lima A, Van der Vleuten CPM. Mini-CEX: A Method Integrating Direct Observation and Constructive Feedback for Assessing Professional Performance. Rev Argent Cardiol. 2011;79:531-6.
7. Amin Z, Seng CY, Eng KH. Practical guide to medical student assessment. World Scientific Publishing; 2006.
8. Bennett D, Kelly M, O'Flynn S. Framework for feedback: the peer mini-clinical examination as a formative assessment tool. Med Educ. 2012; 46(5):512.
9. Brazil V, Ratcliffe L, Zhang J, Davin L. Mini-CEX as a workplace-based assessment tool for interns in an emergency department – Does cost outweigh value? Med Teach. 2012;34(12):1017–23.
10. Cantillon P, Wood D. ABC of Learning and Teaching in Medicine. 2nd ed. West Sussex: John Wiely & Sons; 2010.
11. Cook DA, Beckman TJ. Does scale length matter? A comparison of nine- versus five-point rating scales for the mini-CE. Adv Health Sci Educ. 2009;14:655–64.
12. Cook DA, Beckman TJ, Mandrekar JN, Pankratz VS. Internal structure of mini-CEX scores for internal medicine residents: Factor analysis and generalizability. Adv Health Sci Educ. 2010;15:633–45.
13. Cook DA, Dupras DM, Beckman TJ, Thomas KG, Pankratz S. Effect of Rater Training on Reliability and Accuracy of Mini-CEX Scores: A Randomized, Controlled Trial. J Gen Intern Med. 2008;24(1):74–9.
14. Cruess R, McIlroy JH, Cruess S, Ginsburg S, Steinert Y. The Professionalism Mini-Evaluation Exercise: A Preliminary Investigation. Acad Med. 2006;81(10):S74–8.
15. Dewi SP, Achmad TH. Optimising feedback using the mini-CEX during the final semester pro-

gramme. Med Educ. 2010;44(5):509.
16. Dijksterhuis M, Schuwirth L, Braat D, Scheele F. What's the problem with the mini-CEX? Med Educ. 2011;45(3):317–9.
17. Donato AA, Pangaro L, Smith C, et al. Evaluation of a novel assessment form for observing medical residents: A randomized, controlled trial. Med Educ. 2008;42(12):1234–42.
18. Durning SJ, Cation LJ, Markert RJ, Pangaro LN. Assessing the reliability and validity of the mini-clinical evaluation exercise for internal medicine residency training. Acad Med. 2002;77(9):900–4.
19. Fernando N, Cleland J, McKenzie H, Cassar K. Identifying the factors that determine feedback given to undergraduate medical students following formative mini-CEX assessments. Med Educ. 2008;42(1):89–95.
20. Golnik KC, Goldenhar LM, Gittinger JW Jr, Lustbader JM. The Ophthalmic Clinical Evaluation Exercise (OCEX). Ophthalmol. 2004;111(7):1271– 4.
21. Golnik KC, Goldenhar LM. The Ophthalmic Clinical Evaluation Exercise Reliability Determination. Ophthalmol. 2005;112(10):1649– 54.
22. Han PKJ, Keranen LB, Lescisin DA, Arnold RM. The Palliative Care Clinical Evaluation Exercise (CEX): An Experience-Based Intervention for Teaching End-of-Life Communication Skills. Acad Med. 2005; 80(7):669–76.
23. Hatala R, Ainslie M, Kassen BO, Mackie I, Roberts JM. Assessing the mini-clinical evaluation exercise in comparison to a national specialty examination. Med Educ. 2006;40(10):950–6.
24. Hawkins RE, Margolis MJ, Durning SJ, Norcini JJ. Constructing a validity argument for the mini-clinical evaluation exercise: A review of the research. Acad Med. 2010; 85(9):1453–61.
25. Hill F, Kendall K. Adopting and adapting the mini-CEX as an undergraduate assessment and learning tool. Clin Teach 2007;4(4):244–8.
26. Hill F, Kendall K, Galbraith K, Crossley J. Implementing the undergraduate mini-CEX: A tailored approach at Southampton University. Med Educ. 2009;43(4):326–34.
27. Holmboe ES, Hawkins RE, Huot SJ. Effects of Training in Direct Observation of Medical Residents' Clinical Competence. Ann Intern Med. 2004;140(11):874-81.
28. Holmboe ES, Huot S, Chung J, Norcini J, Hawkins RE. Construct validity of the miniclinical evaluation exercise (miniCEX). Acad Med. 2003;78(8):826–30.
29. Holmboe ES, Yepes M, Williams F, Huot SJ. Feedback and the Mini Clinical Evaluation Exercise. J Gen Intern Med. 2004;19:558–61.
30. Holmboe ES, Hawkins RE. Practical guide to the evaluation of clinical competence. Philadelphia: Mosby/Elsevier; 2008.
31. Kogan JR, Bellini LM, Shea JA. Feasibility, Reliability, and Validity of the Mini-Clinical Evaluation Exercise (mCEX) in a Medicine Core Clerkship. Acad Med. 2003;78(10):S33–5.
32. Kogan JR, Conforti LN, Bernabeo EC, Durning SJ, Hauer KE, Holmboe ES. Faculty staff perceptions

of feedback to residents after direct observation of clinical skills. Med Educ. 2012;46(2):201–15.

33. Kogan JR, Hess BJ, Conforti LN, Holmboe ES. What Drives Faculty Ratings of Residents' Clinical Skills? The Impact of Faculty's Own Clinical Skills. Acad Med. 2010;85(10):S25–8.
34. Liao KC, Pul SJ, Liu1 MS, Yang CW, Kuo HP. Development and implementation of a mini-Clinical Evaluation Exercise (mini-CEX) program to assess the clinical competencies of internal medicine residents: from faculty development to curriculum evaluation. BMC Med Educ. 2013;13:31.
35. Malhotra S, Hatala R, Courneya C. internal medicine residents' perceptions of the Mini-Clinical Evaluation Exercise. Med Teach. 2008;30(4):414–19.
36. Margolis MJ, Clauser BE, Cuddy MM, et al. Use of the mini-clinical evaluation exercise to rate examinee performance on a multiplestation clinical skills examination: A validity study. Acad Med. 2006;81(10):S56–60.
37. Morris A, Hewitt J, Roberts CM. Practical experience of using direct observed procedures, mini clinical examinations and peer observation in pre-registration house officers (FY1) trainees. Postgrad Med J. 2006;82(966):285–8.
38. Nair BR, Alexander HG, McGrath BP, Parvathy MS, Kilsby EC, et al. The mini clinical evaluation exercise (mini-CEX) for assessing clinical performance of international medical graduates. MJA. 2008;189(3):159-61.
39. Norcini JJ, Blank LL, Arnold GK, Kimball HR. The mini-CEX (clinical evaluation exercise): A preliminary investigation. Ann Intern Med. 1995;123(10):795–9.
40. Norcini JJ, Blank LL, Duffy FD, Fortna GS. The mini-CEX: A method for assessing clinical skills. Ann Intern Med. 2003;138(6):476–81.
41. Norcini J, Burch V. Workplace-based assessment as an educational tool: AMEE Guide No. 31. Med Teach. 2007;29(9-10):855–71.
42. Pelgrim EAM, Kramer AWM, Mokkink HGA, van den Elsen L, Grol RPTM, van der Vleuten CPM. In-training assessment using direct observation of single-patient encounters: A literature review. Adv Health Sci Educ. 2011;16:189–99.
43. Pelgrim EAM, Kramer AWM, Mokkink HGA, van der Vleuten CPM. Quality of written narrative feedback and reflection in a modified mini-clinical evaluation exercise: an observational study. BMC Med Educ. 2012;12(97).
44. Senta Z, Jha V, Boursicot KAM., Roberts TE. Evaluating the utility of workplace-based assessment tools for speciality training. Best Pract Res Cl Ob. 2010;24(6):767–782.
45. Sidhu RS, Hatala R, Barron S, Broudo M, Pachev G, Page G. Reliability and Acceptance of the Mini-Clinical Evaluation Exercise as a Performance Assessment of Practicing Physicians. Acad Med. 2009;84(10):S113–5.
46. Singh T, Sharma M. Mini-clinical examination (CEX) as a tool for formative assessment. Natl Med J India. 2010;23(2):100-2.

47. Torre DM, Simpson DE, Elnicki DM, Sebastian JL, Holmboe ES. Feasibility, reliability, and user satisfaction with a PDA-based mini-CEX to evaluate the clinical skills of third-year medical students. Teach Learn Med. 2007;19(3):271–7.
48. Walsh K, Jaye P. The costs and utility of the Mini-CEX. 2013;35(9):789.
49. Weller JM, Jolly B, Misur MP, Merry AF, Jones A, Crossley JGM, *et al*. Mini-clinical evaluation exercise in anaesthesia training. Br J Anaesth. 2009;102(5):633–41.
50. Wilkinson JR, Crossley JGM, Wragg A, Mills P, Cowan G, Wade W. Implementing workplace-based assessment across the medical specialties in the United Kingdom. Med Educ. 2008;42(4):364–73.
51. Yousuf N. Mini clinical evaluation exercise: validity and feasibility evidences in literature. Educ in Med J. 2012;4(1):101–7.
52. Swanwick T. Uderstanding Medical Education: Evidence, Theory and Practice. West Sussex: John Wiely & Sons; 2010

فصل ۲۲

آزمون DOPS

ساختار آزمون DOPS

به طور سنتی، توانمندی دستیاران رشته‌های تخصصی مختلف پزشکی در پروسیجرهای عملی با استفاده از روش‌هایی مانند نظر کلی استادان در انتهای چرخش بالینی[1]، ممیزی پیامدهای بالینی[2] یا با استفاده از لاگ‌بوک ارزیابی می‌شود. این روش‌ها در ارزیابی پروسیجرها محدودیت‌هایی دارند. در نظردهی کلی در پایان چرخش بالینی، عملکرد فراگیر به صورت ذهنی و بدون معیار مشخص مورد قضاوت قرار می‌گیرد و در نتیجه پایایی آن بسیار پایین است. ممیزی پیامدهای بالینی پیامدهای غیرمستقیم مربوط به مهارت‌های پروسیجرال مانند میزان مرگ‌ومیر را اندازه‌گیری می‌کند. این روش گذشته‌نگر است و در زمانی که ممیزی در حال اجرا است، پیامدهای غیرمطلوب رخ داده است که می‌تواند باعث سوگیری شود. لاگ‌بوک تنها بیانگر انجام پروسیجر و کمیت آن است و توانمندی فراگیر در انجام پروسیجر را نشان نمی‌دهد. هدف لاگ‌بوک اطمینان از انجام حداقل تعداد مورد انتظار پروسیجر برای تبدیل شدن به یک پزشک توان‌مند است و در آن فرایند پروسیجر مورد مشاهده قرار نمی‌گیرد و بازخوردی نیز نسبت به عملکرد فراگیر ارائه نمی‌شود.

در پاسخ به این محدودیت‌ها کالج سلطنتی پزشکان انگلیس برای اولین بار در سال ۲۰۰۳ روش ارزیابی DOPS را طراحی کرد و آن را به صورت آزمایشی در برنامه پیش‌دستیاری استفاده کرد. سپس در سال ۲۰۰۵ این روش ارزیابی به صورت رسمی وارد برنامه پیش‌دستیاری شد. در واقع، DOPS بسیار شبیه mini-CEX است و می‌توان آن را نوعی mini-CEX دانست که به طور خاص عملکرد فراگیران در پروسیجرهای عملی بالینی را ارزیابی می‌کند. ارزیابی مهارت‌های پروسیجرال با ابزارهای مشاهده‌ای به این دلیل اهمیت دارد که نقص‌های موجود در مهارت‌ها را مشخص می‌کند و بلافاصله بازخورد ارائه می‌شود.

آزمون DOPS یکی از انواع روش‌های ارزیابی مبتنی بر محل کار است که در آن مهارت‌های پروسیجرال فراگیران پزشکی مورد مشاهده مستقیم و ارزیابی قرار می‌گیرد. در این آزمون یکی از اعضای هیأت علمی مهارت فراگیر را در انجام یک مهارت پروسیجرال روی بیمار واقعی و در محیط بالینی به مدت ۱۵ تا ۲۰ دقیقه ارزیابی می‌کند. پس از مشاهده عملکرد فراگیر با استفاده از مقیاس درجه‌بندی به هر کدام از سؤالات فرم ارزیابی نمره داده می‌شود. هدف اصلی در این روش، ارائه بازخورد بر اساس عملکرد مشاهده‌شده است و در انتها، ۵ تا ۱۰ دقیقه به ارائه بازخورد به فراگیر اختصاص می‌یابد.

تعداد بسیار زیادی از پروسیجرها در رشته‌های تخصصی بالینی مختلف می‌توانند با DOPS ارزیابی شوند. هر مواجهه

1. Supervisor end of rotation rating
2. Clinical outcome audit

می‌بایست برای پوشش مهارت‌های پروسیجرال ضروری برنامه درسی آن رشته و مقطع باشد. این آزمون برای ارزیابی پروسیجرهای کوتاه تشخیصی و درمانی یا بخشی از یک پروسیجر که از تعداد مراحل اندکی تشکیل شده است، مناسب‌تر است.

مواجهه می‌تواند در انواع مختلفی از موقعیت‌های بالینی شامل درمانگاه، اورژانس، بخش بیماران بستری، اتاق عمل و محیط‌های دیگر رخ دهد و طیف وسیعی از پروسیجرهای بالینی را پوشش دهد. نکته قابل توجه این است که در آزمون DOPS پروسیجرها باید بر روی بیمار واقعی و نه مدل‌ها، شبیه‌سازها و جسد انجام شود.

در آزمون DOPS نیز مانند mini-CEX انتظار می‌رود عملکرد فراگیر در طول یک سال، در انجام چند پروسیجر و با استفاده از ارزیابان متفاوت ارزیابی شود. فراگیران معمولاً شش بار یا بیشتر در طول سال ارزیابی می‌شوند.

مزایا و محدودیت‌های آزمون DOPS

مزایای آزمون DOPS

- امکان مشاهده مستقیم مهارت‌های پروسیجرال را فراهم می‌کند.
- امکان ارزیابی کلی عملکرد فراگیر در انجام یک پروسیجر را فراهم می‌سازد.
- کاربردی است و استفاده از آن آسان است.
- امکان تطبیق آن با توجه به شرایط و نیازهای موجود وجود دارد.

محدودیت‌های آزمون DOPS

- تا حدودی مفهوم جدید و ناآشنایی برای اعضای هیأت علمی است و در نتیجه نیاز به آموزش دارد.
- امکان ارزیابی تمام قسمتهای مختلف یک مهارت از طریق یک مواجهه بالینی منفرد وجود ندارد.
- در صورتی که یک پروسیجر، ماهیت تکنیکی و فنی داشته باشد، ممکن است حضور مشاهده کننده یا ارزیاب متخصص ضروری باشد.

ضرورت و کاربرد آزمون DOPS

یکی از انتظارات جامعه، رعایت موارد مربوط به ایمنی بیمار توسط پزشکان است. در برخی از رشته‌های تخصصی پزشکی مانند جراحی و رشته‌های دیگری مانند دندانپزشکی به دلیل شیوع بیشتر پروسیجرهای لازم این مورد محسوس‌تر است. بنابراین، یکی از مهمترین جنبه‌های طبابت پزشکی توانایی انجام پروسیجرهای مختلف به صورت مؤثر و ایمن است. این امر مستلزم مهارتی فراتر از انجام دقیق پروسیجر از لحاظ تکنیکی است و مواردی مانند تصمیم‌گیری، کار تیمی، مهارت‌های ارتباطی و تعهد حرفه‌ای را در بر می‌گیرد. آزمون DOPS به منظور مشاهده و ارزیابی همه توانمندی‌های مربوط به انجام پروسیجرها و نه تنها تکنیک انجام آن طراحی شد. با توجه به این‌که بررسی‌ها نشان می‌دهند در برنامه‌های آموزشی فعلی، زمان کافی برای آموزش پروسیجرهای بالینی اختصاص نمی‌یابد و انجام مهارت‌های پروسیجرال به صورت روتین مورد مشاهده قرار نمی‌گیرند، به نظر می‌رسد اجرای آزمون DOPS بتواند زمینه‌ای را برای رفع این مشکل فراهم آورد.

همانند دیگر روش‌های ارزیابی مبتنی بر محل کار، هدف اولیه و ذاتی این آزمون ارائه بازخورد و در نتیجه افزایش یادگیری است. البته این آزمون هم‌اکنون در بسیاری از مؤسسات آموزشی سراسر دنیا با اهداف تکوینی و تراکمی استفاده می‌شود. یک از ویژگی‌های DOPS این است که در رشته‌های مختلف علوم پزشکی مانند دندانپزشکی، پرستاری و مامایی، دامپزشکی و پیراپزشکی که با آموزش و ارزیابی مهارت‌های پروسیجرال سرو کار دارند کاربرد دارد. به عنوان مثال، در

دانشکده دامپزشکی دانشگاه ناتینگهام[1] دانشجویان سال آخر ملزم به گذراندن ده آزمون DOPS در پروسیجرهای متنوع و با استفاده از انواع مختلف حیوانات هستند. دانشجویان باید در همه ده مورد توانمند ارزیابی شوند تا بتوانند در امتحان کتبی چندگزینه‌ای آخر سال شرکت کنند. در دانشگاه بیرمینگهام[2] نیز در دوره دستیاری بیهوشی، دستیاران موظف به تکمیل ۷۵ مورد DOPS در پروسیجرهای مختلف در کل دوره ۵ ساله دستیاری هستند.

گام‌های طراحی و اجرای آزمون DOPS مطلوب

با توجه به این‌که آزمون DOPS بسیار شبیه آزمون mini-CEX است بسیاری از اصول و قواعدی که در طراحی و اجرای این آزمون باید رعایت شود در فصل قبل در قسمت مربوط به طراحی و اجرای آزمون mini-CEX ذکر شده است. در این‌جا تنها به مواردی که به طور خاص به آزمون DOPS مربوط می‌شود اشاره می‌شود. خلاصه این موارد در جدول ۱-۲۲ آمده است.

جدول ۱-۲۲: خلاصه مراحل طراحی و اجرای آزمون DOPS

ردیف	عنوان	توضیح
۱	تهیه بلوپرینت آزمون	در تهیه بلوپرینت آزمون DOPS پروسیجرهای شایع و مهم (نجات‌دهنده جان بیمار) مدنظر قرار می‌گیرند. اگر چه با تغییر و ارتقای دانش و نیز فنآوری پزشکی، فهرست مهارت‌های پروسیجرال ضروری تغییر خواهند کرد.
۲	تعیین معیارهای ارزیابی پروسیجر	در اکثر فرم‌های آزمون DOPS معیارهایی انجام آمادگی‌های قبل از انجام پروسیجر، توانایی فنی و تکنیکی انجام پروسیجر، درخواست کمک در صورت نیاز، مدیریت بعد از انجام پروسیجر، مهارت‌های ارتباطی و توجه به بیمار/ حرفه‌ای‌گری لحاظ می‌شوند.
۳	تهیه فرم‌های ارزیابی	فرم‌های مورد استفاده در آزمون DOPS ممکن است به صورت کلی طراحی شوند و در ارزیابی مهارت‌های پروسیجرال مختلف مورد استفاده واقع شوند یا فرم‌های مخصوص ارزیابی یک پروسیجر خاص طراحی شوند. نمره‌دهی فرم‌های اصلی آزمون DOPS به صورت مقیاس نمره‌دهی لیکرت است اما فرم‌های چک‌لیست نیز وجود دارد.
۴	تعیین حداقل سطح قابل‌قبول عملکرد در هر حیطه	در این آزمون نیز باید علاوه بر پروسیجرهای مورد سنجش و تعریف آن، معیارهای نمره‌دهی در هر یک از حیطه‌ها و حداقل مورد انتظار از داوطلب در هر سطح نمره‌دهی تعریف شود.
۵	تصمیم‌گیری در مورد تعداد و ویژگی ارزیابان	ارزیابان می‌توانند از میان اعضای هیأت علمی، دستیار تخصصی و یا فوق‌تخصصی، پزشک عمومی یا پرستار که در پروسیجر مورد ارزیابی مهارت دارند باشند.
۶	اطلاع‌رسانی، آشناسازی و آموزش ارزیابان و داوطلبان	توجیه و آموزش ارزیابان و توجیه و آشنایی داوطلبان با فرایند آزمون از دیگر موارد مهم قبل از طراحی آزمون است.
۷	اجرای آزمون	هماهنگی یا درخواست انجام آزمون می‌تواند از طرف داوطلب یا ارزیاب باشد. مدت زمان انجام آزمون بستگی به نوع پروسیجر و پیچیدگی آن دارد و معمولاً طول مدت بازخورد یک سوم زمان مشاهده است. نکته بسیار مهم در انجام DOPS رعایت ایمنی بیمار است و ارزیابان و فراگیران باید تمام تلاش خود را در این زمینه به کار بندند.
۸	بررسی کیفیت آزمون برگزارشده	ضروری است کیفیت آزمون‌های برگزار شده با استفاده از روش‌های کمی و کیفی بررسی شود.

1. Nottingham
2. Birmingham

تهیه بلوپرینت آزمون

موارد مهم در تهیه بلوپرینت آزمون DOPS شامل انتخاب پروسیجرهای مورد سنجش و تصمیم‌گیری در مورد تعداد آن است. به طور معمول، پروسیجرها از مواردی که در توانمندی‌های پایه و اصلی برنامه آموزشی دوره مربوط آمده است، انتخاب می‌شوند. به عنوان مثال، در ارزیابی مهارت‌های پروسیجرال فراگیران در سال اول برنامه پیش‌دستیاری در کشور انگلیس از پروسیجرهایی که در توانمندی‌های پایه در سند پزشکان فردا[1]، مربوط به شورای پزشکی عمومی[2] آمده است استفاده می‌شود. دیدگاه کلی این است که پروسیجرهای شایع و مهم (نجات‌دهنده جان بیمار) انتخاب شود. هر چند، ارزیابی مانورهای نجات‌دهنده جان بیمار مانند کریکوتیروتومی مورد مناقشه است. علاوه بر موارد مذکور، باید به این نکته توجه داشت که با تغییر و ارتقای دانش و نیز فنآوری پزشکی، فهرست مهارت‌های پروسیجرال ضروری تغییر خواهند کرد.

پروسیجرهای مورد بررسی می‌توانند از موارد بسیار ساده مانند خون‌گیری تا پروسیجرهای بسیار پیچیده مانند ترمیم ضایعه با فلاپ پوستی، تحت بی‌حسی موضعی متفاوت باشد. به طور ایده‌ال، عملکرد فراگیران در طول یک سال در پروسیجرهای مختلفی ارزیابی می‌شود. به منظور پوشش اهداف دوره و همچنین دستیابی به پایایی قابل قبول، لازم است برای هر داوطلب تعداد مناسب غیرتکراری از پروسیجرهای متنوع را در نظر گرفت. برنامه درسی جراحی بین دانشکده‌ای[3] ارائه شواهد از ارزیابی حداقل سه پروسیجر متفاوت را در طول سال الزامی می‌داند. در برنامه پیش‌دستیاری انگلیس، انجام شش تا هشت مواجهه در یکسال با ارزیابان متفاوت الزامی است.

پروسیجرهای مورد انتظار در ارزیابی پزشکان در برنامه پیش‌دستیاری در کشور انگلیس

- رگ‌گیری
- کشت خون (محیطی)
- تزریق داخل وریدی
- نمونه‌گیری خون شریانی
- تزریق داخل پوستی
- تزریق وریدی
- مراقبت راه هوایی
- خون‌گیری
- کشت خون (مرکزی)
- ECG
- تزریق زیر جلدی
- تزریق عضلانی
- سوندگذاری مثانه
- گذاشتن لوله معده

تعیین معیارهای ارزیابی پروسیجر

در اکثر فرم‌های آزمون DOPS و از جمله فرم‌های مورد استفاده در برنامه پیش‌دستیاری از یازده معیار برای ارزیابی پروسیجر استفاده می‌کنند که یک مورد به اررزیابی توانایی کلی در انجام پروسیجر و ده مورد دیگر به ارزیابی پروسیجر از جنبه‌های زیر می‌پردازد:

- درک موارد اندیکاسیون، آناتومی و روش انجام پروسیجر
- کسب رضایت آگاهانه
- انجام آمادگی های قبل از انجام پروسیجر
- استفاده از مواد آرامبخش و بی‌حس کننده ایمن
- توانایی فنی و تکنیکی
- تکنیک ضدعفونی
- درخواست کمک در صورت نیاز
- مدیریت بعد از انجام پروسیجر

1. Tomarrow's doctors
2. General Medical Council (GMC)
3. Intercollegiate Surgical Curriculum Programme (ISCP)

- مهارتهای ارتباطی
- توجه به بیمار/ حرفه‌ای‌گری

موارد فوق می‌توانند به منظور استفاده در ارزیابی پروسیجرهای خاص تغییر کنند. در زیر نمونه‌هایی از این معیارهای تغییریافته ذکر شده‌اند:

آزمون DOPS در رشته ارتوپدی

هاردن در سال ۲۰۱۰ برای ارزیابی مدیریت پروسیجرهای جراحی ماژور در ارتوپدی مانند همارتروسکوپی برای شکستگی داخل کپسولی گردن فمور یا جراحی ثابت کردن شکستگی Weber B مچ پا معیارهای ارزیابی زیر را پیشنهاد کرد:

- اخذ رضایت
- آمادگی قبل از عمل
- تکنیک‌های حین عمل
- برنامه‌ریزی قبل از عمل
- باز کردن و بستن
- مراقبت بعد از عمل

آزمون DOPS در رشته گوارش

آزمون DOPS هم اکنون به طور وسیعی در انگلیس مورد استفاده قرار می‌گیرد. یکی از موارد استفاده رایج آن در ارزیابی مهارت‌های پروسیجرال اندوسکوپی فوقانی و تحتانی و ERCP[1] است. گروه مشترک مشاوره‌ای (JAG)[2] اندوسکوپی دستگاه گوارش که مسؤولیت برنامه آموزش اندوسکوپی را بر عهده دارد، اقدام به تهیه فرم‌های متنوع برای ارزیابی پروسیجرهای مختلف اندوسکوپی کرده است. در این فرم‌ها چهار حیطه توانمندی کلی زیر در همه فرم‌ها به صورت مشترک وجود دارد:

- ارزیابی بیمار، کسب رضایت و برقراری ارتباط
- رعایت ایمنی و بیحسی
- مهارت انجام اندوسکوپی
- توانایی تشخیص و درمان

سپس با توجه به پروسیجر خاص به عنوان مثال ERCP یا کولونوسکوپی هر یک از این معیارها به موارد جزئی‌تری تقسیم می‌شود. به طوری‌که در برخی موارد مانند ERCP این آیتم‌ها به بیش از ۲۰ مورد می‌رسد. گروه JAG حتی فرم‌های DOPS جداگانه‌ای با توجه به هدف آزمون (تکوینی یا تراکمی) تهیه کرده است.

1. Endoscopic Retrograde Cholangiopancreatography
2. Joint Advisory Group (JAG)

آزمون DOPS در رشته بیهوشی

دلفینوو همکاران (۲۰۱۳) به روشی کاملاً نظام‌مند فرم‌های آزمون DOPS مورد استفاده در برنامه پیش‌دستیاری را برای استفاده در برنامه دستیاری بیهوشی و ارزیابی پروسیجر لوله‌گذاری در کشور شیلی اصلاح کردند. معیارهای فرم‌ها پس از اصلاح به صورت زیر تغییر یافت:

- درک کامل و جامع از موارد اندیکاسیون، آناتومی و تکنیک انجام پروسیجر
- انجام آمادگی‌های کافی شامل مواد/وسایل مورد استفاده قبل از انجام پروسیجر
- آگاهی از وضعیت بیمار در طول انجام پروسیجر
- تکنیک ضدعفونی
- درخواست کمک در صورت نیاز
- مدیریت بعد از انجام پروسیجر
- تشخیص و درمان عوارض
- مهارت‌های ارتباطی
- در نظر گرفتن راحتی و ایمنی بیمار
- احترام به پرسنل اتاق عمل
- تعهد حرفه‌ای
- توانایی کلی در انجام پروسیجر

تهیه فرم‌های ارزیابی

با توجه به این‌که فرم مورد استفاده در آزمون DOPS بسیار شبیه فرم‌های mini-CEX است، برای طراحی فرم‌های آن

به قسمت طراحی آزمون mini-CEX در فصل سوم مراجعه کنید. نمونه‌ای از فرم‌های مورد استفاده در آزمون DOPS برنامه پیش‌دستیاری در شکل ۱-۲۲ آمده است. همان‌طور که مشخص است فرم‌های طراحی‌شده در برنامه پیش‌دستیاری «کلی» است و می‌تواند در ارزیابی مهارت‌های پروسیجرال مختلف مورد استفاده واقع شود. این در حالی است که بسیاری از طراحان آزمون این دغدغه را دارند که فرم مذکور نتواند مهارت‌های اختصاصی مربوط به پروسیجرهای مختلف را ارزیابی کند. از این‌رو در بسیاری از رشته‌های دستیاری تخصصی و فوق‌تخصصی فرم‌های مخصوص ارزیابی یک پروسیجر خاص مانند لوله‌گذاری داخل تراشه (دلفینو و همکاران ۲۰۱۳) طراحی شده است. ویلکینسون و همکاران (۲۰۰۸) هم‌زمان هر دو نوع فرم کلی و مخصوص هر پروسیجر را در برنامه پیش‌دستیاری به کار بردند و همبستگی خوبی (۰r =/۸۴) بین این دو نوع فرم گزارش کردند.

هرچند نمره‌دهی فرم‌های اصلی آزمون DOPS به صورت مقیاس نمره‌دهی لیکرت است اما فرم‌های چک‌لیست نیز وجود دارد. چک‌لیست به ویژه در ارزیابی یک پروسیجر خاص بیشتر مورد استفاده قرار گرفته است. در این ابزارها یک پروسیجر به گام‌های تشکیل‌دهنده آن شکسته می‌شود (تحلیل وظیفه) و با استفاده از چک‌لیست نمره‌دهی می‌شود (احمد و همکاران[1] ۲۰۱۱).

تعیین حداقل سطح قابل قبول عملکرد در هر حیطه

همانند آنچه در مورد آزمون mini-CEX بیان شد، در این آزمون نیز باید علاوه بر پروسیجرهای مورد سنجش و تعریف آن، معیارهای نمره‌دهی در هر یک از حیطه‌ها و حداقل مورد انتظار از داوطلب در هر سطح نمره‌دهی تعریف شود.

تصمیم‌گیری در مورد تعداد و ویژگی ارزیابان

ارزیابان می‌توانند از میان اعضای هیأت علمی، دستیار تخصصی و یا فوق‌تخصصی، پزشک عمومی یا پرستار باشند. از ملزومات ارزیاب این است که اطلاعات و مهارت کافی هم در انجام آزمون DOPS و هم مهارت مورد ارزیابی داشته باشد. همچنین ارزیاب باید از برنامه آموزشی که فراگیر دریافت کرده است و سطح وی مطلع باشد، به طوری که عملکرد دانشجوی مرزی را در ذهن داشته باشد. در مورد تعداد ارزیابان نیز به نظر می‌رسد تغییرپذیری نمرات آزمون DOPS برخلاف آزمون mini-CEX بیشتر تحت‌تاثیر تفاوت قضاوت ارزیابان قرار دارد تا ویژگی موارد. بنابراین استفاده از تعداد بیشتر ارزیابان توصیه می‌شود تا افزایش تعداد موارد بیماری. هر چند هنوز در خصوص تعداد ارزیابان مطالعات و شواهد قوی در اختیار نیست.

اطلاع‌رسانی، آشناسازی و آموزش ارزیابان و داوطلبان

موارد دیگری که قبل از اجرای آزمون باید در نظر داشت، توجیه و آموزش ارزیابان و توجیه و آشنایی داوطلبان با فرایند آزمون می‌باشد. این موارد به طور مفصل در فصل قبل مورد بحث قرار گرفته است.

اجرای آزمون DOPS

بسیاری از موارد اجرا، شبیه آزمون mini-CEX است و در نتیجه از ذکر آنها صرف‌نظر می‌شود. در این‌جا تنها به ذکر نکات مهم می‌پردازیم. هماهنگی یا درخواست انجام آزمون می‌تواند از طرف داوطلب یا ارزیاب باشد اما معمولاً فراگیر است که ارزیاب، نوع پروسیجر و زمان انجام آن را (در چارچوب برنامه درسی) انتخاب می‌کند. مدت زمان انجام آزمون بستگی به نوع پروسیجر و پیچیدگی آن دارد و معمولاً طول مدت بازخورد یک سوم زمان مشاهده است. نکته بسیار مهم در انجام DOPS رعایت ایمنی بیمار است و ارزیابان و فراگیران باید تمام تلاش خود را در این زمینه به کار بندند. در مواردی که احتمال آسیب به بیمار به دلیل برگزاری آزمون وجود دارد باید از اجرای آن اجتناب ورزید. در صورت برگزاری آزمون در چنین موقعیتی، ارزیاب مسؤول حفظ امنیت بیمار است و در صورت لزوم باید مداخله کند.

1. Ahmed et al.

Direct Observation of Procedural Skill (DOPS)- F_1 Version

لطفا با گذاشتن علامت ضربدر مقابل سؤالات فرم را کامل کنید: ☒ لطفا از قلم مشکی برای تکمیل فرم استفاده کنید.

نام خانوادگی داوطلب	
نام داوطلب	
شماره GMC (دانشجویی)	ثبت شماره GMC الزامی است
محیط بالینی	درمانگاه ☐ بستری ☐ پذیرش اورژانس ☐ پذیرش جراحی ☐
شماره پروسیجر	موارد دیگر
موقعیت ارزیاب	موارد دیگر

دفعات انجام DOPS توسط ارزیاب با هر فراگیر	۰	۱	۲	۳	٤	۵-۹	۹<
	☐	☐	☐	☐	☐	☐	☐

تعداد دفعات انجام پروسیجر توسط فراگیر	۰	۱ تا ٤	۵ تا ۹	۱۰ <	پیچیدگی پروسیجر	کم	متوسط	زیاد
	☐	☐	☐	☐		☐	☐	☐

لطفا حیطه‌های زیر را با استفاده از نمره‌دهی مقابل نمره‌دهی کنید	زیر حد انتظار برای F_1		مرزی	در حد انتظار برای F_1	بالاتر از حد انتظار برای F_1		بدون نمره*
۱. درک موارد اندیکاسیون، آناتومی و روش انجام پروسیجر	☐	☐	☐	☐	☐	☐	☐
۲. کسب رضایت آگاهانه	☐	☐	☐	☐	☐	☐	☐
۳. انجام آمادگی‌های قبل از انجام پروسیجر	☐	☐	☐	☐	☐	☐	☐
٤. استفاده از مواد آرام‌بخش و بی‌حس‌کننده ایمن	☐	☐	☐	☐	☐	☐	☐
۵. توانایی فنی و تکنیکی	☐	☐	☐	☐	☐	☐	☐
٦. تکنیک ضدعفونی	☐	☐	☐	☐	☐	☐	☐
۷. درخواست کمک در صورت نیاز	☐	☐	☐	☐	☐	☐	☐
۸. مدیریت بعد از انجام پروسیجر	☐	☐	☐	☐	☐	☐	☐
۹. مهارت‌های ارتباطی	☐	☐	☐	☐	☐	☐	☐
۱۰. توجه به بیمار/ حرفه‌ای‌گری	☐	☐	☐	☐	☐	☐	☐
۱۱. توانایی کلی در انجام پروسیجر	☐	☐	☐	☐	☐	☐	☐

* بدون نمره: لطفا این مورد را در صورتی علامت بزنید که مهارت مورد نظر مشاهده نشده است و در نتیجه قادر به نمره‌دهی نیستید.

موارد مثبت عملکرد	**موارد پیشنهادی برای ارتقای عملکرد**
اقدامات مورد توافق:	
آیا در مورد این روش تاکنون آموزش دیده‌اید؟	چهره به چهره ☐ مطالعه دستورالعمل‌ها ☐ اینترنت/CD ☐
امضای ارزیاب	تاریخ — زمان مشاهده (دقیقه) — زمان بازخورد (دقیقه)
نام خانوادگی ارزیاب	
شماره پرسنلی ارزیاب	

شکل ۱-۲۲: فرم DOPS طراحی‌شده توسط نظام ملی سلامت انگلیس برای استفاده در برنامه پیش‌دستیاری (برای اطلاعات بیشتر در مورد این فرم و جزئیات توانمندی‌های مورد انتظار به سایت WWW.HCAT.NHS.UK مراجعه کنید).

بررسی کیفیت آزمون برگزارشده

اقدامات مربوط به مرحله پس از اجرا نیز بسیار شبیه دیگر آزمون‌های مبتنی بر محل کار است و بنابراین از بیان آن خودداری می‌شود. در این بخش به نمونه‌ای از پرسشنامه بررسی میزان رضایت که توسط دلفینو و همکاران (۲۰۱۳) استفاده شده است اشاره می‌شود. پرسشنامه فوق به بررسی میزان رضایت داوطلبان و ارزیابان از آزمون DOPS برگزارشده در دستیاری بیهوشی به منظور ارزیابی مهارت لوله‌گذاری داخل نای می‌پردازد. پرسشنامه مشتمل بر نه سؤال است و در مقیاس لیکرت پنج‌تایی از «کاملاً موافقم» تا «کاملاً مخالفم» تهیه شده است (جدول ۲-۲۲).

جدول ۲-۲۲: فرم نظرخواهی از داوطلبان و ارزیابان (دلفینو و همکاران ۲۰۱۳)

ردیف	سؤالات	کاملاً موافقم	موافقم	تا حدودی موافقم	نه موافقم و نه مخالفم	تا حدودی مخالفم	مخالفم	کاملاً مخالفم
۱	آزمون DOPS الزامات لازم برای ارزیابی هر پروسیجر عملی را دارد.							
۲	تعریف سطوح عملکرد مورد انتظار واضح بود.							
۳	تعریف سطوح عملکرد مورد انتظار واقعی بود.							
۴	آزمون DOPS در تشخیص ضعف در پروسیجر لوله‌گذاری داخل تراشه مفید بود.							
۵	آزمون DOPS زمان‌بر بود.							
۶	استفاده از آزمون DOPS می‌تواند تکنیک پروسیجر لوله‌گذاری داخل تراشه را بهبود بخشد.							
۷	آزمون DOPS به طور عادلانه پروسیجر لوله‌گذاری داخل تراشه را می‌سنجد.							
۸	آزمون DOPS سطوح توانمندی در پروسیجر لوله‌گذاری داخل تراشه را اندازه می‌گیرد.							
۹	من تمایل دارم آزمون DOPS را در ارزیابی پروسیجرهای دیگر به کار بندم.							

سودمندی آزمون DOPS

آزمون DOPS ابزار نسبتاً جدیدی است و با اینکه به طور گسترده‌ای در انگلیس استفاده شده است، اطلاعات کمی در مورد سودمندی آن منتشر شده است. هر چند به عنوان یک واریاسیون از آزمون Mini-CEX مطالعاتی که سودمندی این روش را نشان می‌دهند، می‌توانند برای آزمون DOPS نیز به کار روند.

روایی آزمون DOPS

در صورت تهیه بلوپرینت و نمونه‌گیری مناسب از پروسیجرهای برنامه درسی می‌توان روایی محتوای مناسبی برای این آزمون به‌دست آورد. هر چند در عمل ممکن است احتمال ارزیابی برخی از پروسیجرها فراهم نشود.

در این آزمون عملکرد داوطلب در زمان انجام پروسیجر در محیط واقعی و بر روی بیمار واقعی مورد مشاهده قرار می‌گیرد و درنتیجه باور بر این است که از روایی صوری مناسبی به ویژه در مقایسه با روش‌های دیگر ارزیابی پروسیجرها مانند لاگ‌بوک و OSATS از روایی صوری بیشتری برخوردار است.

بررسی‌های مختلف ضریب همبستگی بالایی بین ابزارهای مشاهده‌ای پروسیجرها و شبیه‌سازهای واقعیت مجازی[1] یا ابزارهای تحلیل حرکت[2] در موقعیت‌های واقعی نشان می‌دهند (مانند ضریب اسپیرمن ۰/۸۸ و ۰/۰۵ >P در مطالعه احمد و همکاران ۲۰۱۱). اما بین دانش فراگیران در آزمون بورد جراحی و عملکرد آن‌ها در پروسیجرهای جراحی که با فرم نمره‌دهی کلی مورد ارزیابی قرار گرفتند همبستگی دیده نشده است. بارتون و همکاران[3] (۲۰۱۲) گزارش کردند که بین نمرات آزمون DOPS انجام‌شده برای ارزیابی مهارت پروسیجرال کولونوسکوپی و آزمون چندگزینه‌ای ارتباط مثبت معنی‌دار اما نسبتاً کمی (۰/۲۷) وجود داشت اما بین نمرات این آزمون با تعداد پروسیجرهایی که دستیاران در سال قبل و در کل دوره انجام داده بودند، ارتباط منفی وجود داشت. این ارتباط ضعیف می‌تواند بیان‌کننده این موضوع باشد که آزمون DOPS روش مناسبی برای ارزیابی توانمندی پروسیجرال نیست. هر چند این موضوع نیاز به بررسی دقیق‌تر دارد.

روایی سازه این آزمون مناسب گزارش شده است. به طوری‌که فراگیران سطوح بالا، نمرات بالاتری در این آزمون کسب نمودند و همچنین نمرات آزمون در طول زمان افزایش یافت. دلفینو و همکاران (۲۰۱۳) نشان دادند بین عملکرد فراگیران سال اول با دوم (P =۰/۰۴۲) و بین عملکرد فراگیران سال اول با سوم (P =۰/۰۳۹) تفاوت معنی‌داری در انجام پروسیجر لوله‌گذاری داخل نای وجود داشت اما تفاوت بین عملکرد دستیاران سال دوم و سوم معنی‌دار نبود. این مسأله ممکن است به این دلیل باشد که دستیاران بیهوشی پروسیجر لوله‌گذاری را مکرراً انجام می‌دهند و در نتیجه پس از یک دوره زمانی به خبرگی در آن می‌رسند و در نتیجه ممکن است بعد از مدت زمان خاصی نتوان تفاوتی بین سال‌های مختلف تحصیل مشاهده کرد. بنابراین لازم است مطالعات بیشتر با پروسیجرهای پیچیده‌تر و کمتر شایع‌تر انجام شود تا روایی سازه آزمون DOPS ثابت شود. در خصوص عوامل مخدوش‌کننده تأثیرگذار بر نمره آزمون DOPS، نقش نوع پروسیجر مورد ارزیابی بررسی شده است و نتایج نشان دادند تفاوت معنی‌داری بین نمرات پروسیجرهای قلبی، آندوسکوپی، عصبی–فیزیولوژیکی[4] و کلیوی دیده نشد (ویلکینسون و همکاران ۲۰۰۸).

پایایی آزمون DOPS

هر چند تاکنون مطالعات کمی در مورد سودمندی DOPS انجام شده است، به نظر می‌رسد در مقایسه با دیگر روش‌های مبتنی بر محل کار، این آزمون در قبال تعداد مواجهه کمتر از پایایی بیشتری برخوردار است. در این زمینه پژوهش‌های بیشتر به منظور قطعی نمودن یافته‌ها لازم است.

- یکی از اولین پژوهش‌هایی که پایایی آزمون DOPS را بررسی کرده است، مطالعه ویلکینسون و همکاران (۲۰۰۸) است. در این پژوهش نتایج استفاده از فرم‌های DOPS برای ۲۳۰ دستیار در ۱۷ رشته تخصصی مختلف بررسی شد. بر اساس یافته‌های این پژوهش در صورتی‌که حداقل سه ارزیاب، عملکرد یک فراگیر را مورد مشاهده قرار دهند، در حالی‌که هر کدام حداقل دو پروسیجر را ارزیابی کنند، آزمون از پایایی مناسبی برخوردار است. البته این تعداد مواجهه و ارزیاب به خوبی نمی‌تواند دانشجویان با عملکرد نامطلوب را تشخیص دهد و در نتیجه استفاده از آن با اهداف تراکمی مقایسه‌ای باید با احتیاط صورت گیرد.
- در مطالعه بارتون و همکاران (۲۰۱۲) نیز با دو مورد (پروسیجر) که هر کدام توسط دو آزمون‌گر مشاهده شوند، پایایی ۰/۸۱ به دست آمد. در این مطالعه، واریانس مشاهده‌گران بیشتر از واریانس موارد بود و افزایش ارزیابان نسبت به افزایش تعداد موارد (پروسیجر)، ضریب تعمیم‌پذیری را بیشتر افزایش داد. در مطالعه مذکور مشاهده‌گران قبل از انجام آزمون در یک کارگاه پنج‌ساعته آموزشی شرکت کرده بودند و این احتمال وجود دارد که پایایی خوب آزمون به

1. Virtual reality
2. Motion analysis devices
3. Barton et al.
4. Neurophysiological

دلیل این موضوع باشد. این پژوهشگران نتیجه گرفتند آزمون DOPS در ارزیابی مهارت کولونوسکوپی روا و پایا است.

- دلفینو و همکاران (۲۰۱۳) با شش پروسیجر ضریب پایایی ۰/۸ به دست آوردند و ثبات درونی فرم‌ها خوب گزارش شد (۰/۸۷ تا ۰/۹).

تأثیر آموزشی آزمون DOPS

به صورت نظری می‌توان گفت که آزمون DOPS با ارائه بازخورد، فرصتی برای ارتقای یادگیری فراهم می‌کند و همچنین تأثیر مثبتی بر رویکرد به یادگیری دارد و موجب یک رویکرد عمقی‌تر می‌شود. همان‌طور که پیدا است مطالعات در زمینه تأثیر آموزشی آزمون DOPS نادر هستند و نتایج پژوهش‌های انجام‌شده نیز لزوماً در یک راستا نیستند:

- کوب و همکاران[1] (۲۰۱۳) تأثیر آموزشی دو آزمون DOPS و سؤالات چندگزینه‌ای را بر رویکرد دانشجویان به یادگیری مقایسه کردند. در این پژوهش پرسشنامه فرایند مطالعه[2] برای دانشجویان سال آخر دامپزشکی فرستاده شد. نتایج نشان داد دانشجویان در برابر آزمون DOPS رویکرد «عمقی‌تری»[3] داشتند و در برخورد با چندگزینه‌ای رویکرد مطالعه «سطحی‌تری»[4] در پیش گرفتند. البته در زمان نزدیک به امتحان نهایی که به صورت چندگزینه‌ای برگزار می‌شد رویکرد به مطالعه در آزمون DOPS نیز سطحی شده و در نتیجه دانشجویان یک رویکرد «کسب موفقیت[5]» در پیش می‌گرفتند زیرا موفقیت در آزمون DOPS و چندگزینه‌ای برایشان مطرح بود. در این پژوهش تفاوت معنی‌داری بین رویکرد «سطحی» و «عمقی» به مطالعه و روش آزمون وجود داشت اما بین دو نوع آزمون و رویکرد «کسب موفقیت» تفاوت معنی‌داری وجود نداشت. علاوه بر این، بین رویکرد مذکور و عملکرد در آزمون چندگزینه‌ای ارتباط معنی‌دار مثبتی وجود داشت. البته تأثیر آموزشی احتمالاً بیشتر از آن که به شکل دو آزمون مربوط باشد به دلیل نوع پیامد یادگیری مورد ارزیابی است. علاوه بر این، اهمیت آزمون (تکوینی یا تراکمی) در سرنوشت دانشجویان نیز بر رویکرد مطالعه مؤثر است.
- هر چند نتیجه مطالعه فوق از تأثیر مثبت آزمون حکایت داشت اما در مطالعه دیگری، داوطلبان و ارزیابان در دوره دستیاری بیهوشی این آزمون را ابزار آموزشی مناسبی ارزیابی نکردند (بیندال و همکاران[6] ۲۰۱۳). آن‌ها DOPS را یک تمرین علامت زدن و نه یک فرصت آموزشی معرفی کردند و بیان داشتند این آزمون منعکس کننده توانمندی داوطلبان در انجام صحیح پروسیجر نیست. با توجه به این‌که زمان DOPS انجام شده در این مطالعه بسیار کم بود (در یک سوم موارد کمتر از پنج دقیقه) این احتمال وجود دارد که زمان کافی برای ارائه بازخورد اختصاص داده نشده است. با در نظر گرفتن نحوه اجرای آزمون، تاثیر آموزشی پایین گزارش‌شده منطقی به نظر می‌رسد.
- در مطالعه دلفینو و همکاران (۲۰۱۳)، دو سوم ارزیابان بیان داشتند که ارائه بازخورد برای آن‌ها مشکل بود و نیمی از آن‌ها به طور روتین بازخورد ارائه ندادند. آن‌ها دلیل این موضوع را کمبود وقت، کمبود اعتماد به نفس و دانش ناکافی بیان کردند. به همین ترتیب، نیمی از داوطلبان نیز بیان داشتند هیچ بازخوردی دریافت نکرده‌اند اما در صورت دریافت آن را مؤثر دانستند. بیش از نیمی از داوطلبان و ارزیابان این آزمون را مصنوعی دانستند که ممکن است به دلیل جدید بودن ابزار باشد.

1. Cobe et al.
2. Study Process Questionnaire (SPQ)
3. Deep
4. Surface
5. Achieving
6. Bindal et al.

هزینه و قابلیت اجرای DOPS

این آزمون زمان‌بر نیست و در مورد پروسیجرهای رایج به راحتی اجرا می‌شود. هر چند در عمل با احتساب زمان صرف‌شده برای اخذ رضایت از بیمار و جلب اعتماد وی و ورود اطلاعات در کارپوشه فراگیران، زمان آزمون بیشتر از رقم گزارش‌شده ۱۵ تا ۲۰ دقیقه است.

با توجه به این‌که این آزمون نیاز به منابع خاصی مانند شبیه‌سازها و مدل‌های مصنوعی ندارد، از هزینه اثربخشی قابل قبولی برخوردار است؛ به ویژه در مقایسه با روش‌هایی که در محیط شبیه‌سازی به ارزیابی این پروسیجرها می‌پردازند.

قابلیت اجرای آزمون DOPS در مواردی با چالش مواجه می‌شود. به عنوان مثال، انجام برخی از پروسیجرها خیلی رایج نیست و در نتیجه یافتن موقعیتی برای انجام آن مشکل است. زمانی هم که فی‌البداهه انجام چنین پروسیجرهایی محقق می‌شود، هماهنگی برای حضور ارزیاب چندان آسان نیست و حتی این احتمال وجود دارد که ارزیاب در بیمارستان حضور نداشته باشد. رگ‌گیری و خون‌گیری از جمله پروسیجرهایی هستند که می‌توان به صورت برنامه‌ریزی‌شده آن‌ها را مشاهده کرد. از آنجا که بخش اورژانس امکان مشاهده پروسیجرهای بیشتری را فراهم می‌کند، بهتر است در شروع، اجرای آزمون از این بخش آغاز کرد و بخش بیماران بستری امکان انجام پروسیجرهای از قبل برنامه‌ریزی‌شده را فراهم کند.

انتخاب بیمار مناسب هم موضوع دیگری است که همیشه امکان آن وجود ندارد. همچنین در بسیاری از مواقع، پروسیجرها برای بیماران بدحال مورد نیاز هستند که ممکن است انجام آن را غیرممکن نماید. ممکن است بیماران نیز با انجام پروسیجر برای امتحان موافقت نکنند.

منابع

1. Abbas Khan MA, Gorman M, Gwozdziewicz L, Sobani ZA, Gibson C. Direct Observation of Procedural Skills (DOPS) as an Assessment Tool for Surgical Trainees. J PAK MED STUD. 2013;3(3):137–40.
2. Ahmed K, Miskovic D, Darzi A, Athanasiou T, Hanna GB. Observational tools for assessment of procedural skills: a systematic review. AM J SURG. 2011;202(4):469–80.
3. Amin Z, Seng CY, Eng KH. Practical guide to medical student assessment. World Scientific Publishing; 2006.
4. Barton JR, Corbett S, van der Vleuten CP. The validity and reliability of a Direct Observation of Procedural Skills assessment tool: assessing colonoscopic skills of senior endoscopists. Gastrointest endosc. 2012;75(3):591–7.
5. Bindal N, Goodyear H, Bindal T, Wall D. DOPS assessment: A study to evaluate the experience and opinions of trainees and assessors. Med Teach. 2013;35(6):e1230–4.
6. Cantillon P, Wood D. ABC of Learning and Teaching in Medicine. 2nd ed. West Sussex: John Wiely & Sons; 2010.
7. Cobb KA, Brown G, Jaarsma DADC, Hammond RA. The educational impact of assessment: A comparison of DOPS and MCQs. Med Teach. 2013;35(11):e1598–607.
8. Delfino AE, Chandratilake M, Altermatt FR, Echevarria G. Validation and piloting of direct observation of practical skills tool to assess intubation in the Chilean context. Med Teach. 2013;35(3):231–6.
9. Holmboe ES, Hawkins RE. Practical guide to the evaluation of clinical competence. Philadelphia: Mosby/Elsevier; 2008.
10. Mcleod R, Mires G, Ker J. Direct observed procedural skills assessment in the undergraduate setting. Clin Teach. 2009;9(4):228–32.
11. Morris A, Hewitt J, Roberts CM. Practical experience of using direct observed procedures, mini clinical examinations and peer observation in pre-registration house officers (FY1) trainees. Postgrad Med J. 2006;82(966):285–8.
12. Norcini J, Burch V. Workplace-based assessment as an educational tool: AMEE Guide No. 31. Med Teach. 2007;29(9-10):855–71.
13. Norcini JJ, McKinkley DW. Assessment methods in medical education.Teach Teach Educ. 2007;23(3):239–50.
14. Pelgrim EAM, Kramer AWM, Mokkink HGA, van den Elsen L, Grol RPTM, van der Vleuten CPM. In-training assessment using direct observation of single-patient encounters: A literature review. Adv Health Sci Educ. 2011;16:189–99.

15. Swanwick T. Uderstanding Medical Education: Evidence, Theory and Practice. West Sussex: John Wiely & Sons; 2010.
16. Turnbull J, MacFadyen J, van Barneveld C, Norman G. Clinical work sampling. A new approach to the problem of in-training evaluation. JGIM. 20001;15(8):556–61.
17. Wilkinson JR, Crossley JGM, Wragg A, Mills P, Cowan G, Wade W. Implementing workplace-based assessment across the medical specialties in the United Kingdom. Med Educ. 2008;42(4):364–73.

فصل ۲۳

آزمون Chart Stimulated Recall (CSR)

ساختار آزمون CSR

روش Chart Stimulated Recall (CSR)، از ابزارهای ارزیابی مبتنی بر محل کار است که مبتنی بر پرونده‌های پزشکی بیماران انجام می‌شود. این ابزار بر مرور ساختارمند مهارت‌های تصمیم‌گیری و استدلال بالینی فراگیران در محیط طبابت واقعی متمرکز است. استفاده از پرونده‌های بیماران از طریق ارتباط دادن روش‌های ارزیابی با محیط کار واقعی و معنی‌دار کردن ارزیابی باعث می‌شود ارزیابی مبتنی بر اصول یادگیری بزرگسالان انجام شود. CSR سه حوزه زیر را مورد ارزیابی قرار می‌دهد:

- بالینی (دانش، تصمیم‌گیری و مهارت‌ها)
- تعهد حرفه‌ای (اخلاق و کار گروهی)
- مهارت‌های ارتباطی (با بیماران، خانواده بیمار و همکاران)

در این روش ارزیابی از بخشی از پرونده بیمار یا بخش‌های مختلف آن به منظور ارزیابی مهارت‌های تحلیل، ترکیب و قضاوت بالینی فراگیران استفاده می‌شود. علاوه بر این، ارزیابان قادر خواهند بود عوامل دیگری مانند بیمار، محیط و دیگر عوامل مربوط به سیستم را که بر تصمیم‌گیری فراگیران تأثیرگذار بوده‌اند، بررسی کنند. یادداشت‌های قسمت‌های مختلف پرونده بیمار مربوط به زمان پذیرش بیمار، شرایط حاد، پیشرفت روزانه و یادداشت‌های پیگیری بیمار در درمانگاه به عنوان مبنایی برای ارزیابان مورد استفاده قرار می‌گیرد تا سؤالات بیشتری را به منظور کاوش مهارت‌های استدلال و تصمیم‌گیری بالینی طرح کنند.

روش ارزیابی CSR برای اولین بار در دهه ۸۰ میلادی در بورد طب اورژانس آمریکا[1] طراحی شد و مورد استفاده قرار گرفت. بورد طب اورژانس در این روش ارزیابی یک آزمون شفاهی به عمل آورد که مبنای آن پرونده‌های بیماران بخش اورژانس بود که آزمون‌شوندگان آن را تهیه کرده بودند. ارزیابان از پرونده‌ها به منظور ارزیابی دانش، مهارت‌های حل مشکلات بالینی و تخمینی کلی از توانمندی بالینی دستیاران استفاده کردند. در نهایت بورد طب اورژانس تأیید کرد که CSR روشی پایا و معتبر است و سه تا شش پرونده بیمار می‌تواند پایایی کافی را برای این روش ارزیابی فراهم کند. هر چند، CSR به عنوان یک روش ارزیابی تراکمی گران و زمان‌بر بود. بنابراین، علی‌رغم رضایت استفاده‌کنندگان از این روش ارزیابی، بورد طب اورژانس این روش را از فرایند اعطای مدک بورد حذف کرد (هولمبو و هاوکینز ۲۰۰۸). پس از آن CSR در ارزیابی عملکرد پزشکان در حال طبابت در رشته‌های تخصصی مختلف مانند پزشکی خانواده، طب فیزیکی و طب کار

1. The American Board of Emergency Medicine (ABEM)

مورد استفاده قرار گرفت و روایی و پایایی آن نیز تأیید شد. ACGME استفاده از این روش ارزیابی را در ارزیابی دستیاران پزشکی نیز توصیه نمود (هایدن و همکاران[1] ۲۰۰۲) و هم اکنون بخش قابل‌توجهی از برنامه‌های دستیاری در کشورهای آمریکای شمالی (در کانادا متناسب با چارچوب توانمندی‌های CanMEDS) از این روش به منظور ارزیابی دستیاران، به ویژه در دستیابی به توانمندی «یادگیری و ارتقای مبتنی بر طبابت» استفاده می‌کنند (شیپر و رز[2] ۲۰۱۰). قابل ذکر است هر چند طراحی اولیه این روش ارزیابی با هدف تراکمی و اعطای مدرک بورد صورت گرفت، این ابزار هم اکنون بیشتر با کاربرد تکوینی و یادگیری در برنامه‌های دستیاری به کار می‌رود. فرم‌ها و فرایندهای مورد استفاده در آزمون نیز نسبت به شکل اولیه آن با توجه به دوره دستیاری که آن را مورد استفاده قرار داده‌اند، تغییر یافته است.

مشابه این روش ارزیابی تحت نام «بحث مبتنی بر موارد بالینی»[3] (CBD) در کشور انگلیس به عنوان بخشی از ارزیابی فراگیران در دوره پیش‌دستیاری به کار گرفته شده است. در آزمون CBD فراگیر دو پرونده از بیماران خود را انتخاب می‌کند و به ارزیاب تحویل می‌دهد. ارزیاب یکی از دو مورد را برای بحث انتخاب می‌کند و بر روی یک یا دو جنبه از مهارت‌های بالینی شامل ارزیابی بالینی، بررسی و ارجاع بیمار، درمان، پیگیری و برنامه‌ریزی برای مراجعات بعدی و تعهد حرفه‌ای تمرکز می‌کند. علاوه بر این مستندسازی نیز، که جزء اهداف اولیه آزمون نیست، به دلیل در دسترس بودن پرونده بیمار قابل ارزیابی است. از فراگیران درخواست می‌شود در مورد دلیل انتخاب و انجام عمل خود توضیح دهند. این روش فرصتی را برای ارزیابی کاربرد دانش، تصمیم‌گیری و مسائل اخلاقی فراهم می‌کند. هر مواجهه ۲۰ دقیقه طول می‌کشد که پنج دقیقه آن به بازخورد اختصاص دارد. انتظار می‌رود فراگیران در طول دوره آموزشی در چندین مواجهه با ارزیابان متعدد ارزیابی شوند. در برنامه پیش‌دستیاری انگلیس تعداد مواجهات چهار تا شش مورد در سال است. ارزیابان می‌توانند اعضای هیأت علمی بالینی، دستیاران تخصصی با تجربه، مربیان و پزشکان عمومی باشند. انتظار می‌رود فراگیران از فهرست برنامه درسی پایه موارد را انتخاب کنند. در این روش قضاوت در مورد یک مواجهه صورت می‌گیرد و ارزیابان در مورد کیفیت عملکرد فراگیر و تناسب آن قضاوت می‌کنند. همان‌طور که بعداً اشاره می‌شود، فرم‌ها و فرایند آزمون CBD برگرفته از آزمون CSR است اما متناسب با شرایط برنامه پیش دستیاری اصلاح شده است.

همچنین CSR می‌تواند با دیگر ابزارهای ارزیابی توانمندی تلفیق شود. به عنوان مثال، ترکیب یک روش ارزیابی مشاهده مستقیم مانند mini-CEX با CSR موجب می‌شود درک عمیق‌تری از مهارت تصمیم‌گیری بالینی ارزیابی شونده داشته باشیم. ترکیب CSR و «ممیزی مستندات پزشکی»، ابزار باارزشی در ارزیابی و ارتقای کیفیت مراقبت از بیمار است (جنت و همکاران[4] ۱۹۹۵).

مزایا و محدودیت‌های CSR

مزایای آزمون CSR

- یک ابزار ارزیابی مبتنی بر داده‌های محیط واقعی است
- یک ابزار آموزشی است و به یادگیری کمک می‌کند.
- بازخورد فوری، مرتبط، اختصاصی و مستندشده ارائه می‌دهد.
- مهارت‌های مستندسازی فراگیران را ارتقاء می‌دهد.
- توانمندی‌هایی که با دیگر روش‌های ارزیابی قابل بررسی نیستند مانند «یادگیری مبتنی بر طبابت» و «طبابت مبتنی

1. Hayden et al.
2. Schipper & Ross
3. Case Based Discussion
4. Jennett et al.

بر سیستم» را ارزیابی می‌کند و علاوه بر این، موجب شکل‌گیری این توانمندی‌ها در فراگیران می‌شود.
- توانایی ارزیابی استدلال بالینی، تصمیم‌گیری و تفکر نقاد را دارد.

گام‌های طراحی و اجرای آزمون CSR مطلوب

در نظر گرفتن موارد زیر که به صورت خلاصه در جدول ۱-۲۳ آمده‌اند، به طراحی و اجرای یک آزمون CSR یا CBD مطلوب کمک می‌کند.

جدول ۱-۲۳: خلاصه مراحل طراحی و اجرای آزمون CSR

ردیف	عنوان	توضیح
۱	مشخص نمودن هدف آزمون	اولین قدم در طراحی ابزار تعیین هدف آن (تکوینی یا تراکمی) است به این دلیل که بر بخش‌های دیگر آن مانند تدوین فرم‌ها تأثیرگذار است.
۲	توجه به پوشش برنامه درسی پایه در زمان انتخاب پرونده‌ها	این موضوع به ویژه در مواردی که آزمون با هدف تراکمی به کار می‌رود، به منظور رفع مشکل ویژگی مورد مهم است. لازم است طیفی از موارد بیماری متنوع بر اساس مشکلات و شرایط بیمار، ویژگی‌های دموگرافیک بیمار و محیط طبابت بر اساس برنامه درسی پایه پوشش داده شوند.
۳	طراحی فرم‌های آزمون	ضروری است متناسب با هدف آزمون (تکوینی یا تراکمی) فرم‌های آن طراحی شود. در این زمینه فرم‌های مناسبی برای طراحان آزمون در دسترس است.
۴	تعیین معیارهای بررسی پرونده	در اغلب موارد معیارهایی در اختیار ارزیابان قرار می‌گیرد تا در زمان بررسی اولیه، پرونده‌ها را بر اساس آن معیارها ارزیابی کنند. از آنجا که حین آزمون بر اساس مندرجات پرونده، ارزیاب به پرسش از دستیار می‌پردازد، لازم است که از قبل سؤالاتی را متناسب با پرونده موردنظر طراحی کرده باشد.
۵	اطلاع‌رسانی، آشناسازی و آموزش ارزیابان و داوطلبان	لازم است ارزیابان از بین اعضای هیأت علمی بالینی که به محتوای پرونده‌ها اشراف دارند، انتخاب شوند و به این منظور آموزش دیده باشد. همچنین لازم است دستورالعمل‌های مشخصی در ارتباط با فرایند اجرای آزمون تدوین شود و به اطلاع دست‌اندرکاران رسانده شود.
۶	اجرای آزمون	پرونده‌های انتخاب شده توسط داوطلب جهت مرور اولیه در اختیار ارزیابان قرار می‌گیرد. پس از مرور اولیه بر اساس معیارهای از قبل تهیه شده، پرونده و مواجهه بیمار مورد بحث قرار می‌گیرد. در زمان بحث ارزیاب سؤالاتی را مطرح می‌کند تا به عمق مهارت‌های استدلال بالینی، تصمیم‌گیری یا تعهد حرفه‌ای فراگیر واقف شود. ارزیاب می‌تواند از نتایج مرور اولیه به عنوان مبنایی برای بحث بیشتر استفاده کند.
۷	بررسی کیفیت آزمون برگزارشده	ضروری است کیفیت آزمون‌های برگزار شده با استفاده از روش‌های کمی و کیفی بررسی شود.

مشخص نمودن هدف آزمون

همان‌طور که اشاره شد، CSR یا CBD می‌توانند با اهداف تکوینی یا تراکمی به کار روند. اولین قدم در طراحی ابزار تعیین هدف آن است به این دلیل که بر بخش‌های دیگر آن مانند تدوین فرم‌ها تأثیرگذار است. به عنوان مثال، بخش عمده فرم‌هایی که با هدف تکوینی به کار می‌روند، به ارائه بازخورد اختصاص یافته است. آزمون CSR دانشگاه آلبرتا و ییل[1] با هدف تکوینی و آزمون CBD برنامه پیش‌دستیاری انگلیس با اهداف تراکمی یا تکوینی طراحی شده‌اند. هدف آزمون هر چه باشد لازم است از همان ابتدا به اطلاع فراگیران نیز رسانده شود.

1. Yale

تجربه دانشگاه ییل از آزمون CSR (هولمبو و هاوکینز ۲۰۰۸)

در دانشگاه ییل، آزمون CSR با هدف تکوینی و به عنوان بخشی از فرایند تشخیص دستیاران با ضعف عملکرد به کار می‌رود. در این آزمون، یادداشت‌های مربوط به یک مواجهه (به عنوان مثال زمان پذیرش، یادداشت‌های پیشرفت روزانه، یادداشت‌های پیگیری بیماران در درمانگاه و ...) به دو منظور مورد استفاده قرار می‌گیرند:

- مرور کیفیت پایه یادداشت‌ها با استفاده از یک چارچوب ساده. سؤالات مربوط به این مرور عبارت هستند از:
  - آیا یادداشت‌ها خوانا هستند؟
  - آیا یادداشت‌ها از یک ساختار استاندارد (مثلاً SOAP یا حل مسأله) پیروی می‌کنند؟
- مرور کیفیت یادداشت‌ها به صورت وسیع‌تر با بررسی:
  - ثبات درونی: آیا مسأله یا مشکلی که در شرح‌حال و معاینه فیزیکی بیان شده است، در ادامه یادداشت‌ها به طور منطقی دنبال شده است؟ به عنوان مثال اگر فراگیر در بخش مشکل اصلی، «درد قفسه سینه» را مطرح کرده است، در بخش معاینه فیزیکی باید به معاینه قلبی بپردازد و در بخش ارزیابی و برنامه نیز تشخیص‌های افتراقی و سپس مدیریت درد قفسه صدری را مطرح کند.
  - ناهماهنگی‌ها: ناهماهنگی به این معنی است که اطلاعات در یک بخش با اطلاعات یا تصمیمات مستندشده در بخش دیگر هماهنگی ندارد. در مثال درد قفسه سینه، علائم همراهی که در بخش مشکل اصلی ذکر شده است، مطرح‌کننده بیماری ایسکمیک قلبی است اما فراگیر در بخش ارزیابی، سوزش سر دل را به عنوان محتمل‌ترین تشخیص بدون ارتباطی با بیماری ایسکمیک قلبی مطرح کرده است.

هر چند مشکلات مربوط به ثبات درونی و ناهماهنگی می‌توانند به خطاهای مربوط به مستندسازی مربوط باشند، به احتمال بیشتری نشاندهنده ضعف در دانش و تصمیم‌گیری بالینی هستند. ارزیابان زمانی که این خطاها را مشاهده می‌کنند می‌توانند از آن به عنوان مبنایی جهت بررسی بیشتر دانش پزشکی و مهارت‌های تصمیم‌گیری بالینی فراگیران استفاده کنند. اگر پرونده بیمار مواردی از عدم هماهنگی و ثبات درونی نداشته باشد، باز هم آزمون CSR ابزار مفیدی است. به عنوان مثال، دستیار در پرونده، تشخیص صحیح پرفشاری خون را برای بیمار مطرح کرده است و درمان مناسب مبتنی بر شواهد (مثلاً یکی از داروهای خانواده دیورتیک‌ها) را انتخاب کرده است. سؤال این است که آیا تجویز دیورتیک یک انتخاب طوطی‌وار بوده است یا انتخابی آگاهانه و بر اساس دستورالعمل‌ها؟ آیا دیورتیک بهترین انتخاب برای بیمار بوده است و آیا فراگیر از عوارض جانبی درمان آگاه است و آن را با عوامل خطر بیمار و بیماری‌های همراه دیگرش مرتبط کرده است؟ ارزیاب می‌تواند پاسخ این سؤالات را در قالب آزمون CSR پیدا کند.

توجه به پوشش برنامه درسی پایه در زمان انتخاب پرونده‌ها

این موضوع به ویژه در مواردی که آزمون با هدف تراکمی به کار می‌رود، به منظور رفع مشکل ویژگی مورد مهم است. لازم است طیفی از موارد بیماری متنوع بر اساس مشکلات و شرایط بیمار، ویژگی‌های دموگرافیک بیمار و محیط طبابت بر اساس برنامه درسی پایه پوشش داده شوند. به همین دلیل این مسأله باید به صورت کاملاً روشن مشخص شود و به اطلاع دستیاران نیز رسانده شود. زیرا در این آزمون خود دستیاران پرونده‌های مورد نظر برای آزمون را انتخاب می‌کنند. همچنین به منظور دستیابی به پایایی و روایی مطلوب برای آزمون‌های تراکمی بیش از چهار تا شش پرونده در سال لازم است.

طراحی فرم‌های آزمون

همان‌طور که در ابتدای این فصل گفته شد، به مرور زمان فرم‌های مختلفی با توجه به نیاز برنامه‌های آموزشی مختلف تدوین شده است. شناخته‌شده‌ترین این فرم‌ها، فرم‌های مورد استفاده در برنامه‌های دستیاری کشور کانادا (فرم آزمونCSR مورد استفاده در دانشگاه آلبرتا) و فرم‌های مورد استفاده در آزمون CBD برنامه پیش‌دستیاری انگلیس است. این فرم‌ها هم اکنون مبنای تدوین فرم‌های متنوع دیگری در برنامه‌های دستیاری مختلف قرار گرفته‌اند. بنابراین، به نظر می‌رسد فرم‌های مناسبی برای طراحان آزمون در دسترس است. در شکل ۱-۲۳ نمونه‌ای از فرم CBD ارائه شده است.

تعیین معیارهای بررسی پرونده

ارزیابان قبل از شروع جلسه حضوری با دستیار، باید پرونده‌هایی که توسط دستیار در اختیارشان قرار گرفته است، مطالعه و بررسی کنند. در اغلب موارد معیارهایی در اختیار ارزیابان قرار می‌گیرد تا در زمان بررسی اولیه، پرونده‌ها را بر اساس آن معیارها ارزیابی کنند. در زیر نمونه‌ای از معیارهای به کاررفته در آزمون CSR دانشگاه ییل آمده است.

Case-Based Discussion (CBD) – F_2 Version

لطفا با گذاشتن علامت ضربدر مقابل سؤالات فرم را کامل کنید: ☒ لطفا از قلم مشکی برای تکمیل فرم استفاده کنید.

نام خانوادگی داوطلب	
نام داوطلب	
شماره GMC (دانشجویی)	**ثبت شماره GMC الزامی است**
محیط بالینی	درمانگاه ☐ بستری ☐ پذیرش اورژانس ☐ پذیرش جراحی ☐
مشکل بالینی	تنفسی ☐ گردش خون ☐ گوارشی ☐ نورولوژیک ☐ درد ☐ روانپزشکی ☐ دیگر موارد
تمرکز مواجهه بالینی	ثبت پرونده پزشکی ☐ ارزیابی بالینی ☐ مدیریت ☐ تعهد حرفه‌ای ☐
موقعیت ارزیاب	پیچیدگی بیماری: کم ☐ متوسط ☐ زیاد ☐

لطفا حیطه‌های زیر را با استفاده از نمره‌دهی مقابل نمره‌دهی کنید	زیر حد انتظار برای F_2	مرزی	در حد انتظار برای F_2	بالاتر از حد انتظار برای F_2	بدون نمره*
۱. ثبت پرونده پزشکی	☐	☐	☐	☐ ☐	☐
۲. ارزیابی بالینی	☐	☐	☐	☐ ☐	☐
۳. بررسی و ارجاع	☐	☐	☐	☐ ☐	☐
۴. درمان	☐	☐	☐	☐ ☐	☐
۵. پیگیری و برنامه‌ریزی برای آینده	☐	☐	☐	☐ ☐	☐
۶. تعهد حرفه‌ای	☐	☐	☐	☐ ☐	☐
۷. قضاوت بالینی کلی	☐	☐	☐	☐ ☐	☐

* بدون نمره: لطفا این مورد را در صورتی علامت بزنید که رفتار مورد نظر مشاهده نشده است و در نتیجه قادر به نمره‌دهی نیستید.

موارد مثبت عملکرد	**موارد پیشنهادی برای ارتقای عملکرد**

اقدامات مورد توافق:

	بدون رضایت								رضایت کامل
رضایت فراگیر از CBD	☐ ۱	☐ ۲	☐ ۳	☐ ۴	☐ ۵	☐ ۶	☐ ۷	☐ ۸	☐ ۹	☐ ۱۰
رضایت ارزیاب از CBD	☐ ۱	☐ ۲	☐ ۳	☐ ۴	☐ ۵	☐ ۶	☐ ۷	☐ ۸	☐ ۹	☐ ۱۰

آیا در مورد این روش تاکنون آموزش دیده‌اید؟ چهره به چهره ☐ مطالعه دستورالعمل‌ها ☐ اینترنت/CD ☐

امضای ارزیاب	تاریخ	زمان صرف شده برای بحث (دقیقه)	زمان صرف شده برای بازخورد (دقیقه)

نام خانوادگی ارزیاب	
شماره پرسنلی ارزیاب	

شکل ۱-۲۳: فرم CBD طراحی‌شده توسط نظام ملی سلامت انگلیس برای استفاده در برنامه پیش‌دستیاری (برای اطلاعات بیشتر در مورد این فرم و جزئیات توانمندی‌های مورد انتظار برای F1 و F2 به سایت WWW.HCAT.NHS.UK مراجعه کنید)

معیارهای بررسی پرونده قبل از آزمون CSR در دانشگاه ییل

سازمان‌دهی و شفافیت یادداشت‌ها
آیا یادداشت‌ها سازمان‌دهی شده است؟ آیا ساختار آن مناسب و سازگار است؟ آیا یادداشت‌ها خوانا هستند؟
محتوای یادداشت‌ها
آیا مشکلات بالینی با جزئیات کافی توضیح داده شده‌اند؟ آیا هر گونه اطلاعات ضروری از دست رفته است؟
ثبات درونی
آیا مسائل بالینی توالی منطقی را در طول یادداشت‌ها دارند؟
عدم‌هماهنگی
آیا تصمیمات یا مانورهایی وجود دارند که با بخش‌های دیگر همخوانی ندارند؟
پنج سؤالی را که تمایل دارید از این دستیار در مورد قضاوت و استدلال بالینی بپرسید، یادداشت کنید.
۱)
۲)
۳)
۴)

تدوین راهنمای طراحی سؤالات

از آنجا که حین آزمون بر اساس مندرجات پرونده، ارزیاب به پرسش از دستیار می‌پردازد، لازم است که از قبل سؤالاتی را متناسب با پرونده موردنظر طراحی کرده باشد. در برخی از فرم‌های ارزیابی، سؤالاتی به عنوان نمونه به منظور راهنمای ارزیابان درج شده است. نمونه‌ای از آن را در تجربه دانشگاه ییل مشاهده کردید. نمونه دیگری متعلق به دانشگاه آلبرتا در زیر آمده است:

نمونه سؤالات پیشنهادی در آزمون CSR دانشگاه آلبرتا

مرور کلی مورد بیماری
- حیطه ارزیابی بالینی- متخصص پزشکی خانواده، برقرارکننده ارتباط
  - آیا می‌توانید مرور کلی از مورد بیماری ارائه دهید؟
  - چه ویژگی‌هایی در تظاهرات بیمار شما را به سمت دو تشخیص اول هدایت کرد؟
  - آیا شما در مورد تجربه ناخوشی بیمار (احساسات، عقاید، تأثیر بر عملکرد و انتظارات) سؤال کردید؟ چه چیزی یاد گرفتید؟
  - آیا مواردی از ابهام یا عدم قطعیت در مورد بیماری وجود داشت؟ شما چگونه به آن پرداختید؟
  - آیا مواردی وجود دارد که تمایل داشته باشید در موردشان سؤال کنید؟
- حیطه بررسی و ارجاع- همکار، مدیر
  - به چه دلیل این بررسی‌ها را انجام دادید؟
  - آیا به آزمایش‌های دیگری هم فکر کردید که آن ها را درخواست نکرده باشید؟ چرا؟

چگونه تصمیم گرفتید به متخصص مورد نظر ارجاع دهید؟
- حیطه درمان و مدیریت بیماری- برقرارکننده ارتباط
  - چه عواملی موجب شد که شما درمان مورد نظر را انتخاب کنید؟
  - انتظارات بیمار از درمان چه بود؟
  - فکر می‌کنید با بیمار به یک نتیجه مشترک رسیدید؟
  - درمان‌های دیگری نیز وجود داشتند که شما به آن فکر کردید اما پیشنهاد نکردید؟ اگر پاسخ مثبت است، دلایل عدم استفاده از آن را شرح دهید؟
- حیطه پیگیری
  - آن‌چه شما برای پیگیری در نظر گرفتید مناسب بود؟ آیا شما طرح پیگیری را مستند کردید؟
  - چه عواملی بر تصمیمات شما تأثیرگذار بود؟
  - مراقبت جامع- مدافع سلامت
- پایش بیماری‌های مزمن
  - در مورد بیماری مزمن وی و پیشرفت آن صحبت کردید؟
  - آیا در مورد مناسب بودن راهبردهایی که به منظور پایش بیماری به کار برده‌اید، فکر می‌کنید؟
- پیشگیری و ارتقای سلامت
  - آیا در مورد مداخله‌کننده‌های پیشگیری کننده صحبت کردید (برای مثال، فشار خون، قطع سیگار، مصرف الکل، رژیم غذایی، ورزش، تست‌های غربالگری و ...)؟
  - آیا فکر می‌کنید مداخله‌هایی وجود دارند که لازم است در مورد آن‌ها بحث کنید؟

عوامل مربوط به بیمار- مدافع سلامت

- آیا موضوع به خصوصی در مورد این بیمار وجود داشت که تصمیمات شما را در مورد مدیریت بیمار تحت تأثیر قرار داد (برای مثال، ظرفیت پذیرش بیمار، سابقه پزشکی قبلی، سیستم حمایتی و اشتغال)؟
- آیا مسائل دیگری در مورد این بیمار وجود دارد که شما راغب هستید بیشتر بدانید؟
- عوامل مربوط به طبابت یا سیستم- همکار، مدیر
- آیا موضوع به خصوصی در محیط طبابت وجود داشت که تصمیمات شما را در مورد مدیریت بیمار تحت تأثیر قرار داد (برای مثال، آموزش پرستاران، شبکه مراقبت، فقدان دسترسی به خدمات)؟

چگونه می‌توانید ارائه خدمات به این بیمار را بهبود بخشید؟

اطلاع‌رسانی، آشناسازی و آموزش ارزیابان و داوطلبان

لازم است ارزیابان از بین اعضای هیأت علمی بالینی که به محتوای پرونده‌ها اشراف دارند، انتخاب شوند و به این منظور آموزش دیده باشد. همچنین لازم است دستورالعمل‌های مشخصی در ارتباط با فرایند اجرای آزمون تدوین شود و به اطلاع دست‌اندرکاران رسانده شود.

اجرای آزمون

به طور معمول، فراگیر با انتخاب تعدادی از پرونده‌های بیماران تحت نظر خود که ویزیت کرده است و در آنها یادداشت گذاشته است، آغازگر فرایند آزمون است. پرونده‌ها جهت مرور اولیه در اختیار ارزیابان قرار می‌گیرد. ارزیابان که قبل از شروع آزمون، پرونده‌های انتخاب‌شده را به منظور بررسی معیارهایی که ذکر شد و همچنین طرح سؤالات اختصاصی مناسب مطالعه می‌کنند. معمولاً بخشی از یادداشت‌های موجود در پرونده یا تمام آن برای مرور انتخاب می‌شود.

پس از مرور اولیه، عضو هیأت علمی به فراگیر اطلاع می‌دهد که پرونده‌ها بررسی شده است و سپس مواجهه بیمار مورد بحث قرار می‌گیرد. در زمان بحث ارزیاب سؤالاتی را مطرح می‌کند تا به عمق مهارت‌های استدلال بالینی، تصمیم‌گیری یا تعهد حرفه‌ای فراگیر واقف شود. ارزیاب می‌تواند از نتایج مرور اولیه به عنوان مبنایی برای بحث بیشتر استفاده کند. در برخی از فرم‌های ارزیابی سؤالاتی به عنوان نمونه به منظور راهنمای ارزیابان درج شده است.

توصیه بر آن است که جلسه با سؤالات باز شروع شود تا از پیش‌قضاوت‌ها در مورد فراگیر پرهیز شود و نیز بینش درستی از فرایند تفکر ارزیابی شونده به دست آید. تمام سؤالات با یک رویکرد غیرقضاوتی و بدون سوگیری پرسیده شود. در صورتی که لازم باشد فرایند انجام آزمون می‌تواند ضبط شود و سپس مورد تحلیل قرار گیرد. ممکن است بنا به ضرورت آزمون به صورت تلفنی برگزار شود. در نهایت عملکرد داوطلب مورد قضاوت قرار می‌گیرد. اگر هدف آزمون تراکمی باشد (مانند آزمون CBD)، عملکرد داوطلب بر اساس معیارهای مشخصی نمره‌دهی شده و در نهایت به داوطلب بازخورد داده می‌شود. در صورتی که هدف آزمون تکوینی باشد بر اساس معیارهای مشخصی نسبت به عملکرد داوطلب بازخورد ارائه می‌شود.

بازخورد در آزمون CSR دانشگاه آلبرتا

در دانشگاه آلبرتا، بازخورد «اولیه» بر اساس مرور یادداشت‌های ثبت شده در پرونده ارائه می‌شود و در بخش A فرم درج می‌شود. این بازخورد می‌تواند شامل برخی از موارد زیر یا همه آنها باشد:

- مستند سازی و خوانا بودن
- مرتبط و به هم پیوسته بودن اطلاعات ثبت شده
- ثبت پیگیری بیمار
- ارائه توصیه‌های کلی

پس از برگزاری آزمون و طرح سؤالات و بحث در مورد پرونده بیمار، مجدداً در خصوص عملکرد فراگیر بازخورد ارائه می‌شود و بازخوردها و پیشنهادها در بخش B فرم مستند می‌شوند. این بازخورد می‌تواند شامل برخی یا همه موارد زیر باشد:

- توصیه‌های کلی در مورد ارائه بیمار
- تحلیل اطلاعات و مهارت‌های استدلال
- رویکرد به مدیریت ابهام
- استفاده از پزشکی مبتنی بر شواهد
- نشان دادن طبابت بیمار محور و توانمندی‌های CanMEDS در ارتباط با پزشکی خانواده
- طبابت جامع و ارتقای سلامت
- شواهدی از بازاندیشی در طبابت

بررسی کیفیت آزمون برگزارشده

ضروری است کیفیت آزمون‌های برگزار شده با استفاده از روش‌های کمی و کیفی شبیه دیگر آزمون‌های مبتنی بر محل کار بررسی شود.

سودمندی آزمون CSR

مطالعات اندکی در ارتباط با سودمندی آزمون CSR انجام شده است و بیشتر مطالعات انجام‌شده در مورد روایی ابزار بوده است. این مسأله در مورد آزمون CBD با توجه به جدیدتر بودن آن مصداق بیشتری دارد. هر چند نتایج مطالعات انجام‌شده برای یکی از آزمون‌های فوق را می‌توان به دیگری نیز تعمیم داد.

روایی آزمون CSR

آزمون CSR بر نقش‌های واقعی که پزشکان ایفا می‌کنند، متمرکز است، در نتیجه از روایی صوری بالایی برخوردار است. علاوه بر این، شکل تکامل‌یافته‌ای از آزمون شفاهی در کنار بستر بیمار است. بنابراین به این دلیل که اصول اولیه این روش ارزیابی در اکثر دانشکده‌های پزشکی اجرا می‌شود، هدف آزمون برای ارزیابان و آزمون‌شوندگان هر دو قابل درک است.

در مورد روایی محتوا، از این جهت که روش ارزیابی CSR مهارت استدلال بالینی و تصمیم‌گیری پزشکان در محیط واقعی را ارزیابی می‌کند، ابزاری منحصر به فرد است. البته با شش آزمون در سال، به ویژه اگر شرایط بیماری مورد ارزیابی مشابه باشد، دستیابی به روایی محتوایی مطلوب غیرمحتمل به نظر می‌رسد. در واقع، یکی از مسائلی که در روایی آزمون CSR مطرح است ویژگی مورد و عدم تعمیم‌پذیری نتیجه عملکرد در یک مورد به بیماران دیگر است. جانسون و همکاران[1] (۲۰۱۱) تعداد ۱۲ مواجهه را برای رسیدن به پایایی بیش از ۰/۷۰ لازم دانستند.

چندین مطالعه از روایی معیار و روایی سازه این روش حمایت می‌کند. مطالعات اولیه در این رابطه، در خصوص برنامه‌های صدور یا تمدید گواهی بورد طب اورژانس آمریکا بودند.

- ماچ و همکاران۲ (۱۹۸۳) اطلاعات مربوط به چندین روش ارزیابی را برای گروهی از پزشکان متخصص طب اورژانس واجد شرایط تمدید مدرک بررسی کردند. بر اساس نتایج این مطالعه، CSR با روش‌های ارزیابی دیگر مانند «ممیزی مستندات طبابت» و آزمون شفاهی ارتباط داشت. توزیع نمرات و نتایج رد و قبولی با نمرات آزمون اعطای مدرک قبلی، مربوط به ۱۰ سال پیش، سازگار بود و مهم‌ترین نکته این‌که CSR از دیدگاه پزشکان شرکت‌کننده در این مطالعه معتبرترین روش ارزیابی شناخته شد.
- در مطالعه دیگری نورمن و همکاران (۱۹۸۹) عملکرد دو گروه از پزشکان (گروه داوطلب و گروهی که در طبابت مشکل داشتند) را در آزمون CSR مورد بررسی قرار دادند. این مطالعه نشان داد نتایج CSR با ارزیابی توسط بیمار استانداردشده (r=۰/۷۴) آزمون شفاهی (r=۰/۵۱) ارتباط داشت. مهم‌تر این‌که CSR توانایی تمایز دو گروه از داوطلبان را به خوبی دارا بود.
- مطالعه سولومون و همکاران۳ (۱۹۹۰) نشان داد که جزء CSR آزمون اعطای مدرک با جزء شفاهی همان آزمون (۰/۴۹) ارتباط داشت و زمانی که با آزمون شفاهی تلفیق شد، با آزمون شفاهی و کتبی که ۱۰ سال پیش برگزار شده بود، ارتباط داشت (ضریب همبستگی به ترتیب ۰/۳۷ و ۰/۴۵).

1. Johnson et al.
2. Maatsch et al.
3. Solomon et al.

پایایی آزمون CSR

یکی از مسائلی که در پایایی CSR مطرح می‌شود، ثبات قضاوت ارزیابان در مواجهه‌های مختلف به ویژه زمانی است که آزمون با هدف تراکمی به کار می‌رود. مطالعات اولیه در مورد پایایی CSR مورد استفاده توسط بورد طب اورژانس آمریکا از پایایی مناسب این روش حمایت می‌کند. در مطالعه ماینورز- والیس و همکاران[1] (۲۰۱۱)، هیچ یک از ارزیابان در یک مقیاس صفر تا پنج، نمره زیر سه ارائه نکرد. البته این مطالعه به صورت آزمایشی برگزار شد و شرکت‌کنندگان (داوطلبان) ممکن است از عملکرد خوبی برخوردار بودند اما احتمال سهل‌گیری یا سوگیری ارزیابان مطرح است. از جمله عوامل مؤثر در نمره‌دهی می‌تواند آشنا بودن ارزیابان و فراگیران یا امتناع کردن پزشکان از ارائه بازخورد منفی به همکاران خود باشد. هر چند این عوامل در مورد روش‌های مشاهده مستقیم مانند mini-CEX مورد بررسی قرار گرفته است اما بررسی آن در شرایط اجرای آزمون CSR/CBD نیز توصیه می‌شود.

یکی از روش‌های افزایش پایایی ابزارهای ارزیابی مبتنی بر داده‌های عملکردی، ترکیب آن‌ها با هم یا با دیگر ابزارهای ارزیابی مبتنی بر محل کار است. گولت و همکاران[2] (۲۰۰۷) مشاهده کردند پایایی ترکیب آزمون CSR و «ممیزی مستندات بیمار» بیشتر از پایایی «ممیزی مستندات بیمار» به تنهایی است. «ممیزی مستندات پزشکی» داده‌هایی را در مورد تشخیص‌های بالینی فراهم می‌کند و آزمون CSR منطق زیربنایی این تصمیمات و انتخاب‌هایی که رد شده‌اند را مشخص می‌کند. آزمون CSR عوامل تأثیرگذار بر تصمیم‌گیری پزشکان و نحوه تأثیرگذاری را نیز نشان می‌دهد. همچنین داده‌های جمع‌آوری شده از طریق ممیزی موجب می‌شود طیف اطلاعاتی که در CSR استفاده می‌شود، وسیع‌تر باشد.

تأثیر آموزشی آزمون CSR

هر چند فرض می‌شود آزمون CSR/CBD موجب ارتقای کیفیت مراقبت از بیمار، ارتقای بازاندیشی، یادگیری مبتنی بر طبابت و مهارت استدلال بالینی می‌شود، هنوز پژوهش‌های متقنی در این زمینه وجود ندارند. مطالعات اندکی که در این زمینه انجام شده است به بررسی دیدگاه فراگیران و اعضای هیأت علمی پرداخته‌اند که بر اساس نتایج آن تأثیر آموزشی آزمون بالا گزارش شده است (جانسون و همکاران ۲۰۱۱).

هزینه، قابلیت اجرا و مقبولیت آزمون CSR

قابلیت اجرای آزمون CBD در تمدید گواهی طبابت پزشکان در مطالعه آزمایشی توسط ماینورز- والیس بررسی شد. کل زمان صرف شده برای هر CBD از زمان آماده‌سازی تا بازاندیشی بر عملکرد، ۴۸ دقیقه طول کشید. مشارکت‌کنندگان این آزمون را در ارتقای مراقبت از بیمار مفید دانستند و آن را در فرایند تمدید گواهی طبابت الزامی دانستند. اکثر پزشکان معتقد بودند که بهتر است موارد بیماری به صورت تصادفی انتخاب شوند.

1. Mynors-Wallis et al.
2. Goulet et al.

منابع

1. Anzia JM. A clear step in the right direction Commentary on... Case-based discussion. The Psychiat. 2011;35(6):235 –6.
2. Cantillon P, Wood D. ABC of Learning and Teaching in Medicine. 2nd ed. West Sussex: John Wiely & Sons; 2010.
3. Hayden SR, Dufel S, Shih R. Definitions and competencies for practice-based learning and improvement. Acad Emerg Med 2002;9(11):1242–8.
4. Holmboe ES, Hawkins RE. Practical guide to the evaluation of clinical competence. Philadelphia: Mosby/Elsevier; 2008.
5. Goulet F. Assessment of family physicians' performance using patient charts. Interraters reliability and concordance with chart-stimulated recall interview. Eval Health Prof. 2007;30(4):376–92.
6. Jennett P, Affleck L. Chart audit and chart stimulated recall as methods of needs assessment in continuing professional health education. J Cont Educ Health Prof 1998;18(3):163–71.
7. Jennett PA, Scott SM, Atkinson MA. Patient charts and office management decisions: Chart audit and chart stimulated recall. J Cont Educ Health Prof 1995;15(1):31–9.
8. Johnson G, Booth J, Crossley J, Wade W. Assessing trainees in the workplace: results of a pilot study. Clin med. 2011;11(1):48–53.
9. Munger BS, Krome RL, Maatsch JC, Podgorny G: The certification examination in emergency medicine: An update. Ann Emerg Med. 1982;11(2):91–6.
10. Mynors-Wallis L, Cope D, Brittlebank A, Palekar F. Case-based discussion: a useful tool for revalidation. The Psychiat. 2011;35(6):230–4.
11. Schipper S, Ross S. Structured teaching and assessment A new chart-stimulated recall worksheet for family medicine residents. Canadian Family Physician. 2010;56(9):958–9.
12. Solomon DJ, Reinhart MA, Bridgham RG, Munger BS, Starnaman S. An assessment of an oral examination format for evaluating clinical competence in emergency medicine. Acad Med. 1990;65(9):S43–4.
13. Swanwick T. Uderstanding Medical Education: Evidence, Theory and Practice. West Sussex: John Wiely & Sons; 2010.
14. Williamson JML, Osborne AJ. Critical Analysis of Case Based Discussions. BJMP;2012;5(2):a514.

فصل ۲۴

لاگ‌بوک

ساختار لاگ‌بوک

از آنجا که یکی از دغدغه‌های آموزش در محیط بالینی این است که آیا فراگیران با همه مواردی که جزء الزامات برنامه آموزشی است، مواجه می‌شوند یا خیر، به ابزارهایی نیاز است تا در بررسی درون‌دادهای آموزشی کمک کنند. لاگ‌بوک که به آن «دفترچه ثبت فعالیت‌های یادگیری» نیز می‌گویند، ابزاری است که به منظور پایش محتوای درون‌دادهای آموزشی و تجربیات بالینی، مقایسه تجربیات دانشجویان در مکان‌های مختلف و تعیین میزان دستیابی فراگیران به اهداف برنامه درسی به کار می‌رود. علاوه بر این، لاگ‌بوک ظرفیت ارزشیابی برنامه و فراهم نمودن اطلاعات به منظور تغییر یا تدوین برنامه درسی را دارد. در مقایسه با ارزیابی «پیامد» یا «محصول» عملکرد (مانند آزمون‌های انتهای دوره) لاگ‌بوک با ارزیابی وسعت و عمق تجارب یادگیری و توانایی و عادات کاری دانشجویان، بازگوکننده «فرایند» آموزشی است. هر چند امروزه تأکید بر ارزیابی پیامدها افزایش یافته است، قضاوت در خصوص دستیابی به پیامدها بدون بررسی فرایند آموزشی در عمل امکان‌پذیر نیست زیرا میان محصول و فرایند آموزش ارتباط وجود دارد.

لاگ‌بوک ابزاری است که فراگیران به وسیله آن تجربیات یادگیری خود را مستند می‌کنند. محتوای دفترچه شامل اطلاعاتی در مورد مواجهه با بیمار، موارد بیماری آموزش داده‌شده، پروسیجرهای انجام شده یا اطلاعات مربوط به مکان و زمان رخداد تجربیات است. به بیان دیگر، در این روش ارزیابی محتوای یادگیری (تظاهرات بیماری، مهارت‌ها و پروسیجرها)، نحوه یادگیری (به عنوان مشاهده‌گر، مشارکت‌کننده یا انجام‌دهنده یک فعالیت؛ ثبت شرح‌حال بیمار و دریافت بازخورد از استادان) و نیز محل وقوع یادگیری (شرکت در دوره‌ها و کلاس‌های نظری، راند بالینی، آموزش بر بالین بیمار، درمانگاه و غیره) مستند می‌شوند. به این منظور لازم است بر اساس اهداف آموزشی دوره، موارد و تظاهرات بیماری و مهارت‌های بالینی که انتظار می‌رود فراگیران در انتهای دوره، توانمندی مورد نظر را در آن به دست آورند، در لاگ‌بوک فهرست شوند.

لاگ‌بوک به طور وسیع در آموزش پزشکی به ویژه دوره پزشکی عمومی و مقطع کارآموزی مورد استفاده قرار گرفته است. یکی از دلایل رویکرد دانشکده‌های پزشکی به استفاده از لاگ‌بوک پاسخگویی به الزامات نهادهای اعتباربخشی بوده است. به عنوان مثال، یکی از استانداردهای نهاد GMC در انگلیس در سال ۱۹۹۲ لزوم آگاه‌سازی دانشجویان از اهداف یادگیری بود که موجب جهت‌گیری دانشکده‌ها به سمت استفاده از لاگ‌بوک در دوره‌های کارآموزی شد. اخیراً یکی از الزامات کمیته ارتباطات در آموزش پزشکی[1] (LCME) در آمریکا و کانادا (سال ۲۰۱۲) به این شرح است که دوره‌های بالینی نه تنها باید شرایط و بیماری‌های اصلی و مهم مورد مواجهه فراگیران و نحوه مواجهه آن‌ها را مشخص کنند، بلکه باید دستیابی دانشجویان به آن را نیز تعیین کنند. به این منظور، LCME به سمت اجرای لاگ‌بوک در دانشکده‌های پزشکی در کشورهای آمریکای شمالی حرکت کرده است و هم اکنون تقریباً همه برنامه‌های کارآموزی در این کشورها از روش مذکور به منظور ثبت فعالیت‌های یادگیری کارآموزان استفاده می‌کنند.

1. Liaison Committee on Medical Education

انواع لاگ‌بوک

از لحاظ ساختار، لاگ‌بوک به سه فرم تقسیم می‌شود:

- کتبی: لاگ‌بوک‌های کتبی، دفترچه‌هایی هستند که باید اطلاعات را به صورت دستی در آنها وارد کرد. انواع کتبی شایعترین فرم مورد استفاده، حداقل در مقطع کارآموزی هستند و می‌توانند شامل انواع بدون ساختار یا ساختارمند به همراه چک‌لیست باشند. اجرای این نوع از دفترچه‌ها آسان است اما به فضایی برای ذخیره‌سازی و زمانی برای مرور و تحلیل نیاز دارند.
- قابل تصحیح توسط اسکنر: دفترچه‌هایی هستند مشتمل بر چک‌لیستی از انواع پروسیجر، مشکلات بیمار و تشخیص‌ها که مستقیماً توسط اسکنر برای تحلیل داده‌ها آماده می‌شوند. می‌توانند برخی از مشکلات مربوط به انواع کتبی را برطرف کنند اما به دلیل لزوم رعایت ساختار خاص، نمی‌توانند همه طبقه‌بندی‌ها را داشته باشند و در نتیجه از گزینه «دیگر موارد» بیش از حد استفاده می‌شود. علاوه بر این، ممکن است همه دانشجویان آن را تکمیل نکنند.
- الکترونیکی: لاگ‌بوک‌های الکترونیکی، دفترچه‌های مبتنی بر کامپیوتر، مبتنی بر شبکه، مبتنی بر اینترنت یا مبتنی بر PDA[1] هستند که دانشجویان مستقیماً داده‌ها را وارد آنها می‌کنند. استفاده از فن‌آوری دیجیتال، موجب استفاده از سیستم یکسان در دوره‌های آموزشی مختلف یک دانشکده یا دانشکده‌های مختلف می‌شود. علاوه بر این، جمع‌آوری اطلاعات، ذخیره‌سازی، تحلیل و گزارش داده‌ها آسان‌تر است.

مطالعات کمی به مقایسه انواع لاگ‌بوک در یک دانشکده پرداخته‌اند. سامنر[2] (۲۰۰۱) با مقایسه ورود داده‌ها در لاگ‌بوک کتبی بدون ساختار و مبتنی بر PDA نتیجه گرفت تعداد مشکلات بیشتری در مورد هر بیمار، در دفترچه‌های مبتنی بر PDA گزارش شده است. البته در این مطالعه مقایسه‌ای در مورد مقبولیت ابزار بین دانشجویان انجام نشد. برخی مطالعات گزارش کردند دفترچه‌های الکترونیکی توانایی ورود داده‌ها را افزایش داده‌اند و استفاده و حمل آنها از نظر دانشجویان آسان‌تر بوده است. همچنین مقایسه لاگ‌بوک بدون ساختار با لاگ‌بوک ساختارمند نشان داد در دفترچه‌های ساختارمند به احتمال بیشتری (۸۵ تا ۹۵ درصد در مقابل ۵۰ تا ۶۰ درصد) بیماران با مشکلات خاص فهرست شده‌اند. در مجموع، پژوهش‌های بیشتری به منظور مقایسه انواع لاگ‌بوک از نظر مقبولیت بین دانشجویان، میزان تکمیل آن، هزینه و ویژگی‌های روان‌سنجی توصیه می‌شود. در مورد مقبولیت، مطالعات چند مرکزی توصیه می‌شود، به این دلیل که یافته‌های یک دانشکده ممکن است قابل کاربرد در دانشکده‌های دیگر نباشد.

مزایا و محدودیت‌های لاگ‌بوک

مزایای لاگ‌بوک

- لاگ‌بوک به مثابه یک راهنما است که از همان ابتدا برای فراگیران مشخص می‌کند چه چیزی را و چگونه باید یاد بگیرند.
- فهرستی از اهداف آموزشی، تظاهرات و بیماری‌ها و مهارت‌های بالینی که باید تا انتهای دوره توانمندی لازم در آن را کسب کنند، در اختیار فراگیران قرار می‌دهد.
- فعالیت‌ها و تجربیات یادگیری را مستند می‌کند.
- امکان ارائه بازخورد از عملکرد فراگیر در خصوص میزان دستیابی به اهداف دوره را فراهم می‌کند.

1. Personal Digital Assistant
2. Sumner

- امکان ارزیابی فراگیران و نیز ارزشیابی دوره را فراهم می‌کند.
- ظرفیت تغییر و اصلاح برنامه درسی را دارد.
- موجب یکسان نمودن تجربیات یادگیری فراگیران و نظام‌مند نمودن آموزش می‌شود.
- بخشی از مسؤولیت یادگیری را به فراگیران واگذار می‌کند.

محدودیت‌های لاگ‌بوک

- دانشجویان و استادان با اهداف و اهمیت استفاده از آن آشنایی کافی ندارند و در نتیجه در تکمیل آن مشارکت لازم را ندارند.
- روایی و پایایی دفترچه‌ها به دلیل عدم اطمینان از دقت اطلاعات وارد شده در آن زیر سؤال است؛ فراگیران به غیر از موارد ضروری اطلاعات را در دفترچه‌ها وارد نمی‌کنند و موارد مواجهه را کمتر از میزان واقعی ثبت می‌کنند. بنابراین از اطلاعات موجود در لاگ‌بوک نمی‌توان به آسانی در ارزشیابی برنامه آموزشی به ویژه اهداف آموزشی و تغییر یا تأیید آن استفاده نمود.
- به طور معمول، تعداد حداقل مهارت‌ها، پروسیجرها و مواجهه‌های مورد نیاز به صورت سلیقه‌ای تعیین می‌شود که تأییدکننده عملکرد آتی فرد نیست.
- حجم تجربیات ثبت‌شده لزوماً با عملکرد فرد در دوره در ارتباط نیست. در واقع، هنوز شواهد محکمی در برقراری ارتباط بین فرایندهای ثبت شده در دفترچه‌ها و پیامدهای آموزشی یافت نشده است.

گام‌های طراحی و اجرای لاگ‌بوک مطلوب

رعایت یک سری از موارد که در جدول ۱-۲۴ خلاصه شده‌اند، به طراحی و اجرای لاگ‌بوک مطلوب کمک می‌کند.

جدول ۱-۲۴ : خلاصه مراحل طراحی و اجرای لاگ‌بوک

ردیف	عنوان	توضیح
۱	تعیین محتوای لاگ‌بوک بر اساس اهداف دوره	لازم است اهداف دوره یا بخش بالینی که قرار است دفترچه‌ها در آن مورد استفاده واقع شوند، بر اساس اهداف برنامه درسی تعیین شوند. سپس بر اساس آن، تجربیات یادگیری که فراگیران باید کسب کنند، مشخص شود.
۲	تصمیم‌گیری در خصوص فرمت لاگ‌بوک	از لحاظ ساختار، لاگ‌بوک به سه فرم کتبی، قابل تصحیح توسط اسکنر و الکترونیکی وجود دارد که لازم است قبل از طراحی لاگ‌بوک به توجه به امکانات، نیروی متخصص دردسترس و فرهنگ مؤسسه درخصوص آن تصمیم‌گیری شود.
۳	تعریف حداقل سطح مشارکت مورد انتظار از فراگیران در هر یک از فعالیت‌های یادگیری	به منظور یکسان نمودن تجربیات و اطمینان از دستیابی به اهداف آموزشی دوره، میزان مشارکت مورد انتظار از فراگیران در هر یک از تجربیات یادگیری باید مشخص گردد که می‌تواند به صورت مشاهده تجربیات یادگیری، مشارکت در انجام آن یا انجام مستقل آن تعریف شود.
۴	تعیین حداقل تعداد مهارت، پروسیجر و مواجهه مورد نیاز	به طور معمول، تعداد موارد مورد انتظار از هر فعالیت یادگیری، توسط متخصصان و پس از رسیدن به اجماع تعیین می‌شود.
۵	تصمیم‌گیری در مورد تناوب تکمیل اطلاعات	زمان تکمیل اطلاعات می‌تواند به صورت روزانه، هفتگی یا در پایان دوره و در برخی از موارد تلفیقی از این موارد باشد.
۶	تصمیم‌گیری در مورد تناوب و نحوه ارائه بازخورد در تکمیل اطلاعات	تکمیل اطلاعات می‌تواند با راهنمایی و بازخورد استادان مشاور یا بدون آن باشد. در صورت راهنمایی برای تکمیل اطلاعات، زمان و نحوه آن باید مشخص شود. ارائه بازخورد می‌تواند به صورت کتبی یا به صورت حضوری باشد.
۷	تهیه راهنماهای تکمیل لاگ‌بوکw	توصیه می‌شود در ابتدای دفترچه، بخشی به عنوان راهنمای استفاده از آن تعبیه شود. هر چند لازم است توضیحات شفاهی و آموزش‌های لازم در مورد نحوه استفاده و تکمیل دفترچه‌ها از همان ابتدای دوره در اختیار فراگیران قرار گیرد.
۸	تکمیل لاگ‌بوک توسط فراگیر و بررسی آن	دفترچه‌ها در ابتدای دوره همراه با راهنمایی و اطلاع‌رسانی کافی در اختیار فراگیران قرار می‌گیرد. اطلاعات مطابق دستورالعمل از قبل تهیه شده و در انتهای یک دوره زمانی مشخص جمع‌آوری می‌شوند و اطلاعات ثبت شده در آن مورد بررسی قرار می‌گیرد.
۹	بررسی کیفیت لاگ‌بوک	ضروری است با استفاده از روش‌های کمی و کیفی کیفیت لاگ‌بوک‌های تکمیل شده مورد بررسی قرار کیرد.

تعیین محتوای لاگ‌بوک بر اساس اهداف دوره

همان‌طور که اشاره شد، هدف اصلی این روش ارزیابی، نظام‌مند و یکسان کردن تجربیات آموزشی است. بنابراین لازم است اهداف دوره یا بخشی که قرار است دفترچه‌ها در آن مورد استفاده واقع شوند، بر اساس اهداف برنامه درسی تعیین شوند. سپس بر اساس آن، تجربیات یادگیری که فراگیران باید کسب کنند، مشخص شود. در این ارتباط، دفترچه‌ها با الگوهای مختلف بر اساس نیازها و اهداف برنامه آموزشی طراحی شده است. به عنوان مثال، برخی از دفترچه‌ها به ویژه در بخش‌های جراحی تنها شامل پروسیجرهای ضروری دوره است، در حالی‌که دفترچه‌های دیگر می‌توانند شامل انواع مختلفی از تجربیات دیگر مانند بیماری‌ها و تظاهرات بالینی ضروری، پروسیجرها و مهارت‌های بالینی مورد نیاز،

فرصت‌های یادگیری (راند بالینی، گزارش صبحگاهی، آموزش بر بالین بیمار و ...)، آزمون‌های دوره و نیز بخش‌هایی برای دریافت بازخورد و پیشنهادها باشد. محتوای دفترچه هر چه باشد باید بر اساس اهداف دوره و توسط متخصصان دوره آموزشی و در یک فرایند رسیدن به اجماع و توافق تعیین شود.

بخش‌های لاگ‌بوک دانشگاه ناتینگهام در انگلیس (دنیک[1] ۲۰۰۰):

- فهرستی از اهداف آموزشی
- فهرستی از موارد بیماری ضروری که باید مشاهده شود
- فهرستی از مهارت‌های بالینی که باید کسب شود
- فهرستی از منابع مطالعه
- فضایی به منظور مشاهدات فردی
- فضایی به منظور دریافت پیشنهادها و بازخورد میان ترم
- معیارهای ارزیابی
- فرم بازخورد و ارزیابی نهایی

1. Dennick

تجربه دانشکده پزشکی گرونینگن[1] در هلند: مقایسه تئوری و عمل

راخوبار-کریگر و بندر[2] (۱۹۹۷) ۲۴۰ دفترچه تکمیل‌شده توسط کارآموزان در بخش‌های بالینی طب داخلی، بیماری‌های اعصاب، جراحی، اطفال، زنان و زایمان و روان‌پزشکی را در دانشکده پزشکی گرونینگن بین سال‌های ۱۹۸۹ تا ۱۹۹۳ بررسی کردند (۴۰ دفترچه از هر چرخش بالینی). دفترچه‌ها شامل ثبت موارد بیماری و نشانه‌ها و علائم فیزیکی هر مورد بیماری که دانشجویان مشاهده می‌کردند، بود. سپس محتوای دفترچه‌ها با اهداف برنامه درسی کشوری که در سال ۱۹۹۴ اعلام شد مقایسه شد. نتایج مقایسه در سه گروه طبقه‌بندی شد:

- همپوشانی کامل بین اهداف برنامه درسی و محتوای دفترچه‌ها
- همپوشانی ناقص: به صورت اهداف آموزشی که در دفترچه‌ها مستنداتی در مورد آن وجود نداشت
- همپوشانی مازاد: به صورت تجربیات آموزشی که در دفترچه‌ها ثبت شده است اما در اهداف آموزشی وجود نداشت

نتایج نشان داد در هیچ‌یک از کارآموزی‌ها، میانگین تجربیات ثبت شده در دفترچه‌ها حد نصاب مورد نیاز اهداف برنامه درسی را رعایت نکردند، هر چند در اکثر کارآموزی‌ها میزان تجربیات آموزشی ثبت شده در دفترچه‌ها از سقف الزامات برنامه درسی بیشتر بود. علاوه بر آن نتایج نشان داد که تجربیات دانشجویان در مورد نوع و تعداد موارد بیماری بسیار متنوع بود. به عنوان مثال، دانشجوی X، با ۲۰ مورد بیمار مبتلا به انواع متفاوت کانسر مواجهه داشته و دانشجوی ۷، تنها با یک نوع از بیماری کانسرِ مواجهه داشته است. در عین حال، دانشجوی ۷ با ۲۰ مورد بیماری از همان یک نوع کانسر مواجهه داشته است، در حالی که دانشجوی X ابداً با بیمار مبتلا به این نوع از کانسر مواجهه نداشته است. بنابراین، عدم انطباق بین تجربه‌های ضروری و آن‌چه در عمل رخ می‌دهد دو جنبه داشت:

- در حالت اول، دانشجویان مواردی را تجربه کردند که در اهداف برنامه درسی تنها در حد کسب دانش تئوری در آن مورد بیماری ضروری ذکر شده بود. به عنوان مثال، بخش نظری tap مایع پلور جزء الزامات اهداف دوره کارآموزی بود اما در عمل بیش از ۷۰ درصد دانشجویان تجربه مشاهده این مهارت را داشتند. به طور مشابهی، دانشجویان در بسیاری از موقعیت‌ها تشخیص و درمان بیماری را تجربه کرده بودند که تنها کسب دانش نظری آن در اهداف برنامه درسی ذکر شده بود.
- در حالت دوم، دانشجویان فرصت مواجهه با بیمارانی را که در برنامه درسی جزء الزامات بود نداشتند. هر چند این احتمال وجود داشت که در بخش دیگری این فرصت برای آن‌ها فراهم شود.

البته با توجه به نتایج این مطالعه نمی‌توان در مورد تطابق یا عدم تطابق تجربیات کارآموزان و اهداف آموزشی نتیجه‌گیری کرد به این دلیل که تکمیل دفترچه‌ها قبل از اعلام اهداف برنامه درسی رخ داده است. اما در مورد موارد بیماری و شرایطی (تجربیات) که دانشجویان با آن مواجه هستند، اطلاعات خوبی در اختیار قرار می‌دهد.

1. Groningen
2. Raghoebar-Krieger & Bender

تصمیم‌گیری در خصوص فرمت لاگ‌بوک

همان‌طور که اشاره شد لاگ‌بوک از لحاظ ساختار به سه فرم کتبی، قابل تصحیح توسط اسکنر و الکترونیکی وجود دارد که با توجه به امکانات در دسترس و فرهنگ مؤسسه می‌توان در مورد آن تصمیم‌گیری کرد.

تعریف حداقل سطح مشارکت مورد انتظار از فراگیران در هر یک از فعالیت‌های یادگیری

با توجه به ماهیت تصادفی آموزش و یادگیری در محیط بالینی این احتمال وجود دارد که فراگیران در برخی از فعالیت‌هایی که تنها کسب دانش نظری در آنها مورد نیاز است، به طور مکرر شاهد آن باشند، درگیر اجرای آن شوند و حتی مستقلاً آنها را انجام دهند اما برخی از فعالیت‌هایی را که باید به صورت مستقل اجرا کنند، اصلاً در طول دوره تجربه نکنند. بنابراین به منظور یکسان نمودن تجربیات و اطمینان از دستیابی به اهداف آموزشی دوره، میزان مشارکت مورد انتظار از فراگیران در هر یک از تجربیات یادگیری باید مشخص گردد. به عنوان مثال، در لاگ‌بوک کارورزان طب اورژانس بیمارستان امام خمینی (دانشگاه علوم پزشکی تهران)، میزان مشارکت در سه سطح تعریف شده است (فرهمند و اصل سلیمانی ۱۳۸۹):

- مشاهده انجام پروسیجر
- مشارکت در انجام پروسیجر
- انجام مستقل پروسیجر

تعیین حداقل تعداد مهارت، پروسیجر و مواجهه مورد نیاز

به طور معمول، تعداد موارد مورد انتظار از هر فعالیت یادگیری، توسط متخصصان و پس از رسیدن به اجماع تعیین می‌شود. اگرچه اکثراً در مورد تعداد بیماری‌ها و پروسیجرهای لازم، به طوری که پیش‌بینی‌کننده عملکرد آتی فرد باشد، شواهد محکمی در اختیار نیست:

تجربه دانشکده پزشکی گرونینگن در هلند (راخوبار-کریگر و بندر ۲۰۰۱)

لاگ‌بوک در این دانشکده، توسط تعدادی از اعضای هیأت علمی و بر اساس اهداف برنامه درسی ملی اعلام شده از طرف وزارت بهداشت کشور هلند در سال ۱۹۹۴، به منظور ارزیابی در دوره کارورزی طراحی شد. دفترچه‌ها با هدف پوشش اهداف زمان فارغ‌التحصیلی تدوین شد. اهداف برنامه درسی در سه بخش طبقه‌بندی شده بودند:

- اهداف عمومی: شامل دانش، مهارت و نگرش مورد نیاز برای هر پزشک
- مشکلات مربوط به بیمار: فهرستی از مشکلات که لازم است یک پزشک توانایی مواجهه و مدیریت آن را کسب کند
- اهداف مربوط به هر تخصص: شامل فهرستی از مهارت‌ها و بیماری‌های مربوط به هر رشته تخصصی

دو مورد اول مربوط به یک دوره خاص نبودند و در طول دوره پزشکی به دست می‌آمدند. بنابراین، لاگ‌بوک بر اساس اهداف مربوط به هر تخصص طراحی شد. به عنوان مثال، در روتیشن کارورزی بیماری‌های داخلی، بیماری‌ها به ۱۳ مجموعه (برای مثال، بیماری‌های خون) و هر مجموعه به چندین زیر مجموعه (مانند کم خونی) و هر زیرمجموعه به چندین بیماری (به عنوان مثال، کم خونی فقر آهن) تقسیم شد. در نهایت ۲۲۸ بیماری فهرست شد که ۹۲ مورد آن به عنوان موارد ضروری که دانشجویان ملزم به کسب تجربه در آن مورد هستند در نظر گرفته شدند.

علاوه بر این، برای هر یک از موارد ضروری سطح مشارکت در فعالیت‌های یادگیری مشخص شد. فعالیت‌های یادگیری مربوط به موارد بیماری در سه سطح تعریف شد:

- **سطح ۱**: شرکت در دوره‌ها و سخنرانی‌های برگزار شده
- **سطح ۲**: شرکت در راند بالینی و آموزش بر بالین بیمار
- **سطح ۳**: تشخیص و درمان بیمار به صورت ثبت شده در پرونده بیمار

فعالیت‌های یادگیری مربوط به پروسیجرها و مهارت‌های عملی نیز در دو سطح به صورت سطح ۱ به عنوان «مشاهده‌گر» و سطح ۲ «انجام مهارت» تعریف شد.

با توجه به این‌که فهرست موارد بیماری تعیین‌شده طولانی بود و امکان تجربه همه موارد در بخش داخلی امکان پذیر نبود، در مرحله بعد به منظور اطمینان از تجربه طیف وسیعی از بیماری‌ها توسط فراگیران، حداقل یک مورد از هر طبقه به عنوان نماینده آن طبقه انتخاب شد. علاوه بر آن حداقل تعداد مورد نیاز برای اطمینان از دستیابی به تجربه کافی در مورد هر نماینده تعیین شد. توضیحات لازم در مورد نحوه تکمیل دفترچه‌ها در اختیار کارورزان قرار گرفت. به عنوان مثال، به آن‌ها توضیح داده شد اگر بیماری با درد قفسه سینه پذیرفته شد اما دیابت و پرفشاری خون هم داشت، دانشجویان هر سه مورد را وارد دفترچه‌ها کنند.

تصمیم‌گیری در مورد تناوب تکمیل اطلاعات

زمان تکمیل اطلاعات می‌تواند به صورت روزانه، هفتگی یا در پایان دوره باشد و در برخی از موارد به صورت تلفیقی از این موارد است. به عنوان مثال، موارد بیماری مربوط به درمانگاه بیماران سرپایی به صورت روزانه و موارد بیماری بستری در بخش به صورت هفتگی در دفترچه ثبت می‌شوند.

تصمیم‌گیری در مورد تناوب و نحوه ارائه بازخورد در تکمیل اطلاعات

تکمیل اطلاعات می‌تواند با راهنمایی و بازخورد استادان مشاور یا بدون آن باشد. در صورت راهنمایی برای تکمیل اطلاعات، زمان و نحوه آن باید مشخص شود. راهنمایی می‌تواند به صورت هفتگی یا یک‌بار در وسط دوره و سپس در انتهای دوره باشد. به طور معمول، راهنمایی به صورت بررسی اطلاعات وارد شده و ارائه بازخورد در مورد آن صورت می‌گیرد. راهنمایی باعث می‌شود دقت اطلاعات ثبت شده افزایش یابد. ارائه بازخورد می‌تواند به صورت کتبی در فضایی که در دفترچه‌ها به همین منظور در نظر گرفته شده یا به صورت حضوری باشد. در برخی موارد نیز، امضای استاد راهنما جهت تأیید مهارت انجام شده یا مشاهده مواجهه ضرورت دارد.

تهیه راهنماهای تکمیل لاگ‌بوک

برای اینکه دانشجویان بدانند چگونه باید لاگ‌بوک خود را تکمیل کنند، لازم است راهنماهایی به همین منظور برای ایشان تدوین شود و از ابتدای دوره در خصوص قسمت‌های مختلف آن توجیه شوند. توصیه می‌شود در ابتدای دفترچه، بخشی به عنوان راهنمای استفاده از آن تعبیه شود. هر چند لازم است توضیحات شفاهی و آموزش‌های لازم در مورد نحوه استفاده و تکمیل دفترچه‌ها از همان ابتدای دوره در اختیار فراگیران قرار گیرد. اطلاع‌رسانی فراگیران منجر به کاهش داده‌های از دست رفته و افزایش دقت داده‌های گزارش شده می‌شود. این مسأله نیز که ثبت ناقص اطلاعات یا ارائه اطلاعات نادرست چه عواقبی دارد، باید به اطلاع فراگیران برسد.

تکمیل لاگ‌بوک توسط فراگیر و بررسی آن

دفترچه‌ها در ابتدای دوره همراه با راهنمایی و اطلاع‌رسانی کافی در اختیار فراگیران قرار می‌گیرد. اطلاعات مطابق دستورالعمل از قبل تهیه شده به صورت روزانه، هفتگی یا در پایان دوره توسط دانشجویان تکمیل می‌شود. ممکن است طی این فرایند از راهنمایی و بازخورد استادان مشاور نیز استفاده شود. در نهایت دفترچه‌ها در انتهای یک دوره زمانی مشخص جمع‌آوری می‌شوند و اطلاعات ثبت شده در آن مورد بررسی قرار می‌گیرد.

بررسی کیفیت لاگ‌بوک‌های تکمیل شده

تمام مواردی که در طراحی و اجرا ذکر شد، در ارتقای کیفیت لاگ‌بوک بسیار مؤثر است. به عنوان مثال، تکمیل لاگ‌بوک‌ها با راهنمایی استاد مشاور و بررسی دوره‌ای آن می‌تواند موجب ارتقای آن شود. با این وجود بررسی میزان تکمیل لاگ‌بوک‌ها و بررسی موانع احتمالی آن می‌تواند در ارتقای کیفیت آن کمک کننده باشد. همچنین توجه به سنجش روایی و پایایی لاگ‌بوک اهمیت دارد. این کار از طرق مختلف امکان‌پذیر است که در قسمت مربوط به سودمندی لاگ‌بوک مورد بحث قرار خواهد گرفت.

سودمندی لاگ‌بوک

روایی لاگ‌بوک

انتخاب تجربیات یادگیری بر اساس اهداف آموزشی دوره و از طریق به اجماع رسیدن متخصصان، روایی صوری و محتوایی دفترچه را افزایش می‌دهد. هر چند هنوز در مورد تعداد حداقل مواجهه مورد نیاز شواهد محکمی وجود ندارد.

چالش مهمی که در مورد اطلاعات درج شده توسط فراگیران در دفترچه‌ها وجود دارد، دقت آن است. یکی از سؤالات مهم در مورد لاگ‌بوک این است که اطلاعات آن تا چه اندازه قابل اعتماد است و آیا می‌توان با توجه به آن تجربیات یادگیری را اصلاح کرد. مطالعات متعددی نشان داده‌اند که دانشجویان در تکمیل اطلاعات خود در دفترچه‌ها، همه تجربیات را وارد نمی‌کنند و توافق بین اطلاعات وارد شده توسط فراگیران و استادان بخش کم تا متوسط بوده است. این یافته‌ها استفاده از اطلاعات موجود در دفترچه‌ها را به عنوان مبنایی برای اصلاح اهداف و برنامه آموزشی مورد تردید قرار می‌دهد. در مورد دقت اطلاعات وارد شده در دفترچه‌ها در بخش پایایی ابزار با جزئیات بیشتری اشاره می‌شود.

در خصوص روایی معیاری باید گفت که انتظار می‌رود که بین حجم اطلاعات ثبت شده در لاگ‌بوک و عملکرد داوطلب در دیگر روش‌های ارزیابی ارتباط قابل ملاحظه‌ای وجود داشته باشد. به این دلیل که احتمالاً تجربه و مواجهه بیشتر، منجر به بهبود توانمندی فراگیر می‌شود و همچنین احتمالاً تعداد بیمار بیشتر منجر به ارتباط بیشتر با استادان بخش و دریافت بازخورد بیشتری می‌شود. این در حالی است که نتایج همه مطالعات، حمایت‌کننده این ارتباط نیست.

- مطالعه لمپ و همکاران[1] (۲۰۰۸) نشان داد فراگیرانی که به واسطه لاگ‌بوک به سمت انتخاب بیماران (۱۰ بیمار ضروری در بخش اورژانس) هدایت شده بودند، در مقایسه با سیستم سنتی انتخاب بیمار عملکرد بهتری در آزمون شناختی پایان بخش داشتند.
- نیومایر و همکاران[2] (۱۹۹۸) نشان دادند ارتباطی بین فعالیت‌های مختلف آموزشی درج شده در لاگ‌بوک با نمرات آزمون پایان ترم (تشریحی و چندگزینه‌ای) وجود ندارد.
- مطالعه‌ای در بخش نورولوژی نشان داده است که بین تعداد مواجهه‌های ثبت شده در دفترچه‌ها و نمره‌دهی کلی انتهای بخش ارتباطی وجود نداشته است (پواسون و همکاران[3] ۲۰۰۹).
- ضریب همبستگی بین حجم تجربیات بالینی گزارش‌شده در دفترچه‌ها و نمرات آزمون انتهای دوره کارآموزی ضعیف (۰/۴) بود. دلیل ارتباط ضعیف می‌تواند این باشد که نمرات آزمون‌های انتهای دوره تحت تاثیر عوامل دیگری مانند مهارت‌های ارتباطی، مهارت‌های نوشتاری و ... قرار گرفته باشد. در این راستا، توصیه می‌شود پژوهش‌های بیشتری انجام شود که عوامل تأثیرگذار دیگر بر عملکرد انتهای دوره را حذف کنند (هانگ و همکاران[4] ۲۰۱۲).
- شادنی و همکاران[5] (۱۹۹۶) به بررسی ارتباط متغیرهای مختلف ثبت شده در دفترچه‌ها و نمرات آزمون OSCE دانشجویان پرداختند. دانشجویانی که تجربیات یادگیری بیشتری در درمانگاه داشتند، در حالی که پیش از ورود به بخش نمرات بالاتری داشتند، نمرات OSCE پایین‌تری گرفتند. بازخوردهای ارائه‌شده با کیفیت بالا در زمان پذیرش بیماران اورژانسی با نمرات OSCE ارتباط داشتند اما بازخوردهای ارائه شده در خصوص پیگیری بیمار با نمرات OSCE پایین‌تری همراه بودند.

در مجموع، بر اساس شواهد موجود می‌توان گفت ارتباط بین اطلاعات ثبت‌شده در دفترچه‌ها (به عنوان یک ابزار اندازه‌گیری

1. Lamp et al.
2. Neumayer et al.
3. Poisson et al.
4. Haung
5. Chatenay

فرایند آموزشی) با پیامدهای آموزشی ضعیف است. البته ابتدا باید دقت و اعتبار داده‌های ثبت شده در دفترچه‌ها تأیید شود. به طور خلاصه، هنوز روایی ابزار به عنوان آن‌چه قرار بوده اندازه‌گیری کند، مورد شک و تردید است.

پایایی لاگ‌بوک

برخلاف سایر ابزارهایی که تاکنون دیدیم، وقتی صحبت از پایایی لاگ‌بوک می‌شود، بیشتر منظور دقت اطلاعات وارد شده است و نه دقت ابزار در ارائه نمرات دانشجویان. به عبارت دیگر در اینجا، بیشتر از اینکه ثبات در تصحیح و نتایج ارزیابی و نمرات مدنظر باشد، ثبات در ثبت داده‌های عملکردی مطرح است. به این علت که عموماً لاگ‌بوک بیشتر برای ارزیابی تکوینی به کار می‌رود و هدف آن تراکمی نیست که به دنبال تصحیح آن، نمره داده شود و در خصوص رد و قبول دانشجو تصمیم‌گیری شود.

یکی از روش‌های تعیین پایایی لاگ‌بوک، تخمین پایایی بین مشاهده‌گران متفاوت است. منظور از مشاهده‌گران متفاوت، پزشکان باتجربه (مانند استادان بالینی یا دستیاران ارشد) به عنوان یک گروه و فراگیرانی که مسؤول تکمیل دفترچه‌ها هستند، به عنوان گروه دیگر هستند. معمولاً دفترچه‌های تکمیل‌شده توسط پزشکان با تجربه به عنوان استاندارد در نظر گرفته می‌شود و با محاسبه ضریب توافق، میزان توافق بین دو گروه به عنوان پایایی بین مشاهده‌گران در نظر گرفته می‌شود. هدف این است که مشخص شود آیا مشاهده‌گران متفاوت، با استفاده از دفترچه‌های مشابه و در گروه آموزشی یکسان، اطلاعات یکسانی را ثبت می‌کنند یا خیر.

در صورت وجود عدم توافق، می‌توان با محاسبه حساسیت[1] و ویژگی[2] تعیین کرد چه منابعی از خطا موجب آن شده است. کیفیت داده‌های ثبت شده از لحاظ عدم ثبت موارد مشاهده نشده، ویژگی نام دارد و به ثبت تمام مواردی که امکان مشاهده آن در دوره وجود داشت، حساسیت گفته می‌شود. از بین این دو، به صورت کلی، ویژگی شاخص مهمتری برای کیفیت داده‌ها است.

در زیر نتایج برخی از پژوهش‌هایی که به بررسی پایایی لاگ‌بوک پرداخته‌اند، به طور خلاصه آمده است:

- در این ارتباط، لینکز و همکاران[3] (۱۹۸۸) توافق کم تا متوسطی (ضریب کاپا: ۰/۳۳ تا ۰/۴۹) را در مستند کردن مشکلات، شرایط و مهارت‌ها در لاگ‌بوک، بین دانشجویان و استادان راهنما مشاهده کردند. هر چند آن‌ها ذکر کردند که دفترچه‌های دانشجویان اطلاعات دقیقی در اختیار برنامه‌ریزان آموزشی قرار می‌دهد.
- راخوبار–کریگر و همکاران (۲۰۰۱) میزان تطابق اطلاعات وارد شده در دفترچه‌ها توسط استادان بخش (دو نفر) و کارورزان را در دوره کارورزی داخلی با هم مقایسه کردند. اطلاعات به تفکیک هر هفته مقایسه شد. میزان توافق بین استادان و دانشجویان برای همه بیماری‌ها پایین به دست آمد (ضریب جاکارد[4]: ۰/۲۳) و همچنین برای بیماری‌های ضروری نسبتاً پایین (ضریب جاکارد: ۰/۳۶) گزارش شد.
- توافق بین مشاهده‌گران (ضریب کاپا) در فرمت الکترونیکی مورد استفاده در دانشکده دارموت[5] ۰/۶۸ بود (دنتون و همکاران[6] ۲۰۰۶).
- در زیر نتایج برخی از پژوهش‌هایی که به بررسی دقت اطلاعات وارد شده در لاگ‌بوک پرداخته‌اند، به طور خلاصه آمده است:
- در مطالعه راخوبار–کریگر و همکاران (۲۰۰۱)، حساسیت دفترچه‌ها برای مجموع موارد به طور متوسط ۰/۳۶ و برای

1. Sensitivity
2. Specificity
3. Links et al.
4. Jaccard
5. Dartmouth
6. Denton et al.

موارد ضروری ۰/۵۱ بود. ویژگی دفترچه‌ها برای مجموع موارد و موارد ضروری به ترتیب ۰/۹۶ و ۰/۹۳ بود. یافته‌های این مطالعه نشان داد که داده‌های لاگ‌بوک از ویژگی بالا و حساسیت پایین برخوردار بودند.

- در دفترچه‌های الکترونیکی میزان موارد حذف‌شده توسط دانشجویان در مقایسه با مشاهده‌گران با تجربه، ۴۰ درصد و ویژگی آن بسیار خوب و ۹۵ درصد گزارش شد (دنتون و همکاران ۲۰۰۷).
- در مطالعه دیگری، بررسی دقت داده‌ها به صورت مقایسه اطلاعات ثبت‌شده در لاگ‌بوک (به صورت مجموعه‌ای از فرم‌های اسکن‌شده) و اطلاعات پرونده بیماران (ثبت شده توسط پزشکان بخش) انجام شد و میزان توافق در مورد تشخیص‌های اصلی ۷۷ درصد بود (رتنر و همکاران[1] ۲۰۰۱).
- بر اساس نتایج یک مطالعه مروری، دقت اطلاعات ثبت‌شده در لاگ‌بوک‌ها از ۳۶ درصد تا ۷۷ درصد متغیر است (دنتون و همکاران ۲۰۰۶).

به نظر می‌رسد دو عامل مهمی که در میزان توافق بین مشاهده‌گران مؤثر هستند، میزان تجربه فراگیران و پیچیدگی موارد بیماری است. به این معنی که دانشجویان به این دلیل که توانایی شناسایی یک مشکل را ندارند، آن را در دفترچه ثبت نمی‌کنند، یا وقتی با یک بیمار پیچیده روبه رو می‌شوند یا فهرستی بلند از موارد را در دفترچه الکترونیکی خود می‌بینند، موارد کم اهمیت را گزارش می‌کنند یا تنها به موارد درج شده در بالای فهرست توجه می‌کنند.

در مورد دفترچه‌های الکترونیکی مشکلات نرم‌افزاری نیز می‌تواند عدم ورود اطلاعات را تشدید کند. علاوه بر این، کیفیت بازخوردهای ارائه شده نیز می‌تواند در افزایش تعداد بیماران مشاهده‌نشده مؤثر باشد. بنابراین، با توجه به پایین بودن حساسیت ابزار، اطلاعات موجود در لاگ‌بوک نباید مبنای تصمیم‌گیری‌های مهم قرار گیرد یا به عنوان ابزاری با هدف تراکمی به کار گرفته شود (دنتون و همکاران ۲۰۰۶).

تأثیر آموزشی لاگ‌بوک

بخشی از اثرات لاگ‌بوک مربوط به فرد (به عنوان مثال، فراگیران و استادان) و بخش دیگر آن مربوط به فرایند برنامه آموزشی است. پژوهش‌ها نشان داده‌اند دانشجویانی که به تکمیل لاگ‌بوک پرداخته‌اند در مقایسه با گروه کنترل، با بیماران بیشتری مواجهه داشته‌اند، پروسیجرهای جراحی بیشتری انجام داده‌اند و با بیماران بدحال بیشتری روبه رو شده‌اند (دنتون و دارنینگ ۲۰۰۹).

تأثیر لاگ‌بوک بر برنامه‌های آموزشی به این صورت بوده است که از نتایج آن به منظور تغییر محتوای برنامه درسی استفاده شده است اما ارزیابی مجدد در خصوص موفقیت‌آمیز بودن این تغییرات صورت نگرفته است.

ارزشیابی برنامه بر اساس اطلاعات لاگ‌بوک (مطالعه مروری دنتون و همکاران ۲۰۰۶)

لاگ‌بوک می‌تواند به منظور مقایسه تجربیات یادگیری در یک یا چند دوره آموزشی (مانند یک چرخش بالینی کارآموزی) و نیز میزان دستیابی فراگیران به اهداف آموزشی برنامه درسی مورد استفاده قرار گیرد. در اکثر مطالعاتی که با این هدف انجام شده‌اند، فراگیران در معرض تجربیات آموزشی یکسانی در راستای اهداف آموزشی دوره قرار نگرفته‌اند.

- مقایسه یک دوره کارآموزی در مکان‌های متفات: بسیاری از دانشکده‌ها یک دوره آموزشی مشخص مانند کارآموزی یا کارورزی رشته خاصی را در بخش‌ها یا بیمارستان‌های مختلفی ارائه می‌دهند که از لحاظ ویژگی‌های دموگرافیک بیماران، توزیع موارد بیماری و موارد دیگر می‌تواند متفاوت باشد. بخش بیماران بستری و بیماران سرپایی نیز از لحاظ محتوایی می‌تواند بسیار متفاوت باشد. مقایسه دفترچه‌های بخش‌های بستری با سرپایی و کارآموزی در مکان‌های مختلف نشان داد:
  - تشخیص‌های اصلی بیماران ویزیت شده در درمانگاه از بیماران بستری متفاوت بود.
  - فراگیران در درمانگاه بیشتر آموزش بیمار، مشاوره و پیگیری بیمار را یاد گرفتند و در بخش بیشتر پاتوفیزیولوژی بیماری‌ها را آموختند.
  - آموزش سرپایی فرصت‌های یادگیری بیشتری در اختیار فراگیران قرار داد.
  - دانشجویان در درمانگاه‌های سطح شهر نسبت به بخش بیماران بستری تعداد بیماران بیشتری مشاهده کردند و تعداد پروسیجر بیشتری انجام دادند.

1. Rattner et al.

- ب‌ین سایت‌های مختلف کارآموزی از نظر اطلاعات دموگرافیک بیماران مشاهده شده و تشخیص‌های رایج تفاوت وجود داشت اما بین تجربیات یادگیری سال‌های پیاپی در مدت پنج سال تفاوتی وجود نداشت.
- آموزش سرپایی (با مقایسه امتحانات پایان بخش) موجب کسب دانش کمتری شد.
- فراگیران در بخش‌های کارآموزی بیمارستان‌های مختلف مدت زمان حضور متفاوتی در بیمارستان داشتند و زمان صرف‌شده در مراقبت از بیمار و مطالعه مستقل نیز در بیمارستان‌های مختلف متفاوت بود. با ارتقای سال تحصیلی فراگیران زمان کمتری را صرف مراقبت از بیمار و زمان بیشتری را صرف فعالیت‌های آموزشی می‌کردند.
- بیماری‌های مشاهده‌شده از لحاظ شدت بیماری در محل‌های مختلف، متفاوت بود.
- مقایسه دوره‌های کارآموزی مختلف: مقایسه دفترچه‌های کارآموزی‌های مختلف نشان داد:
- میزان مشارکت فراگیران در تجربیات یادگیری متفاوت بود.
- درمانگاه‌های پزشکی خانواده مکمل چرخش بالینی طب داخلی بود.
- چرخش بالینی پزشکی خانواده، بیماری‌های اطفال و طب داخلی مکمل یکدیگر بودند؛ فراگیران در بخش اطفال با علائم و نشانه‌های بیشتری مواجه شدند، در پزشکی خانواده، مشاوره و پروسیجرهای بیشتری را انجام دادند و در بخش داخلی با تشخیص‌های اختصاصی بیشتری مواجه شدند.
- در بخش‌های جراحی، هدایت مستقیم فراگیران بیشتر دیده شد.
- تعیین میزان دستیابی به اهداف دوره یا اهداف برنامه درسی:
- فراگیران طیف وسیعی از اهداف آموزشی را تجربه کردند اما توزیع تجربیات مناسب نبود به طوری که برخی از اهداف ضروری را کمتر یا به ندرت تجربه کردند و برخی از اهداف غیرمهم را به کرات تجربه کردند.
- پایش زمان صرف‌شده توسط فراگیران در محیط بالینی: عموماً مسؤولان دوره‌های آموزشی به دنبال این هستند که بدانند فراگیران در محیط بالینی وقت خود را صرف چه اموری می‌کنند و آیا از آن استفاده بهینه می‌کنند. محتویات دفترچه‌های ثبت فعالیت‌های یادگیری فراگیران نشان داد:
- دانشجویان در بخش‌های بالینی، مدت بیشتری را صرف تعامل هدایت‌نشده با بیمار می‌کردند، در حالی‌که در درمانگاه بیشتر به ارزیابی بیمار می‌پرداختند.
- دانشجویان تنها بخشی از زمان خود را صرف شرح‌حال‌گیری و معاینه بیمار می‌کردند و در بیشتر موقعیت‌های آموزشی غیرفعال بودند، بازخورد و هدایتی دریافت نمی‌کردند و پروسیجرهای کمی انجام می‌دادند.
- همان‌طور که مشاهده می‌شود، اطلاعات ثبت شده در لاگ‌بوک می‌تواند داده‌های غنی در خصوص ارزشیابی برنامه‌های آموزشی در اختیار برنامه‌ریزان قرار دهد در صورتی که اعتبار داده‌های ثبت شده به نحوی تضمین شود.

هزینه، قابلیت اجرا و مقبولیت لاگ‌بوک

پژوهش‌های مختلف در مورد میزان مقبولیت لاگ‌بوک نتایج متفاوتی را گزارش کرده‌اند. به طوری‌که میزان تکمیل دفترچه‌ها از ۵۸ درصد تا ۹۰ درصد متغیر بوده است. از جمله عوامل احتمالی تأثیرگذار بر مقبولیت و تکمیل دفترچه‌ها می‌توان به موارد زیر اشاره کرد:

- فرمت (نوشتاری، فرم‌های اسکن‌شده یا الکترونیکی): به نظر می‌رسد فرم الکترونیکی ابزار به دلیل سهولت استفاده و صرف زمان کمتر در زمان ورود اطلاعات از قابلیت اجرا و مقبولیت بیشتری برخوردار باشد (آلدرسون و اسؤالد[1] ۱۹۹۹).
- نحوه تکمیل (با یا بدون نظارت و راهنمایی): مطالعات نشان می‌دهند اگر فرایند تکمیل و نمره‌دهی اطلاعات در دفترچه‌ها تحت هدایت و راهنمایی یک استاد راهنما باشد مقبولیت بین دانشجویان و در نتیجه دقت اطلاعات بیشتر است.
- زمان تکمیل آن (در طول یا انتهای دوره): ورود روزانه اطلاعات، نسبت به ورود اطلاعات در انتهای دوره از دقت بالاتری برخوردار است. چالشی که در این ارتباط وجود دارد این است که ورود روزانه اطلاعات فعالیتی وقت‌گیر و طاقت‌فرسا است و هدایت فرایند توسط استاد مشاور هزینه‌بر است.
- دستورالعمل‌های شفاف و آموزش و توجیه فراگیران: این مسأله هم بر نحوه تکمیل و هم بر دقت داده‌های جمع‌آوری شده تأثیرگذار است. توجیه فراگیران در مورد فواید ابزار می‌تواند مقاومت آن‌ها در مقابل تکمیل اطلاعات را کاهش دهد. همچنین اگر دانشجویان ارتباط بین نتایج اطلاعات وارد شده و فراهم کردن منابع یادگیری را ببینند به تکمیل آن ترغیب می‌شوند. به عنوان مثال، اگر نتایج دفترچه‌های تکمیل شده حاکی از در دسترس نبودن موارد بیماری یا موقعیت‌های یادگیری خاص است، در صورت فراهم نمودن این موقعیت‌ها (مثلاً تمرین روی مانکن یا در بخش‌های دیگر) توسط مسؤول دوره می‌تواند منجر به افزایش میزان تکمیل اطلاعات توسط دانشجویان شود.

1. Alderson & Oswald

- میزان در دسترس بودن بیماران و موقعیت‌های یادگیری برای همه دانشجویان: یکی از دلایل پایین بودن میزان تکمیل دفترچه‌ها در دسترس نبودن موارد بیماری یا عدم مواجهه فراگیران با موقعیت یادگیری است؛ نه عدم تمایل دانشجویان به ورود آن در لاگ‌بوک.

راخوبار-کریگر و همکاران (۲۰۰۱)

این پژوهشگران با هدف تعیین شرایط مناسب‌تر به منظور دستیابی به اطلاعات دقیق‌تر در لاگ‌بوک، تکمیل اطلاعات در سه موقعیت را بررسی کردند:
- گروه ۱: ورود اطلاعات در انتهای دوره کارورزی بیماری‌های داخلی بدون راهنمایی
- گروه ۲: تکمیل روزانه اطلاعات بدون راهنمایی
- گروه ۳: تکمیل روزانه اطلاعات به همراه راهنمایی هفتگی

در مجموع فراگیران مواجهه وسیعی با موارد بیماری ضروری تعیین شده در بخش بالینی (داخلی) داشتند. بالاترین میزان متوسط بیمار گزارش شده مربوط به گروه ۱ و کمترین مربوط به گروه ۲ بود. تفاوت میانگین‌ها بین دو گروه بدون راهنما (۱و ۲) و یک گروه دارای راهنما (۳)، و دو گروه تکمیل روزانه اطلاعات (۲ و ۳) و یک گروه تکمیل اطلاعات در انتهای دوره (۱) معنی‌دار بود (P=۰/۰۰۰۱).

۵۰ درصد واریانس تکمیل داده‌ها، مربوط به زمان ورود داده‌ها و ۷ درصد مربوط به راهنمایی فراگیران در تکمیل اطلاعات بود. علاوه بر تفاوت بین گروه‌ها، در تجارب دانشجویان درون گروه‌ها نیز تفاوت وجود داشت.

پژوهشگران نتیجه گرفتند گروه اول که در انتهای دوره و بدون راهنمایی دفترچه‌ها را تکمیل کردند، تجربیات خود را بیش‌تر از واقعیت بیان کردند. گروه دوم که در طول دوره و بدون راهنما بودند تجارب خود را کمتر از واقعیت جلوه دادند و گرو سوم بهترین تخمین را از واقعیت داشتند. به این دلیل که گروه اول پس از آزمون، دفترچه‌ها را تکمیل کردند و دانش کسب شده برای امتحان بر نحوه ورود اطلاعات تأثیرگذار بوده است. همچنین این احتمال وجود دارد که گروه دوم به دلیل نداشتن استاد راهنما دفترچه‌های خود را مرتب به روز نکرده و برخی از اطلاعات را وارد نکرده‌اند. بنابراین زمان ورود داده‌ها و هدایت و راهنمایی آن بر دقت ورود داده‌ها تأثیرگذار است و در نظر داشتن این عوامل در طراحی لاگ‌بوک به افزایش مقبولیت ابزار و افزایش دقت ورود اطلاعات کمک می‌کند.

تجربه لاگ‌بوک در ارزیابی کارورزن طب اورژانس دانشگاه علوم پزشکی تهران (فرهمند و سلیمانی ۱۳۸۹)

گروه آموزشی طب اورژانس دانشگاه علوم پزشکی تهران به منظور یکسان‌سازی تجربیات کارورزان در بیمارستان امام خمینی اقدام به طراحی لاگ‌بوک نمود. ابتدا اهداف آموزش عملی توسط شورای آموزش کارورزی گروه طب اورژانس تصویب شد و بر اساس آن دفترچه ثبت مهارت‌های پروسیجرال طراحی شد. پس از اجماع نظر بر سر مهارت‌های عملی، در مرحله بعد میزان مشارکت کارورزان در هر یک از این مهارت‌ها به صورت انجام مستقل، انجام با مشارکت یا مشاهده آن انجام دهد. روایی محتوایی دفترچه‌ها با اجماع نظر استادان بخش به دست آمد.

روند اجرا به این صورت بود که در ابتدای چرخش بالینی، دفترچه‌ها در اختیار کارورزان قرار گرفت و توضیحات لازم در مورد نحوه تکمیل دفترچه‌ها و تأثیر آن در نمره نهایی ارائه شد. هر کارورز، مهارت مورد نظر را تحت نظارت دستیار یا عضو هیأت علمی مسؤول بیمار انجام می‌داد و در صورت موفقیت در آن، اطلاعات مربوط را در دفترچه خود ثبت می‌کرد. دستیار یا عضو هیأت علمی جهت تأیید دفترچه را امضاء می‌کردند.

در وسط دوره دفترچه‌ها جمع‌آوری و توسط مسؤول آموزش کارورزی ارزیابی شد و به کارورزان در مورد نواقص تکمیل دفترچه‌ها بازخورد کتبی داده شد. تکمیل دفترچه ۴ نمره از ۲۰ نمره ارزیابی پایان بخش را به خود اختصاص می‌داد. در انتهای دوره لاگ‌بوکها بر اساس میزان تکمیل ارزیابی نهایی شدند و به چهار گروه خیلی خوب (تکمیل ۹۰ تا ۱۰۰ درصد)، خوب (تکمیل ۷۵ تا ۸۹ درصد)، متوسط (تکمیل ۵۰ تا ۷۴ درصد)، بد (تکمیل ۲۵ تا ۴۹ درصد) و خیلی بد (تکمیل کمتر از ۲۵ درصد) تقسیم شده و از صفر تا چهار نمره‌دهی شدند.

این مطالعه با بررسی دفترچه‌های تکمیل شده (۳۸۸ دفترچه) به مدت یک‌سال نشان دادند که ۷۵/۹۸ درصد از کارورزان حداقل یک بار، ۴۹/۳۶ درصد یک‌بار کمتر از تعداد مورد نظر و ۳۲/۲۵ درصد به طور کامل هر یک از فعالیت‌های مورد نظر بخش را انجام داده، مشارکت داشته یا آن را مشاهده کردند. نتایج مطالعه نشان می‌دهد اکثر کارورزان با فعالیت‌های مورد نظر بخش مواجه شده‌اند. هر چند، این اولین تجربه کارورزان دانشکده در تکمیل لاگ‌بوک بوده است. پژوهشگران موارد زیر را به منظور افزایش میزان تکمیل دفترچه‌ها پیشنهاد کردند:
- آشنایی بیشتر کارورزان با ابزار
- اعمال نظارت بیشتر بر نحوه تکمیل (در بسیاری موراد اطلاعات مربوط به فعالیت به صورت ناقص تکمیل شده بود یا به تأیید راهنمای مربوط نرسیده بود به عنوان تکمیل نشده در نظر گرفته شد)
- اصلاح دفترچه‌ها بر اساس نتایج به دست آمده (اهدافی که در بخش‌های دیگر امکان آموزش و تجربه آن توسط کارورزان وجود دارد از فهرست اهداف خارج شود)

به طور خلاصه، ویژگی‌های یک لاگ‌بوک مناسب ارزان‌قیمت بودن، قابل‌اجرا بودن و مقبولیت نزد فراگیران است و باید امکان جمع‌آوری سریع و صحیح اطلاعات مربوط و بازخورد به موقع را فراهم کند. ویژگی‌های یک لاگ‌بوک ایده‌آل در زیر خلاصه شده است:

- از روایی محتوایی مناسب برخوردار باشد: اندازه‌گیری آن‌چه مسؤولان دوره می‌خواهند اندازه بگیرند.
- از ویژگی بالایی برخوردار باشد: حداقل ۹۰ درصد
- از حساسیت متوسطی برخوردار باشد: حداقل ۷۰ درصد
- به طور وسیعی مورد پذیرش دانشجویان باشد: نرخ پاسخ‌دهی حداقل ۹۰ درصد
- اجازه تحلیل آسان و به موقع اطلاعات را فراهم کند.
- توانایی جمع‌بندی اطلاعات به طبقات مشترک را داشته باشد.
- قابلیت اجرای آسان برای مسؤولان دوره داشته باشد.
- مقرون به صرفه باشد.

با وجود تمام شواهدی که در این فصل ارائه و مورد بررسی قرار گرفت، سؤالات بدون پاسخی بسیاری در مورد لاگ‌بوک و به خصوص دقت آن وجود دارد. طرح این سؤالات در انتهای فصل می‌تواند به عنوان راهنمایی برای پژوهش‌های بیشتر در این زمینه باشد:

- دقت لاگ‌بوک چگونه باید تعریف شود؟ آیا فهرستی از تشخیص‌ها که توسط اعضای هیأت علمی به صورت واقعی در پرونده‌ها ثبت می‌شود دقیق‌ترین است؟
- اگر اطلاعات موجود در لاگ‌بوک غیردقیق است، دلیل آن چیست؟ دانشجویان قادر به تشخیص فهرست موارد به صورت دقیق و کامل نیستند؟ یا آن‌ها توانایی این کار را دارند اما وظایف دیگر آن‌ها را از این عمل باز می‌دارد و تلاش کمی را برای تکمیل دفترچه‌ها به کار می‌برند؟ یا ساختار و فرمت لاگ‌بوک بر تکمیل آن تأثیر می‌گذارد؟
- در صورتی‌که تجربیات یادگیری به جای فهرستی از موارد بیماری یا پروسیجرها به فرمت مهارت‌های پایه مانند جمع‌آوری، گزارش‌دهی و تفسیر اطلاعات به دست آمده از بیماران تغییر کند چه تأثیری بر میزان دقت اطلاعات وارد شده خواهد داشت؟

منابع

1. Alderson TS, Oswald NT. Clinical experience of medical students in primary care: Use of an electronic log in monitoring experience and in guiding education in the Cambridge Community Based Clinical Course. Med Educ. 1999;33(6):429–33.
2. Amin Z, Seng CY, Eng KH. Practical guide to medical student assessment. World Scientific Publishing; 2006.
3. Chatenay M, Maguire T, Skakun E, Chang G, Cook D, Warnock GL. Does volume of clinical experience affect performance of clinical clerks on surgery exit examinations? Am J Surg. 1996;172(4):366–72.
4. Dennick R. Case study 2: use of logbooks. Med Educ. 2000;34(1):S66–S68.
5. Denton GD, DeMott C, Pangaro LN, Hemmer PA. Narrative review: Use of student-generated logbooks in undergraduate medical education. Teach Learn Med. 2006;18:153–4.
6. Denton GD, Durning SJ. Internal medicine core clerkships experience with core problem lists: results from a national survey of clerkship directors in internal medicine. Teach Learn Med. 2009;21(4):281–3.
7. Denton GD, Hoang T, Prince L, Moores L, Durning S. Accuracy of medical student electronic logbook problem list entry. Teach Learn Med. 2007;19(4):347–51.
8. Dong T, Artino AR, Durning SJ, Denton GD. Relationship between clinical experiences and internal medicine clerkship performance. Med educ. 2012;46(7):689–97.
9. Elnicki DM, Zalenski D, Mahoney J. Associations Between Medical Student Log Data and Clerkship Learning Outcomes. Teach Learn Med. 2012;24(4):298–302.
10. Huang GC, Almeida JM, Roberts DH. Reaching the limits of mandated self-reporting: Clinical logbooks do not predict clerkship performance. Med Teach. 2012;34(3):e185–e8.
11. Lampe CJ, Coates WC, Gill AM. Emergency medicine subinternship: Does a standard clinical experience improve performance outcomes? Acad Emerg Med. 2009;15(1):82–5.
12. Links PS, Foley F, Feltham R. The educational value of student encounter logs in a psychiatry clerkship. Med Teach. 1988;10(1):33–40.
13. McCoy CP, Stenerson MB, Halvorsen AJ, Homme JH, McDonald FS. Association of Volume of Patient Encounters with Residents' In-Training Examination Performance. JGIM. 2013;1-7.
14. Neumayer L, McNamara RM, Dayton M, Kim B. Does volume of patients seen in an outpatient setting impact test scores? Am J Surg. 1998;175(6):511–4.
15. Patil NG, Lee P. Interactive logbooks for medical students: Are they useful? Med Educ. 2002;36(7):672–7.
16. Poisson SN, Gelb DJ, Oh MF, Gruppen LD. Experience may not be the best teacher: Patient logs do not correlate with clerkship performance. Neurology. 2009;72(8):699–704.

17. Raghoebar-Krieger HMJ. The objectives-based logbook: a tool for evaluation of medical education [PhD thesis]. University of Groningen, the Netherland; 2001.
18. Raghoebar-Krieger HMJ, Bender W. A comparison of the Dutch Blueprint standards (theory) with the experiences of students in clerkships in Groningen (practice). J Cancer Educ. 1997;12(2):85–8.
19. Raghoebar-Krieger HMJ, Sleijfer D, Bender W, Stewart RE, Popping R. The reliability of logbook data of medical students: an estimation of interobserver agreement, sensitivity and specificity. Med Educ. 2001;35(7):624–31.
20. Rattner SL, Louis DZ, Rabinowitz C, et al. Documenting and comparing medical students' clinical experiences. JAMA. 2001;286(9):1035–40.
21. Sumner W. Student documentation of multiple diagnoses in family practice patients using a hand-held student encounter log. Proceedings American Medical Informatics Association Annual Symposium 2001:687–90.
22. Wimmers PF, Schmidt HG, Splinter TA. Influence of clerkship experiences on clinical competence. Med Educ. 2006;40(5):450–8.
۲۳. فرهمند ش، اصل سلیمانی ح. کتابچه گزارش روزانه (Logbook) کارورزان بخش اورژانس بیمارستان امام خمینی (ره) چگونه تکمیل شده است؟ مجله ایرانی آموزش در علوم پزشکی. ۱۳۸۹؛۱۰(۱):۵۵-۶۳.

فصل ۲۵

کارپوشه

ساختار کارپوشه

همان‌طور که در بخش اول کتاب بیان شد، امروزه شرایط طبابت برای پزشکان متفاوت از گذشته است. آن‌ها با بیمارانی مواجه هستند که دارای دانش و ادعای بیشتری هستند، ملزم به کاربرد شواهد و نتایج تحقیقاتی در کار طبابت روزمره خود هستند و باید با تیم حرفه‌ای پزشکی و جامعه همکاری کنند. پاسخ به این شرایط پیچیده نیازمند کسب توانمندی‌هایی مانند ارتباط مؤثر، سازمان‌دهی، کار گروهی و تعهد حرفه‌ای است. گاهی از این توانمندی‌ها به عنوان «مهارت‌های نرم»[1] در مقابل «مهارت‌های سخت»[2] نام برده می‌شود. با توجه به نتایج بررسی‌هایی که از وضعیت نامطلوب این توانمندی‌ها در میان پزشکان حکایت داشت، امروزه تلاش انجمن‌های حرفه‌ای و دولتمردان بر آموزش و ارزیابی این مهارت‌ها در دانشکده‌های پزشکی به عنوان یک ضرورت متمرکز شده است.

در پاسخ به تغییرات مذکور، در دو دهه اخیر تمرکز برنامه‌های درسی پزشکی از کسب دانش به دستیابی به توانمندی‌ها تغییر کرده است. چالش موجود یافتن ابزاری است که بتواند به صورت تکوینی شکل‌گیری توانمندی‌ها را به صورت یکپارچه، منسجم و طولی حمایت کند و به صورت تراکمی دستیابی به آن را ارزیابی کند. به نظر می‌رسد کارپوشه بتواند این هدف را تأمین کند. اخیراً به کارپوشه به عنوان یک روش ارزیابی عملکرد و رشد حرفه‌ای توجه بسیاری می‌شود. علت اصلی این موضوع این است که بسیاری از اصول زیربنایی کارپوشه اساس آموزش مبتنی بر توانمندی را نیز شکل می‌دهد و کارپوشه می‌تواند اهداف آموزشی و توانمندی‌هایی را که توسط روش‌های سنتی قابل ارزیابی نیستند، بررسی کند.

کارپوشه به عنوان یک ابزار آموزشی ابتدا در رشته‌های هنر و معماری به کار رفت و هدف از آن، جمع‌آوری مستندات و شواهدی مانند عکس‌ها، تصاویر، نقاشی، داستان، انشاء و دیگر کارهای انجام‌شده توسط کارآموزان این رشته‌ها بود. سپس این ابزار در آموزش به صورت کلی و سپس در آموزش علوم پزشکی به طور خاص به عنوان ابزاری برای ارتقا و رشد حرفه‌ای به کار رفت. در واقع، هدف اولیه کارپوشه یادگیری بوده است اما استفاده از آن به عنوان روشی برای ارزیابی به دلیل امکان ارزیابی در طول زمان و در محیط واقعی جذاب بوده است. به بیان دیگر، کارپوشه ابتدا با هدف توانمندسازی حرفه‌ای، یادگیری و ارزیابی تکوینی به کار رفت و سپس جایگاه خود را در ارزیابی رسمی و تراکمی در آموزش پزشکی پیدا کرد. به‌طوری‌که امروزه کارپوشه‌های ساختارمند در آموزش پزشکی رواج پیدا کرده‌اند و تا حد زیادی جایگزین لاگ‌بوک شده‌اند.

از زمان معرفی کارپوشه در آموزش پزشکی در سال ۱۹۹۰ این ابزار جایگاه خود را در ارزیابی و نیز رشد حرفه‌ای باز کرده است.

---

1. Soft skills
2. Hard skills

در ۱۰ تا ۱۵ سال گذشته شاهد معرفی کارپوشه در طیف آموزش پزشکی شامل آموزش پزشکی عمومی، دستیاری، آموزش مداوم و آموزش پزشکان در حال طبابت بوده‌ایم.

تعاریف متعددی از کارپوشه با توجه به هدف و کاربرد آن ارائه شده است. در حالت کلی، کارپوشه به مجموعه‌ای از فعالیت‌های فراگیر گفته می‌شود که نشان‌دهنده کوشش، پیشرفت و موفقیت تحصیلی او در یک زمینه خاص است. در این روش، فراگیر برای نشان دادن توانایی‌ها و پیشرفت خود، نمونه‌ای از کارهایش را که در طول یک مدت معین، مانند یک نیم‌سال یا یک سال تحصیلی انجام داده است، جمع‌آوری می‌کند و برای ارزیابی در اختیار ارزیاب قرار می‌دهد. بنابراین کارپوشه برخلاف آزمون‌های دوره‌ای که عکس‌های جداگانه‌ای از عملکرد فراگیر فراهم می‌کنند، تصویری یک‌پارچه از چگونگی پیشرفت فراگیر فراهم می‌آورد. در واقع، این روش ارزیابی، داستان پیشرفت یادگیری فراگیر در طول یک دوره زمانی و رشد و موفقیت درازمدت وی را بیان می‌کند.

کارپوشه مجموعه‌ای برنامه‌ریزی‌شده و هدفمند از مدارک و شواهد است که نحوه پیشرفت فراگیر و اقداماتی را که برای رسیدن به آن انجام داده است، شامل می‌شود. این مجموعه حاوی منتخبی از فعالیت‌های فراگیر در مدت زمانی معین است. انتخاب کارها با دقت صورت می‌گیرد تا مجموعه آن در کل، هدف مورد نظر از تشکیل کارپوشه را تحقق بخشد. بنابراین محتویات کارپوشه را بهترین کارهای فراگیر یا مثال‌هایی از کارهای انجام‌شده توسط او تشکیل می‌دهند. محتویات مورد نظر می‌توانند گزارش پروژه تحقیقی، مقاله تحقیقی، تجربه مواجهه با بیمار، نمرات آزمون‌ها و ... باشند.

در حوزه آموزش پزشکی تعاریف مختلفی برای کارپوشه ارائه شده است. تعاریف زیر حوزه آموزش پزشکی از پزشکی عمومی تا پزشکان در حال طبابت را پوشش می‌دهد:

- کارپوشه، مجموعه‌ای هدفمند از کارهای دانشجو است که تلاش، پیشرفت و دستاوردهای او را در یک حوزه مشخص، به خود دانشجو و دیگران نشان می‌دهد. دانشجو باید در انتخاب محتوا، تعیین معیارهای انتخاب، معیارهای قضاوت کیفیت و شواهد بازاندیشی مشارکت داشته باشد (رکیز[1] ۱۹۹۵).
- کارپوشه مجموعه‌ای از اوراق و انواع دیگر مستندات است که نمایانگر رخداد یادگیری، همراه با تفاسیر بازاندیشی دانشجو بر یادگیری خود و میزان دستیابی به پیامدهای یادگیری است (دیویس[2] و همکاران ۲۰۰۱).
- کارپوشه مجموعه‌ای از شواهد جمع‌آوری‌شده در طول زمان است که نشان‌دهنده آموزش پزشکان و دستاوردهای طبابت است (ویلکینسون و همکاران ۲۰۰۳)

یکی از وجوه متمایزکننده کارپوشه، مخصوصاً نسبت به لاگ‌بوک، این است که در کارپوشه، تأمل و بازاندیشی بر یادگیری خود نسبت به پیامدهای یادگیری وجود دارد. این ویژگی، کارپوشه را از لاگ‌بوک متمایز می‌کند.

بازاندیشی می‌تواند بر موارد مشکل‌دار (نقاط ضعف فراگیر)، آنچه فراگرفته شده است، آنچه باید آموخته شود یا بر برنامه‌ریزی برای یادگیری آینده متمرکز باشد. این سیستم زمانی خوب کار می‌کند که بین فراگیر و استاد راهنما، تعامل وجود داشته باشد و از مستندات جمع‌آوری‌شده به عنوان تسهیل‌گر در یادگیری بیشتر استفاده شود. نکته قابل توجه این است که کارپوشه نباید به مجموعه‌ای از فعالیت‌های انجام شده و تجارب تبدیل شود، بلکه تأمل نقادانه بر این مستندات و یادگیری که رخ داده است، مهم است. فراگیران متناسب با توانایی بازاندیشی خود می‌توانند بر تجربیات یادگیری خود در سه سطح شناختی توصیفی، تحلیلی و قضاوتی بازاندیشی کنند.

1. Reckase
2. Davis

کاربرد کارپوشه

کارپوشه به منظور تحقق هدف‌های مختلفی به کار می‌رود. در این بخش ابتدا به کاربردهای کارپوشه در آموزش به صورت کلی و سپس در آموزش پزشکی به صورت خاص می‌پردازیم. مهمترین کاربردهای کارپوشه در آموزش عبارتند از:

- **ارزیابی:** کارپوشه می‌تواند هم با هدف ارزیابی تکوینی و هم با هدف ارزیابی تراکمی به کار رود. اگر کارپوشه با هدف تراکمی به کار می‌رود، لازم است محتوای کارپوشه‌های فراگیران مختلف مقایسه‌پذیر باشد و یک روند دقیق و صحیح برای قضاوت و نمره‌دهی به کار برده شود. زمانی که هدف اصلی کارپوشه، القا و ارتقای یادگیری است، نگرانی در مورد مقایسه‌پذیر بودن محتوای کارپوشه‌های مختلف وجود ندارد. در این‌جا تمرکز بر این است که فراگیران یاد بگیرند در مورد کیفیت کار خود تأمل کنند.
- **تقویت و گسترش مهارت‌های تفکر و راهبردهای یادگیری:** کارپوشه می‌تواند امکان گسترش، آموزش و ارزیابی مهارت‌های فکری و راهبردهای یادگیری در فراگیران را فراهم کند. این مهارت‌ها عبارت هستند از:
  - تفکر
  - حل مسأله
  - تصمیم‌گیری
  - راهبردهای شناختی: تحلیل، تفسیر، بسط معنایی، سازمان‌دهی
  - راهبردهای فراشناختی: طرح‌ریزی، هدایت یادگیری، بازاندیشی

  علاوه بر مهارت‌های فکری سطح بالا، کارپوشه امکان ارزیابی ویژگی‌های فردی یادگیرندگان را نیز فراهم می‌آورد. نمونه‌ای از این صفات عبارت هستند از:
  - انعطاف‌پذیری در کارها
  - توانایی سازگار شدن
  - پذیرش انتقاد
  - همکاری با دیگران
  - اشتیاق نسبت به یادگیری
- **فارغ‌التحصیلی:** کارپوشه می‌تواند به منظور فارغ‌التحصیلی به کار رود. در این صورت مشتمل بر فعالیت‌های فراگیران در سال‌های آخر تحصیل است که می‌تواند شامل نمرات آزمون‌ها، فعالیت‌های علمی، فعالیت‌های اجتماعی، فعالیت‌های خارج از برنامه درسی، نوارهای صوتی، فیلم‌های تصویری، فایل‌های کامپیوتری و ... باشد.
- **نشان دادن موفقیت‌های فعلی و پیشرفت تحصیلی فراگیران:** در صورتی‌که هدف کارپوشه نشان دادن موفقیت‌های جاری باشد، نسخه نهایی کارهای انجام‌شده توسط فراگیر در آن قرار می‌گیرد. در حالتی که هدف کارپوشه، هدف نشان دادن رشد و پیشرفت تحصیلی باشد، فرایند انجام یک فعالیت در طول زمان را باید نشان دهد. به عنوان مثال اگر فعالیت انجام شده توسط فراگیر یک پروژه تحقیقاتی باشد و هدف کارپوشه نشان دادن پیشرفت تحصیلی فراگیران باشد، مراحل کار طی شده در کارپوشه به صورت زیر است: مرور متون و تهیه پیش‌نویس اولیه، ارزیابی آن توسط فراگیر، ارزیابی معلم و هم‌کلاسی‌ها و اظهارنظرهای آنان، نسخه اصلاح شده، باز هم ارزیابی مجدد توسط افراد مختلف و در نهایت نسخه نهایی. در حالی‌که اگر هدف تنها نشان دادن موفقیت‌های فعلی باشد، صرفاً نسخه نهایی پروژه تحقیقاتی در کارپوشه قرار می‌گیرد.
- **نمایش فعالیت‌ها:** هدف از این نوع کارپوشه این است که فراگیر یاد بگیرد شاخص‌ترین کارهایی را که به بهترین

نحو نمایان‌گر دانش و توانایی اوست، انتخاب کند. در این انتخاب دیگران نیز می‌توانند به او کمک کنند.

- **ارائه اسناد و مدارک:** در این نوع کارپوشه، برخلاف نوع قبلی، تنها بهترین کارهای فراگیر جمع‌آوری نمی‌شوند. بلکه مجموعه‌ای از فعالیت‌های وی که معرف کمیت و کیفیت یادگیری اوست، گردآوری می‌شود.
- **نشان دادن کارهای به اتمام رسیده و اقدامات در حال انجام:** اگر ارزیابی تراکمی است، در کارپوشه کارهای آماده و به ثمررسیده برای ارائه به مخاطبان گردآوری می‌شود. در صورتی‌که هدف تکوینی است، اقدامات در حال انجام در کارپوشه ذکر می‌شود و فراگیر در مورد عملکرد خود تفکر می‌کند.
- **گزینش و تأیید توانمندی:** این نوع کارپوشه بیشتر در ارزشیابی و توانمندسازی معلمان مورد استفاده قرار گرفته است. به این معنی که معلمان در بدو استخدام یا در طول خدمت باید شواهد و مستنداتی از دانش حرفه‌ای، دانش معلمی، یادگیری مادام‌العمر، دانش نظریه‌ها و راهبردهای آموزشی و ... را به منظور تأیید توانمندی خود نشان دهند. این نوع از کارپوشه امروزه در آمریکا در آموزش پیش از خدمت معلمان و استخدام و ارزشیابی آن‌ها بسیار کاربرد دارد. به دلیل نوع کاربرد، محتوای این نوع کارپوشه باید استاندارد و کاملاً مشخص و تعریف‌شده باشد.
- **تشویق و ارتقاء کارکنان:** این نوع کارپوشه برای تشویق و ارتقای کارکنان مؤسسات به کار می‌رود. در نتیجه از قبل معیارهای مشخصی در اختیار کارکنان قرار می‌گیرد و آن‌ها بر اساس این معیارها محتوای کارپوشه را تهیه می‌کنند. در این کارپوشه شواهد و مدارکی گردآوری می‌شوند که نشان‌دهنده شایستگی فرد در برابر معیارهای ارائه شده است.

در متون آموزش پزشکی، کاربرد کارپوشه در ارزیابی دانشجویان مانند بالا به تفکیک بیان نشده است، بلکه کارپوشه با سه هدف عمده به صورت زیر مطرح شده است:

- **کارپوشه به عنوان روشی برای توانمندسازی فردی و پایش و نظارت بر پیشرفت:** کارپوشه می‌تواند به عنوان ابزاری برای یادگیری فردی بدون حمایت مربی یا تسهیل‌گر به کار رود. در این حالت، فراگیر به طور پیوسته یادداشت‌های بازاندیشی بر تجارب و عملکرد خود را ثبت می‌کند. این نوع کارپوشه شامل محتوای بسیار غنی است که به فراگیر اجازه می‌دهد رشد و پیشرفت خود را در طول زمان پایش کند. چالشی که در این نوع کارپوشه وجود دارد، این است که عدم وجود نظارت از طرف منتور می‌تواند منجر به استفاده کم افراد از کارپوشه شود. مطالعات نشان داده‌اند یکی از علل موفقیت در اجرای کارپوشه ایجاد شبکه‌ای از منتورها و مربیان و ایجاد ارتباط منظم با فراگیران است. هر چند، کارپوشه با هدف پایش و نظارت بر پیشرفت فردی در متون آموزش پزشکی مطرح شده است اما در عمل کارپوشه‌ها با نظارت یک مربی و تسهیل‌گر اجرا شده‌اند. بنابراین در ادامه فصل حاضر، تأکید ما بر کارپوشه با کاربرد یادگیری و ارزیابی خواهد بود.
- **کاربرد تکوینی، کارپوشه به عنوان ابزاری برای یادگیری:** جمع‌آوری مستندات تنها زمانی مفید خواهد بود که این مستندات مورد استفاده قرار گیرد. بنابراین نقش معلم در این مرحله کمک به فراگیر است تا بر آن‌چه برای وی اتفاق افتاده است تفکر و تمرکز کند. آن‌چه که مهم است این است که در انتهای این فرایند، برنامه‌ای برای یادگیری مشخص شود. مشخص کردن گام‌های یادگیری نظارت بر یادگیری را تسهیل می‌کند. بلافاصله پس از معرفی این ابزار در آموزش پزشکی مفهوم کارپوشه یادگیری[1] مطرح شد. در کارپوشه یادگیری نقش منتور بسیار مهم است. وی در تعامل با فراگیر یادگیری را تسهیل می‌کند. در این نوع کارپوشه مهارت بازاندیشی ارتقا می‌یابد به این دلیل که جمع‌آوری نمونه‌ای از کارها و فعالیت‌های انجام شده باعث می‌شود فراگیر به عقب برگردد و ببینید چه کاری را انجام داده است و چه توانمندی هایی را کسب نکرده است. علاوه بر آن، چون این کار در طول زمان انجام می‌شود موجب برنامه‌ریزی و پایش رشد حرفه‌ای می‌شود. بنابراین کارپوشه‌هایی که با هدف یادگیری به کار می‌رود بر بازاندیشی بر فعالیت‌های انجام شده متمرکز است. مراحلی که برای تأمین هدف یادگیری کارپوشه طی می‌شود عبارت هستند از:

1. Portfolio learning

- گام اول: در این مرحله فراگیر یک تجربه را تعیین می‌کند و به توصیف آن می‌پردازد و منتور فعالانه گوش می‌دهد، قضاوت نمی‌کند، تفسیر نمی‌کند و تنها بخش‌هایی از توصیفات فراگیر که مربوط به آن تجربه نیست را برای فراگیر مشخص می‌کند تا کنار گذاشته شوند.
- گام دوم: در مرحله بعدی فراگیر بر تجربه خود تأمل می‌کند تا مشخص شود چه چیزهایی فراگرفته است و مربی توضیح می‌دهد و شفاف‌سازی می‌کند، خلاصه می‌کند، بر آن‌چه فراگیر گفته است بازاندیشی می‌کند اما تفسیر نمی‌کند.
- گام سوم: در این مرحله فراگیر نیازهای یادگیری جدید را تعیین می‌کند و برای دستیابی به این نیازها برنامه‌ریزی می‌کند. مربی فرایند را تسهیل می‌کند و پیشنهادهایی برای دستیابی به نیازها ارائه می‌دهد اما خاطر نشان می‌کند که گزینه‌های دیگری نیز می‌توانند وجود داشته باشند.

سؤالات پیشنهادی برای طرح در کارپوشه یادگیری توسط مربی

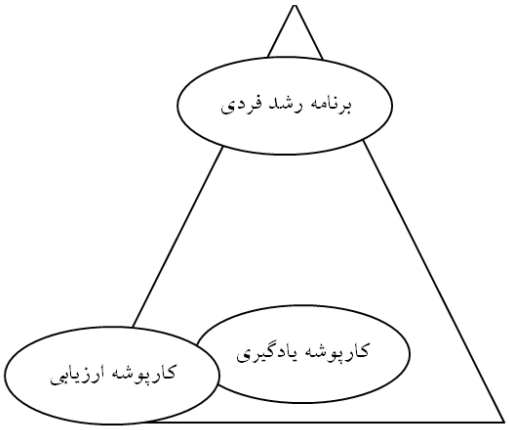
- چه چیزی رخ داده است؟
- چه چیزی توجه شما را جلب کرده است؟ شما را هیجان زده کرده است؟ شما را نگران کرده است؟
- چه چیزی می‌توان از این تجربه آموخت؟
- چگونه آن را یاد خواهید گرفت؟
- چگونه می‌فهمید که آن را یاد گرفته‌اید؟

- **کارپوشه به عنوان ابزاری برای ارزیابی تراکمی:** کارپوشه با فراهم آوردن امکان بررسی عملکرد در محیط واقعی و در طول زمان، ابزاری بسیار مناسب برای ارزیابی است. مواردی از کارپوشه که با هدف ارزیابی به کار می‌رود، بر جمع‌آوری شواهدی مبتنی بر دستیابی به توانمندی‌ها متمرکز است. هر چند باید خاطر نشان کرد در کارپوشه با هدف ارزیابی، ممکن است محتوای مورد انتظار کاملاً متفاوت از کارپوشه با هدف یادگیری باشد. برخی از مطالعات نشان داده‌اند که کارپوشه می‌تواند با هر دو هدف تراکمی و تکوینی (به عنوان مثال در پرستاری) با موفقیت به کار رود (جاسبر[1] ۱۹۹۵). در حالی‌که در مطالعات دیگر نتایج متفاوت بوده است. به عنوان مثال، پزشکان عمومی از گذاشتن یادداشت‌های بازاندیشی که منبع بسیار مهمی برای یادگیری هستند، به دلیل احتمال تأثیر آنها در نمره پایانی امتناع ورزیدند. یکی از راه‌کارها این است که بخش یادداشت‌های بازاندیشی از فرایند ارزیابی تراکمی خارج شود اما باز هم این احساس که یادداشت‌های آن‌ها می‌تواند نگرش ارزیابان را نسبت به فراگیران تغییر دهد، بر این بخش از کارپوشه تأثیر می‌گذارد.

به طور خلاصه می‌توان گفت کارپوشه ابزاری است که می‌تواند با کاربردهای مختلف مورد استفاده واقع شود. مهم این است که بدانیم کارپوشه‌ای که برای هدف خاص، فراگیران مشخص و استفاده در محیط ویژه‌ای طراحی شده است ممکن است برای کاربرد در محیط‌های دیگر مناسب نباشد. در شکل ۱–۲۵ ارتباط بین کارپوشه‌ها با سه کاربرد مختلفی که در بالا اشاره شد، نشان داده شده است. همان‌طور که پیش‌تر بیان شد در عمل، کارپوشه‌ها برای دستیابی به بیش از یک هدف مورد استفاده قرار می‌گیرند و در این صورت مطابق شکل، به مرکز مثلث شیفت می‌کنند.

در این‌جا یکی از چالش‌های کارپوشه مطرح می‌شود به این صورت که آیا قابل پذیرش است که کارپوشه‌ای تهیه کنیم که هم با هدف ارزیابی و هم با هدف یادگیری به کار رود. به این دلیل که هدف ارزیابی ممکن است کیفیت بازاندیشی را تحت تأثیر قرار دهد و فراگیران را از این کار باز دارد. در این حال فراگیران ممکن است در برابر قرار دادن تجربه‌های کمتر موفق یا ضعیف خود و بازاندیشی بر راهبردهای مقابله با عدم موفقیت از ترس تأثیر در نمره ارزیابی مقاومت کنند. از طرف دیگر در کارپوشه‌هایی که با هدف ارزیابی به کار نمی‌روند دانشجویان از وقت و انرژی که صرف کرده‌اند هیچ پاداشی دریافت نمی‌کنند و در نتیجه آن را جدی نمی‌گیرند. در اکثر موارد در آموزش پزشکی، کارپوشه به صورت تلفیقی از دو هدف یادگیری و ارزیابی به کار می‌رود.

1. Jasper



شکل ۱-۲۵: اهداف و محتوای کارپوشه (ون‌تارت‌ویژک و دریسن ۲۰۰۹)

انواع کارپوشه

با توجه به اهداف وسیعی که این روش ارزیابی به صورت بالقوه توان پوشش آن را دارد، کارپوشه طیف وسیعی از محتوا، ساختار و پیچیدگی را شامل می‌شود. گاهی محتوای کارپوشه بسیار عمومی است و مجموعه وسیعی از اطلاعات و مستندات جمع‌آوری‌شده را تشکیل می‌دهد. در موارد دیگر کارپوشه اختصاصی است و تنها مستندات مربوط به هدف خاصی که برای آن تهیه شده است را شامل می‌شود. محتوا و ساختار کارپوشه می‌تواند ساختارمند بوده و بر اساس معیارهای از پیش‌تعیین شده باشد، یا ممکن است بیشتر بر اساس اهداف و پیشرفت فردی فراگیران باشد. ارزیابی محتوای آن نیز می‌تواند از به صورت کاملاً فردی تا ارزیابی توسط افراد بیرونی متغیر باشد. در مجموع، انواع کارپوشه‌هایی که در آموزش پزشکی استفاده شده است را از نظر محتوا می‌توان در دو طبقه اصلی «کارپوشه یادگیری» و «کارپوشه ساختارمند» قرار داد. در زیر این دو نوع کارپوشه با جزئیات بیشتری بیان می‌شوند. سپس در نهایت «کارپوشه الکترونیکی» به دلیل رواج استفاده از آن در ارزیابی دانشجویان علوم پزشکی در سال‌های اخیر مورد بحث قرار می‌گیرد.

کارپوشه یادگیری

اگر چه به دلیل ماهیت فردی این نوع از کارپوشه‌ها، نمی‌توان اجزای دقیقی را به عنوان محتوای آن ارائه داد اما در یک تقسیم‌بندی کلی شامل موارد زیر است:

- **تجربه:** چه چیزی اتفاق افتاده است، چه چیزی انجام شده است، چه چیزی دیده شده است، چه چیزی نوشته شده است، چه چیزی ساخته شده است و ...
- **یادگیری:** بازگو کردن هر آن‌چه فراگرفته شده است و برای انجام یا تغییر فعالیت‌های آینده اهمیت دارد
- **شواهد:** مستند کردن کاربرد دانش در موقعیت مناسب
- **نیازهای یادگیری:** تعیین نقاطی که باید در قدم بعدی به آن پرداخت
- **فرصت‌های یادگیری:** برنامه عملیاتی آموزشی مسیر رسیدن به نیازهای یادگیری را مشخص می‌کند

به عنوان مثال، یک کارپوشه یادگیری که در محیط بالینی مورد استفاده قرار می‌گیرد، می‌تواند شامل موارد زیر باشد:

- **رخدادهای بحرانی از مواجهه با بیمار:** رخدادهای بحرانی وقایعی در کار روزمره هستند که به دلیل خوب بودن یا

بد بودن در ذهن می‌مانند. طیف وسیعی از تجربیات یک پزشک یا دانشجوی پزشکی می‌تواند به عنوان رخدادهای بحرانی در کارپوشه قرار گیرد. فرایند کار به این صورت است که دانشجو خلاصه‌ای را از آن‌چه اتفاق افتاده است یادداشت می‌کند. این تجربه توسط استاد به عنوان منبعی برای یادگیری مورد استفاده قرار می‌گیرد. بدین ترتیب، نیازهای یادگیری مشخص شده و نحوه دستیابی به آن به منظور ارزیابی پیشرفت دانشجو مستند می‌شود. در نتیجه یادگیری بر اساس تجارب و نیازهای دانشجو است. علاوه بر جنبه‌های شناختی موضوع، بحث می‌تواند در مورد واکنش احساسی دانشجو و نگرش وی به مسأله نیز صورت گیرد. به عنوان مثال، دانشجوی پزشکی که برای اولین بار با بیمار فیبریلاسیون دهلیزی مواجه می‌شود و از استفاده از ضد انعقاد برای بیمار مطمئن نیست، تجربه رخداده را یادداشت می‌کند. سپس در جلسه یادگیری با استاد خود مورد را مطرح می‌کند و مشکلات و نیازهای یادگیری دانشجو مشخص می‌شود. بر اساس این نیاز برای یادگیری برنامه‌ریزی می‌شود و در جلسه یا جلسات بعدی دستیابی به برنامه یادگیری بررسی می‌شود. این به معنی یادگیری بر اساس تجربه واقعی است و در نتیجه به دانشجو در صورت مواجهه با بیمار مشابه در آینده کمک خواهد کرد.

- **یادداشت‌های توصیفی از تجربیات بالینی معمول:** این مورد نیز مانند رخدادهای بحرانی است با این تفاوت که در رخدادهای بحرانی فراگیر احساس نیاز و کمبود می‌کند، در حالی‌که ممکن است مواردی وجود داشته باشد که فراگیر از آن آگاه نبوده است و توصیف تجربیات معمول آنها را مشخص کند. محتوای این بخش می‌تواند با توجه به اهداف و توانمندی‌های دوره یا مواردی که دانشجویان مشکلات بیشتری در آن دارند، تجویزی و از پیش تعیین‌شده باشد. مثالی از مورد آخر می‌تواند مدیریت زخم توسط کارآموزان جراحی باشد که اغلب دانشجویان ضعف‌هایی در آن دارند. بنابراین، توصیف تجربه مربوط به آن می‌تواند مشکلات دانشجو و در نتیجه نیازهای یادگیری او را مشخص کند.
- **ضبط ویدئویی تعامل با بیمار یا موارد مشابه:** این مورد بخش مهمی از کارپوشه را تشکیل می‌دهد زیرا فراگیر می‌تواند پیشرفت خود در زمینه تعامل، مشاوره و ارتباط با بیمار را در طول زمان پیگیری کند. ضبط ویدئویی می‌تواند از جلسات مواجهه با بیمار واقعی یا بیمار استانداردشده یا در محیط شبیه‌سازی باشد.
- **منابع آمادگی و مطالعه برای آزمون:** آزمون‌ها و امتحانات، بخش غیرقابل اجتنابی از آموزش هستند. در این بخش فراگیران تلاش‌های خود را به منظور آمادگی برای آزمون در طول دوره با قرار دادن منابع مطالعه‌شده در کارپوشه و بازاندیشی بر آن مستند می‌کنند.
- **یادداشت‌های بازاندیشی:** یادداشت‌های بازاندیشی باعث می‌شوند فراگیر نسبت به آن‌چه انجام می‌دهد، واکنش نشان دهد، نقاط ضعف خود را دریابد و برای رفع آن برنامه‌ریزی کند. یادداشت‌های بازاندیشی فردی است و محتوا و ساختار آن باید توسط فراگیران تعیین شود. چالشی که در این‌جا مطرح می‌شود، دسترسی دیگران به این یادداشت‌ها است. مطالعات نشان می‌دهند اگر دیگران به این یادداشت‌ها دسترسی داشته باشند، فراگیران موارد کمتر شخصی و متفاوتی را در کارپوشه خود قرار می‌دهند. در واقع، یادداشت‌های بازاندیشی جایی است که لازم است مالکیت فراگیران بر کارپوشه حفظ شود.
- **برنامه‌ریزی برای یادگیری و بازاندیشی بر آن:** توصیف جلسات یادگیری و نیازهای یادگیری که در این جلسات مشخص می‌شوند، بسیار مفید است. اما مهم‌تر از آن بازاندیشی بر نقاط قوت و ضعفی که در این جلسات مشخص می‌شوند و برنامه‌ریزی برای یادگیری بیشتر و رفع این نیازها است. یادداشت‌های بازاندیشی در سه موردی که تاکنون بیان شد، باید کوتاه اما موجز باشند. طولانی بودن یادداشت‌ها باعث افزایش بار کاری می‌شود و بررسی و مرور آن را مشکل خواهد کرد.

دیگر مواردی که می‌توان در کارپوشه یادگیری بالینی قرار داد عبارت است از:

- رزومه

- پروژه‌ها، ارائه‌ها و مقالات
- مقالات مرور شده با استفاده از اصول پزشکی مبتنی بر شواهد
- ممیزی مراقبت از بیمار
- بازخوردهای دریافت شده
- منابع مربوط به مدیریت خود و دیگران
- منابع انتخاب شده از طرف فراگیر با این اعتقاد که نشان دهنده مهارت وی است
- هر منبعی که فراگیر معتقد است منبعی برای یادگیری آینده وی است

نمونه‌ای از محتوای کارپوشه یادگیری در محیط غیربالینی

- نتایج فعالیت‌های فراگیر در طول سال یا نیم‌سال تحصیلی
- فعالیت‌های خارج کلاسی
- اظهارنظرهای دیگر افراد مانند معلمان، هم‌کلاسی‌ها و ... در مورد فراگیر
- مواردی که به طور اختصاصی برای کارپوشه تهیه می‌شود: مانند اهداف، بازاندیشی‌ها و توضیحات. به این صورت که فراگیر اهداف و انتظارات خود را از تهیه کارپوشه می‌نویسد؛ در مورد فعالیت‌ها و رشد و پیشرفت خود تأمل می‌کند و در مورد هر یک از کارها توضیحاتی می‌دهد. به عبارت دیگر کارپوشه باید بخشی به عنوان شرح کارپوشه داشته باشد که نشان می‌دهد هر یک از کارها چیست؛ چه ارزشی دارد و به چه منظوری در کارپوشه قرار گرفته است.
- هر چیزی را که معلم و فراگیر مفید تشخیص دهند، می‌توان در کارپوشه قرار داد مانند پروژه‌ها و کارهای تحقیقاتی، مقالات، یادداشت‌های روزانه، نتایج آزمون‌ها و خودارزیابی‌ها

کارپوشه ساختارمند

امروزه کارپوشه‌هایی که در آموزش پزشکی استفاده می‌شوند، ساختارمندتر از انواع سنتی کارپوشه یادگیری هستند. دلایلی که برای این امر وجود دارد عبارت است:

- استانداردسازی قابل‌دفاع برای رسیدن به سطح قابل‌قبول پایایی در صورت استفاده برای اهداف تراکمی
- نشان دادن پاسخگویی اجتماعی برای رسیدن به سطح قابل قبول توانمندی توسط پزشکان در صورت استفاده برای اهداف تراکمی

کارپوشه ساختارمند یعنی محتوایی که توسط برنامه‌ریزان و فراگیران هر دو، تهیه می‌شود تا ارزیابی مبتنی بر پیامد را پیاده کند. یک کارپوشه ساختارمند در آموزش پزشکی باید شواهدی را نشان دهد که فراگیران سطح مورد نظر از توانمندی و عملکرد را به دست آورده‌اند. یک کارپوشه‌ی مؤثر ساختار مشخص اما انعطاف‌پذیر دارد. به عبارت دیگر، به هر فراگیر فرصت توصیف توانمندی و پیشرفت منحصر به خودش را می‌دهد اما با وجود این، دارای دستورالعمل‌های شفاف است. توجه به این موضوع لازم است که وجود راهنماهای واضح خوب است اما نباید با جزئیات بسیار و تجویزی باشند تا آزادی عمل را از فرد سلب نکنند. در زیر نمونه‌ای از چارچوب‌های ساده برای کارپوشه ساختارمند آمده است:

- صفحه عنوان شامل نام، مقطع و سال تحصیلی فراگیر و استاد ناظر
- صفحه فهرست مطالب شامل فهرست محتوای کارپوشه و شماره صفحات مربوط به هر یک
- فهرستی از اهداف یادگیری یا توانمندی‌ها، که در کارپوشه دستیابی به آن مستند می‌شود
- شواهد و مستندات، که بر اساس حیطه‌های اهداف یادگیری دسته‌بندی می‌شوند
- یک مرور بازاندیشی کوتاه که نشان می‌دهد از مرور قبلیِ کارپوشه تاکنون چه پیشرفتی در یادگیری اتفاق افتاده است و مستندات مربوط به آن را نشان می‌دهد

ویژگی‌های اصلی کارپوشه ساختارمند

- دارای کاربرد تکوینی و تراکمی
- حاوی اطلاعات کمی و کیفی
- فردی اما در عین حال استاندارد
- مبتنی بر داده‌های محیط واقعی

محتوای کارپوشه ساختارمند در آمریکا بر اساس استانداردهای ACGME، در کانادا بر اساس چارچوب توانمندی Can-MEDS و در انگلیس بر اساس توانمندی‌های GMC تعیین می‌شود. پس از مشخص شدن اهداف یادگیری یا چارچوب توانمندی‌های مورد انتظار، فراگیر باید مستندات و شواهد را بر اساس دستیابی به این اهداف و توانمندی‌ها مرتب کند. بخش عمده‌ای از این مستندات نتایج ارزیابی فراگیران در آزمون‌های به عمل آمده بر اساس چارچوب توانمندی مورد استفاده است. به عنوان مثال، فهرست ابزارهای ارزیابی بر اساس معیارهای ACGME که می‌تواند در کارپوشه دستیاران تخصصی و فوق تخصصی قرار گیرد در جدول ۱-۲۵ آمده است.

جدول ۱-۲۵: نمونه‌هایی از ابزارهای ارزیابی کارپوشه بر اساس حوزه‌های توانمندی ACGME در آموزش دستیاری

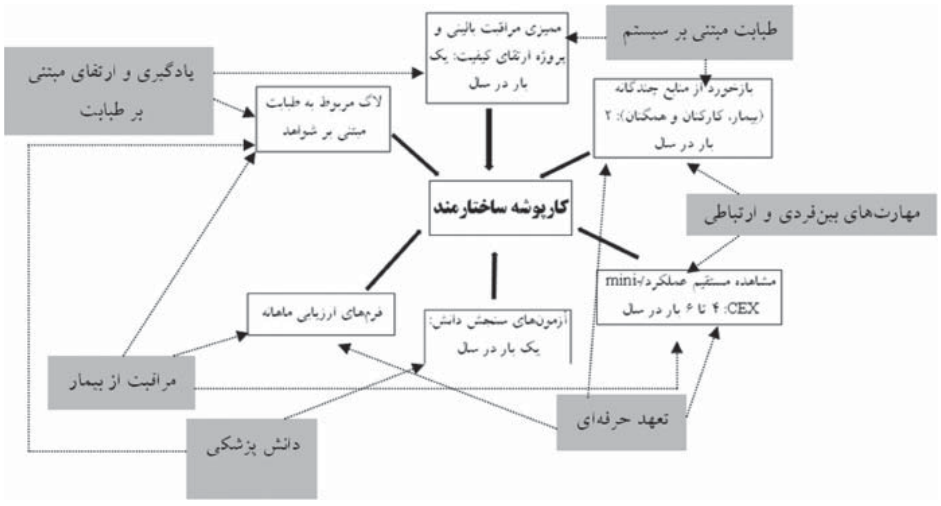
حوزه توانمندی	ابزارهای ارزیابی
مراقبت از بیمار	فرم‌های ارزیابی اختصاصی به هر چرخش بالینی (چک‌لیست و نمره‌دهی گلوبال) مشاهده مستقیم شرح‌حال‌گیری، معاینه فیزیکی و ارتباطات رخدادهای بحرانی[1] لاگ‌بوک (موارد بیماری/ پروسیجر) بهترین نمونه پیگیری بیمار[2] یا تعامل با بیمار به انتخاب فراگیر
دانش پزشکی	امتحانات برگزار شده توسط دانشکده یا بخش موضوعات ارزیابی نقادانه دفترچه‌های ثبت مجلات پزشکی مبتنی بر شواهد[3] دفترچه‌های ثبت سؤالات بالینی (تهیه شده توسط فراگیر) Chart-stimulated recall (CSR)
مهارت‌های بین فردی و ارتباطی	مشاهده مستقیم شرح‌حال‌گیری و ارتباط با بیمار ارزیابی از منابع چندگانه (ارزیابی ۳۶۰ درجه) بازاندیشی فراگیر بر بازخورد از بیمار و همگنان روایت مصاحبه با همکاران و پرسنل پرستاری
تعهد حرفه‌ای	ارزیابی از منابع چندگانه (ارزیابی ۳۶۰ درجه) رخدادهای بحرانی و کارت‌های تحسین[4]
یادگیری و ارتقای مبتنی بر طبابت	برنامه یادگیری فردی پروژه ارتقای کیفیت شامل خود ممیزی طبابت مجموعه سؤالات بالینی خود ارزیابی و بازاندیشی مکالمه و بازخورد با استاد مشاور ثبت شده در وبلاگ
طبابت مبتنی بر سیستم	پروژه بررسی سیستم بهداشتی از دیدگاه بیمار پروژه بررسی خطاهای سیستم، شامل تحلیل رخدادهای بحرانی پروژه طراحی مجدد سیستم‌های خرد[5] ارزیابی مهارت‌های کار گروهی

1. Critical incidents
2. Case log of "best" workups
3. Evidence-based medical journal log
4. praise cards
5. Microsystem

بدون درنظر گرفتن چارچوب توانمندی مورد استفاده، اجزای ارزیابی مندرج در کارپوشه ساختارمند شامل حداقل یک روش ارزیابی از هر یک از پنج دسته روش‌های ارزیابی زیر است:

- **روش‌های پایه[1]:** روش‌های ارزیابی پایه در حداقل موارد شامل استفاده از فرم نمره‌دهی کلی به صورت طولی و ارزیابی ماهانه توسط اعضای هیأت علمی است. هرچند این فرم‌ها مشکلاتی دارند اما اگر به صورت مناسب توسط اعضای هیأت علمی مورد استفاده قرار گیرند، اطلاعات باارزشی را در اختیار می‌گذارند. مهم این است که ارزیابان آموزش ببینند و آیتم‌ها و معیارهای موجود در فرم به خوبی توصیف شود.
- **مشاهده مستقیم مهارت‌های بالینی:** شامل ابزارهایی است که به تفصیل در فصل‌های پیشین به آن پرداخته شد.
- **یادگیری مبتنی بر طبابت و داده‌ها:** این دسته به معنی استفاده فعال فراگیر از داده‌های مربوط به عملکرد خود به صورت فردی است. مثالی از این مورد «ممیزی مستندات پزشکی» مربوط به بیمار است که می‌تواند با خود ارزیابی، بازاندیشی و برنامه ارتقای کیفیت همراه شود.
- **ارزیابی از منابع متعدد:** به این معنا که نظرخواهی از بیمار و دیگر اعضای غیر پزشک تیم پزشکی باید در کارپوشه قرارگیرد.
- **خودارزیابی و بازاندیشی:** ارزیابی‌های این دسته به دو صورت است. مجموعه اول، توصیفاتی داستان‌وار است از آنچه که فراگیران یادگرفته‌اند، دلایل تغییر، برنامه‌ای که برای بهبود عملکرد دارند و نحوه تأثیر فرایند کارپوشه بر یادگیری‌شان. مجموعه دوم، فعالیت‌های خودارزیابی شامل خودممیزی از مستندات پزشکی، پاسخ به سؤالات بالینی و طرح‌های یادگیری خاص است.

شکل ۲-۲۵ نشان می‌دهد که چگونه با ابزارهای متنوع در دسترس می‌توان، حداقل الزامات ارزیابی شش حوزه توانمندی ACGME را فراهم نمود. همان‌طور که ملاحظه می‌شود یک ابزار می‌تواند بیش از یک حوزه توانمندی را مورد ارزیابی قرار دهد. به عنوان مثال فرم ارزیابی کلی ماهانه به بهترین نحو برای ارزیابی مراقبت از بیمار، دانش پزشکی و تعهد حرفه‌ای مناسب است.



شکل ۲-۲۵: تلفیق ابزارهای ارزیابی در کارپوشه ساختارمند (هولمبو و هاوکینز ۲۰۰۸)

1. Foundational methods

ساختارمند کردن و استانداردسازی محتوای کارپوشه، پایایی آن را افزایش می‌دهد اما این مورد باید در برابر آزادی عمل فراگیر برای انتخاب و درج محتوای مورد نظر خود به منظور رشد حرفه‌ای متعادل شود.

تجربه کالج پزشکی(CCLCM) Cleveland Clinic Lerner College of Medicine (دانافر و همکاران ۲۰۰۷)

ساختار کارپوشه در این دانشکده در مقطع پزشکی عمومی بر اساس نه توانمندی زیر تعریف شده است:

- پژوهش
- دانش پزشکی (علوم پایه و بالینی)
- ارتباطات
- مهارت‌های بالینی
- استدلال بالینی
- تعهد حرفه‌ای
- ارتقای فردی
- سیستم‌های مراقبت سلامتی
- بازاندیشی بر طبابت

پس از مشخص شدن چارچوب توانمندی‌ها، گروهی از متخصصان به روش دلفی اصلاح شده برای هر توانمندی استاندارد مناسب را تعریف کردند. این استانداردها برای هر توانمندی در پایان سال اول، دوم و زمان فارغ التحصیلی (پایان سال پنجم) با روند تکاملی تعریف شدند. کوریکولوم دانشکده در فاصله سال‌های ۳ تا ۵ فردی است بنابراین هر دانشجو می‌تواند در هر زمان به استانداردهای انتهایی برسد. در زیر، به عنوان نمونه تعریف توانمندی مهارت‌های ارتباطی و استانداردهای مرتبط با آن آمده است.

تعریف	استانداردهای سال ۱	استانداردهای سال ۲	استانداردهای سال ۳
مهارت‌های ارتباطی زبانی، غیر زبانی و نوشتاری مؤثری را در طیف وسیعی از فعالیت‌های پژوهشی و پزشکی نشان دهد.	• مهارت‌های ارتباطی شفاهی و نوشتاری مؤثری را در پژوهش‌های علوم پایه نشان دهد. • مهارت‌های ارتباطی مؤثری را در محیط بالینی نشان دهد.	• مهارت‌های ارتباطی شفاهی و نوشتاری مؤثری را در پژوهش‌های بالینی نشان دهد. • مهارت‌های ارتباطی مؤثری را با بیمار در محیط بالینی نشان دهد. • مهارت‌های ارتباطی شفاهی و نوشتاری مؤثری را در زمان ارائه مواجهه بیمار نشان دهد. • مهارت‌های ارتباطی مؤثری را در محیط‌های یادگیری مختلف نشان دهد	• از مهارت‌های ارتباطی شفاهی و نوشتاری مؤثر در زمان تعامل با افراد و گروه‌های مختلف در محیط‌های رسمی‌وغیررسمی استفاده کند. • در هر شرایطی طبابت بیمار محور ارائه دهد. • مهارت‌های ارتباطی نوشتاری منسجم، دقیق و از نظر گرامری صحیح نشان دهد. • در زمان تعامل با دیگران، توانایی بازاندیشی بر پاسخ‌ها و اصلاح مناسب آن را نشان دهد.

در دانشکده پزشکی کلیولند سیستمی برای ارزیابی طراحی شد که هدف کلی آن تربیت پزشکانی با توانمندی بازاندیشی با تمایل به یادگیری مادام‌العمر به همراه رویکرد نقادانه به خود ارزیابی و ارتقای فردی بود. به منظور دستیابی به این هدف پیش‌فرض‌ها و موارد زیر در طراحی سیستم در نظر گرفته شد:

- کوریکولوم مبتنی بر توانمندی نمی‌تواند به هدف یادگیری خود راهبر برسد در حالی که ارزیابی دانشجویان تغییر نکند و تنها بر ارزیابی آن‌چه معلمان در کلاس می‌گویند متمرکز باشد.
- در سیستم ارزیابی باید به خود ارزیابی دانشجویان بها داده شود و ارتباط بین دانشجویان و استادان بیشتر شود.
- ارزیابی با هدف ارتقای یادگیری استفاده شود و نیز دستیابی به استانداردها را در زمان فارغ التحصیلی تضمین کند.
- ضرورت دستیابی به طبابت بازاندیشانه لزوم یک روش ارزیابی دانشجو محور را طلب می‌کند.
- کارپوشه با هر دو هدف تکوینی و تراکمی به کار گرفته شود.
- در ارزیابی تراکمی از سیستم نمره‌دهی استفاده نمی‌شود زیرا هدف سیستم ارزیابی، مقایسه توانمندی فراگیران با یکدیگر نخواهد بود.
- در کارپوشه از ابزارهای مختلف ارزیابی استفاده می‌شود:
  - ابزاری شبیه بازخورد از منابع چندگانه یا ارزیابی ۳۶۰ درجه (شامل بازخورد از همگنان، استادان ناظر و خود ارزیابی با چارچوب ثابت اما متناسب با هر مرحله تحصیل دانشجو از نظر دستیابی به توانمندی ها)
  - OSCE
  - آزمون‌های چند گزینه‌ای هفتگی به صورت خودارزیابی
  - سؤالات تشریحی هفتگی با کاربرد دانش
  - آزمون‌های پیشرفت تحصیلی بالینی
  - روش‌های ذکر شده به منظور بازخورد تکوینی استفاده شدند تا نمره‌دهی تراکمی. به عنوان مثال، فراگیران در مورد عملکردشان در OSCE بازخورد دریافت می‌کنند. علاوه بر این، بر اساس تجربیات مختلفی که دانشجو با آن مواجه می‌شود، مستندات مربوط به آن را جمع‌آوری کرده و در نتیجه عملکرد در محیط واقعی در کارپوشه قرار داده می‌شود. از جمله این مستندات گزارش‌های کارهای آزمایشگاهی، نقشه‌های مفهومی، دفترچه‌های مورد بیماری و یادداشت‌های دانشجو هستند. تنها آزمون تراکمی که در کارپوشه قرار داده می‌شود آزمون‌هایی مانند مراحل او ل و دوم USMLE است که برای فارغ‌التحصیلی و دریافت مدرک لازم است.

کارپوشه الکترونیک

می‌توان از فن‌آوری اطلاعات به منظور سازمان دادن، ذخیره کردن و ارائه کارپوشه استفاده کرد. اصطلاحات مورد استفاده، کارپوشه الکترونیکی یا کارپوشه مبتنی بر کامپیوتر است. اخیراً رویکرد به استفاده از کارپوشه الکترونیکی در دانشکده‌های پزشکی افزایش یافته است که دلایل آن در زیر مورد بحث قرار می‌گیرد.

- امکان برقراری ارتباط بین اجزای کارپوشه وجود دارد. در کارپوشه الکترونیکی می‌توان بین شواهد، مرور و بازاندیشی ارتباط[1] ایجاد کرد. این موضوع به ویژه از این نظر مفید است که دانشجو می‌تواند بازاندیشی را با مستنداتی که در جای دیگری ذخیره شده‌اند نشان دهد و یا با ارتباط دادن مستندات و بازاندیشی با یکدیگر مروری شماتیک از پیشرفت و دستاوردهای خود نشان دهد. استفاده از گزینه ارتباط باعث می‌شود مطالب کارپوشه بهتر سازمان‌دهی شود و دسترسی به منابع، مستندات و محتوا برای فراگیر، منتور و ارزیاب راحت‌تر شود.
- کاربری آن آسان است. در مقایسه با کارپوشه‌های دستی، کارپوشه الکترونیکی مشکل حجم زیاد را ندارد، جای کمتری اشغال می‌کند، مشکل ذخیره سازی را حل می‌کند و حمل آن برای ارزیاب و فراگیر آسان‌تر است. در کارپوشه دستی فقط یک نسخه از آن وجود دارد و زمانی که در اختیار یکی از طرفین است طرف دیگر آن را در اختیار ندارد و احتمال مفقود شدن آن وجود دارد. در کارپوشه الکترونیکی انتقال اطلاعات بین معلمان مختلف و مؤسسات مختلف سهل‌تر است، سریعتر می‌توان محتوا را به روز کرد و چندین کاربر در یک لحظه واحد می‌توانند به آن دسترسی داشته باشند و امکان ارائه بازخورد فوری وجود دارد. علاوه بر این امکان بارگذاری مطالب نوشتاری، تصاویر، صدا و ویدئو وجود دارد و می‌توان گزارش‌های چند رسانه‌ای تهیه کرد.
- به دلیل افزایش حس مالکیت و کاربری آسان موجب افزایش انگیزه دانشجویان می‌شود.

در عین حال کارپوشه الکترونیکی دارای معایبی نیز هست که مهمترین آن عبارت هستند از:

- برخی از منتورها دوست ندارند مطالب را از روی صفحه کامپیوتری مرور کنند و بنابراین باید پرینت بگیرند. علاوه بر این گاهی برخی مطالب را نمی توان پرینت گرفت.
- نیاز به تبحر کار با کامپیوتر دارد و برخی از فراگیران و منتورها این مهارت را ندارند.
- به زیرساخت‌های فناوری با ثبات و با کیفیت نیاز دارد که در همه جا در دسترس نیست.

هم اکنون سیستم‌های کارپوشه الکترونیکی با کاربری آسان در اختیار هستند اما ممکن است مؤسسات با توجه به شرایط خود کارپوشه‌های مخصوص خود را طراحی کنند. چندین برنامه تجاری برای کارپوشه‌های الکترونیکی درست شده است که معروف‌ترین آن Aurbach's Grady portfolio[2] است. این برنامه‌ها به معلمان این امکان را می‌دهد تا الگوی مورد نظر خود را با توجه به نیازشان طراحی کنند. هر کدام از موارد فوق مزایا و معایبی دارند. در نتیجه استفاده‌کنندگان باید با توجه به نیازها و امکانات خود در این ارتباط تصمیم‌گیری کنند.

در مقایسه‌ای که دریسون و همکاران[3] (۲۰۰۷) بین کارپوشه الکترونیکی و کاغذی انجام دادند، نتایج نشان داد در نوع الکترونیکی، فراگیران دسترسی بیشتری به محتوا و انواع فرم‌ها داشتند و وقت بیشتری را صرف کارپوشه خود کردند. استادان به صورت متفق‌القول استفاده از کارپوشه الکترونیکی را آسان‌تر از نوع کاغذی دانستند و سریع‌تر می‌توانستند مستندات را از ارتباط‌های موجود بازیابی کنند. دسترسی به محتوا راحت بود و این امکان وجود داشت که از هر مکانی به آن دسترسی پیدا کرد. علاوه بر این، کارپوشه الکترونیکی امکان دسترسی به کار دیگران را نیز فراهم می‌نمود. در نتیجه افراد می‌توانستند با ملاحظه کار همگنان، استاندارد کار خود را بالاتر ببرند. دانشجویان معتقد بودند بازخورد همگنان از معلمان مفیدتر بود. همچنین در استفاده از کارپوشه الکترونیکی ارتباطات با همکلاسی‌ها بیشتر شد. در نتیجه این احتمال وجود

1. Hyperlink
2. http://www.aurbach.com/gp3/index.html
3. Driessen et al.

دارد که کارپوشه الکترونیکی امکان تبادل اطلاعات را به صورتی که در انواع دیگر کارپوشه امکان‌پذیر نیست فراهم کند. مطالعات دیگر با تأیید موارد مذکور، دسترسی آسان و کاهش استفاده از کاغذ را به عنوان فواید و مشکلات عدم دسترسی به کامپیوتر در محیط بالینی را به عنوان چالش‌های کارپوشه در محیط بالینی ذکر کرده‌اند.

دریسون و همکاران (۲۰۰۷)

پژوهشگران در یک مطالعه کارآزمایی تصادفی در دانشجویان سال اول پزشکی دو نوع فرمت کاغذی و الکترونیکی کارپوشه یادگیری را با هم مقایسه کردند. دو ارزیاب به صورت مستقل محتوای هر کارپوشه را از لحاظ کیفیت شواهد و مستندات و بازاندیشی بررسی کردند. همبستگی بین ارزیابان ۰/۷۱ تا ۰/۹۱ بود. نمرات در همه موراد به جز میزان تلاشی که کاربران کارپوشه الکترونیکی صرف کردند یکسان بود. این مطالعه شواهد محکمی در اختیار گذاشت که استفاده‌کنندگان کارپوشه الکترونیکی وقت بیشتری را برای تکمیل کارپوشه خود صرف می‌کنند (P=۰/۰۵). هر دو گروه داوطلبان میزان رضایت مشابهی از کارپوشه داشتند اما کارپوشه الکترونیکی تأثیر بیشتری بر افزایش انگیزه دانشجویان داشت. محتوای هر دو نوع کارپوشه از کیفیت یکسان برخوردار بود و تنها میزان محتوای تصویری کارپوشه الکترونیک بیشتر بود.

به طور کلی عوامل موثر در استفاده از کارپوشه الکترونیکی را می‌توان به صورت زیر طبقه‌بندی کرد:

- **عوامل مربوط به رسانه الکترونیکی:** عوامل مربوط به رسانه الکترونیکی شامل مزایا و محدودیت‌هایی است که به خود رسانه الکترونیکی مطرح است و به تفصیل در بالا به آن اشاره شد.
- **انتقال/دقت اطلاعات در سیستم:** به لحاظ تئوری، رسانه الکترونیکی از این جهت که امکان دسترسی به همه موارد و مستندات مرتبط مربوط به گذشته را فراهم می‌کند بیشتر با الزامات یادگیری مادام‌العمر منطبق است. اما در عمل نتایج مطالعات موردی نشان می‌دهند مشکلاتی در انتقال مستندات بین کارپوشه‌های مختلف و نیز بین مؤسسه‌های مختلف آموزشی وجود دارد. همچنین دغدغه‌هایی در مورد حفظ امنیت اطلاعات این موضوع را تشدید می‌کند.
- **مهارت‌ها/تجارب فناوری اطلاعات کاربران:** تجربه افراد از فناوری اطلاعات، رابطه مثبتی با درک آن‌ها از یادگیری در رسانه‌های الکترونیکی دارد. جر و همکاران[1] (۲۰۰۶) نظر کاربران را در مورد استفاده از کارپوشه الکترونیکی قبل و بعد از مداخله بررسی کردند. قبل از اجرای مداخله ۳۹ درصد کاربران اظهار داشتند اگر مجبور نباشند از کارپوشه الکترونیکی استفاده نخواهند کرد و پس از اجرای آن ۸۷ درصد معتقد بودند که نوع الکترونیکی را ترجیح می‌دهند. نیمی از این افراد اظهار داشتند میزان استفاده خود را با همان میزان ادامه می‌دهند و نیمی دیگر اظهار داشتند میزان استفاده خود را از این پس افزایش می‌دهند. برخی از کاربران در این مورد که ممکن است از بیرون یادگیری آنها کنترل شود محتاط بودند. پژوهشگران نتیجه گرفتند هر چند هر رسانه آموزشی نیاز به آموزش دارد به ویژه برای افرادی که با فناوری آشنا نیستند اما کارپوشه به این موضوع از دیدگاه آموزشی نیاز ندارد.

مزایا و محدودیت‌های کارپوشه

مزایای کارپوشه

- کارپوشه امکان مشارکت فعال فراگیران در یادگیری خود از طریق بازاندیشی را فراهم می‌کند، آن‌ها را به کسب مهارت یادگیری مستقل و اندیشمندانه تشویق و ترغیب می‌کند و چارچوبی برای یادگیری مادام‌العمر و ارتقای مداوم حرفه‌ای فراهم می‌آورد. مهارت بازاندیشی لازمه یادگیری مادام‌العمر است. با توجه به این‌که قسمت عمده حرفه پزشکی در یک محیط طبابت مستقل شکل می‌گیرد و نه طی آموزش ساختارمند؛ و همچنین توانمندی‌های پزشکان به مرور زمان کاهش می‌یابد (هولمبو و هاوکینز ۲۰۰۸)، استفاده از کارپوشه در دوران تحصیل، آن‌ها را برای یادگیری

1. Kjaer et al.

مادام‌العمر و حفظ توانمندی‌ها آماده می‌کند. برخی از برنامه‌های آموزش مداوم نیز از این خاصیت کارپوشه بهره برده‌اند. در دوران تحصیل نیز تعیین محتوای کارپوشه توسط مسؤول دوره و فراگیران با هم موجب یادگیری فعال و ارتقای مهارت خودارزیابی فراگیران می‌شود.

- کارپوشه مبتنی بر تجربیات واقعی فراگیر است که به نوبه خود موجب برقراری ارتباط بین تئوری و عمل می‌شود. کارپوشه یک ارزیابی واقعی از عملکرد فراگیران به عمل می‌آورد، زیرا محتوای کارپوشه، مجموعه جامع از شواهدی است در مورد آنچه فراگیر واقعاً انجام می‌دهد و نه آن‌چه می‌تواند انجام دهد. این موضوع روایی صوری و پیش‌بین کارپوشه را زیاد می‌کند به این دلیل که بین عملکرد فعلی فراگیر و آن‌چه در آینده در محیط طبابت واقعی قادر به انجام آن است، ارتباط برقرار می‌کند. فردی بودن کارپوشه مهمترین عامل واقعی بودن آن است و در صورتی که این عامل با استانداردسازی و ساختارمند کردن بیش از حد کارپوشه از بین برود، روایی کارپوشه تحت تاثیر قرار می گیرد.
- کارپوشه امکان ارزیابی در یک چارچوب، با معیارها و اهداف یادگیری تعیین شده و واضح را فراهم می‌آورد.
- کارپوشه امکان گردآوری شواهد مربوط به یادگیری در محیط‌های مختلف را فراهم می‌آورد.
- کارپوشه امکان ارزیابی تکوینی و تراکمی بر اساس اهداف یادگیری فردی یا اهداف تعیین‌شده توسط دیگران را فراهم می‌کند.
- برخلاف بسیاری از روش‌های ارزیابی دیگر، کارپوشه شامل مجموعه‌ای طولی از مستندات در طی زمان است بنابراین مستندات کارپوشه نمایانگر رشد و پیشرفت حرفه‌ای هستند. با توجه به آن‌چه که تاکنون بیان شد، کارپوشه ارزیابی جامعی از عملکرد فراگیر در طول زمان، در محیط‌های مختلف و در موضوعات مختلف فراهم می‌کند.
- کارپوشه قضاوت کمی و کیفی را با هم ترکیب می‌کند و در نتیجه ارزیابی جامعی از عملکرد فراگیر فراهم می‌آورد.
- کارپوشه یک روش ارزیابی فردی است. ماهیت دانشجو محور بودن کارپوشه باعث می‌شود که دانشجو نقش فعالی در فرایند ارزیابی داشته باشد. فردی بودن کارپوشه به این معنی است که دانشجو در انتخاب شواهد، دخالت تجربیات فردی در کارپوشه و فرایند مرور (که منتورها ویژگی‌های فردی دانشجو را درمی‌یابند، و به دانشجو فرصت می‌دهند تا بر عملکرد خود بازاندیشی کنند و از آن دفاع کنند) مشارکت فعال داشته باشند. فردی بودن کارپوشه امکان استفاده از سبک‌های یادگیری متفاوت بر اساس ترجیحات فراگیر را فراهم می‌کند. در مورد ارزش‌های فردی، احساسات، سبک تعامل و روش‌های مواجهه با تجربیات خاص بازخورد فردی ارائه می‌دهد. یک تصویر فردی از آن چه دانشجو می‌داند و قادر به انجام آن است فراهم می‌کند. تجربیات را با تفاسیر فردی همراه می‌سازد بنابراین یادگیری را مبتنی بر موقعیت (شرایط) می‌کند.
- کارپوشه قابلیت ارزیابی پیامدهای یادگیری از قبیل رشد فردی، یادگیری خودراهبر، توانایی بازاندیشی، خودارزیابی رشد فردی و تعهد حرفه‌ای که به آسانی توسط روش‌های دیگر ارزیابی قابل سنجش نیستند را دارد.
- کارپوشه ارتباط بین دانشجویان و استادان را افزایش می دهد. امکان بحث و گفتگو بین این دو را فراهم می‌کند و به دانشجویان نشان می‌دهد که یادگیری یک فرایند دو طرفه است.
- کارپوشه معلمان را ترغیب به ارزیابی مجدد راهبردهای تدریس می‌کند و انتظارات استادان را در ارتباط با توانایی تفکر و حل مسأله افزایش می‌دهد.

محدودیت‌های کارپوشه

- طراحی و اجرای کارپوشه به نحوی که هر دو کاربرد تکوینی و تراکمی را داشته باشد، در عمل مشکل است.
- فردی بودن کارپوشه در عین حال که یک مزیت محسوب می‌شود، این مشکل را به وجود می‌آورد که متناسب کردن آن با نیازهای یک سیستم یا روش ارزیابی که خواستار استاندارد کردن ارزیابی هستند، مشکل است.

کارپوشه‌های ساختارمند درجات قابل‌دفاعی از استانداردسازی را دارا هستند. با استفاده از این رویکرد برخی از معیارها و وظایف مثل سیاست‌های ردی و قبولی مشخص می‌شود و آموزش ارزیابان و داوطلبان راحت‌تر انجام می‌شود. با وجود این، ممکن است بر فرایند و محتوای کارپوشه کنترل بیش از حد ایجاد شود. به منظور غلبه بر این مشکل احتمالاً می توان در کارپوشه ساختارمند نوع کاری را که قرار است در کارپوشه قرار گیرد، تعیین نمود اما لازم است ساختار، عمق، حجم و نمایش اطلاعات با توجه به نیازهای فردی انعطاف‌پذیر باشد.

گام‌های طراحی و اجرای یک کارپوشه مطلوب

رعایت یک سری از موارد که در جدول ۲-۲۵ خلاصه شده‌اند، به طراحی یک کارپوشه مطلوب کمک می‌کند.

جدول ۲-۲۵ : خلاصه مراحل طراحی و اجرای کارپوشه

ردیف	عنوان	توضیح
۱	تشکیل کمیته مرکزی	ضروری است طراحی، اجرا و نظارت بر اجرای کارپوشه تحت سرپرستی یک کمیته مرکزی باشد تا یکپارچگی سیستم حفظ شود.
۲	تعیین هدف کارپوشه	کارپوشه با اهداف مختلفی به کار می‌رود. هدف کارپوشه بر بخش‌های دیگر آن از جمله محتوا و ساختار آن تأثیرگذار است. بنابراین، ضروری است هدف کارپوشه از همان ابتدا مشخص شود.
۳	برقراری ارتباط بین پیامدهای دوره و محتوای کارپوشه	تدوین بلوپرینت، بین اهداف یادگیری و محتوای کارپوشه ارتباط برقرار می‌کند و از درج محتوای غیرضروری و حجیم شدن کارپوشه پیشگیری می‌کند. علاوه بر این، تضمین می‌کند دانشجو محتوای لازم را در کارپوشه قرار دهد.
۴	تعیین محتوا و ساختار کارپوشه	لازم است در مورد میزان استانداردسازی و ساختارمند بودن کارپوشه، فرمت آن (کاغذی یا الکترونیکی) و ظاهر آن تصمیم‌گیری شود
۵	ایجاد فرهنگ حمایت از فراگیران به صورت نظام‌مند	فراگیران در تهیه کارپوشه، مشارکت در آن و بازاندیشی بر یادگیری خود نیاز به حمایت دارند. این موضوع به ویژه برای فراگیرانی که برای اولین بار درگیر این فعالیت می‌شوند، حائز اهمیت بیشتری است. ماهیت حمایت متناسب با نیازهای فراگیران و هدف کارپوشه می‌تواند از نوع هدایت و مشاوره یا منتورینگ باشد. ضروری است راهنماها و دستورالعمل‌های شفاف در خصوص فرایند حمایت دانشجو شامل فرد یا افراد مسؤول، نحوه ارتباط فراگیر و حامی (یا منتور)، ترتیب و توالی جلسات مرور و زمان اختصاص یافته به آن و مواردی از این دست مشخص شوند.
۶	حمایت فراگیران در تکمیل کارپوشه	لازم است فراگیران در مورد کارپوشه، چگونگی تهیه آن و اهداف آن در ابتدا و به صورت مستمر حمایت شوند. همچنین ضروری است منتورها درخصوص انتخاب و ورود فعالیت‌ها به آن‌ها کمک کنند.
۷	آموزش دست‌اندرکاران	اطلاع‌رسانی و آموزش دست‌اندرکاران (فراگیران، منتورها و ارزیابان) می‌تواند از طریق دستورالعمل‌های کتبی، جلسات توجیهی حضوری و برگزاری کارگاه باشد. در کارگاه آموزشی دغدغه‌ها رفع می‌شود، دستورالعمل‌ها تدوین می‌شود و شرکت‌کنندگان مراحل مربوط به مرور کارپوشه و برنامه‌ریزی عملیاتی را تمرین می‌کنند. یکی از مهم‌ترین بخش‌های کارگاه، حضور افرادی است که قبلا از کارپوشه استفاده کرده‌اند و تجربیات خود را در اختیار دیگران قرار می‌دهند. یکی از بهترین منابعی که می‌توان در کارگاه آموزشی مورد استفاده قرار داد، نمونه‌های کارپوشه واقعی است.
۸	تعیین سطح دسترسی به کارپوشه	با توجه به ماهیت فردی و خصوصی برخی از قسمت‌های محتوای کارپوشه، به ویژه یادداشت‌های بازاندیشی و توصیه‌ها، باید تصمیم گرفته شود که چه کسانی و چگونه این بخش‌ها را ببینند.

ادامه جدول ۲-۲۵ : خلاصه مراحل طراحی و اجرای کارپوشه

ردیف	عنوان	توضیح
۹	طراحی رویکرد نظام‌مند برای ارتقای بازاندیشی فراگیران بر تجربیات خود	شواهد نشان می‌دهند بازاندیشی بر تجربیات در برخی از موارد از کیفیت لازم برخوردار نیست. در نتیجه باید در دستورالعمل‌های تدوین‌شده، در مورد نحوه بازاندیشی با جزئیات پرداخته شود. طراحی یک رویکرد نظام‌مند به ارتقای توانایی بازاندیشی دانشجویان در طول زمان و تشویق دستیابی به سطوح بالاتر بازاندیشی با بالاتر رفتن سطح دانشجو مفید به نظر می‌رسد.
۱۰	تعیین ملاک‌های مرور کارپوشه و فرایند مرور	منتورها باید پیشرفت دانشجو را در دستیابی به اهداف برنامه درسی در طول زمان بررسی و مرور کنند. ملاک‌های مرور و ارزیابی باید در اختیار فراگیران قرار گیرد. در جلسات مرور، منتور ممکن است پس از مرور کارپوشه نظرات خود را به صورت نقاط ضعف و قوت در کنار هر یک از کارها به صورت کتبی بنویسد. اما توصیه می‌شود جلسات به صورت بحث و تبادل نظر با دانشجو باشد و در انتها دانشجو برنامه یادگیری خود را برای آینده تنظیم کند.
۱۱	ارزیابی کارپوشه	روش‌ها و فنون نمره‌دهی کارپوشه شامل چک‌لیست، مقیاس درجه‌بندی، روش کلی و روش تحلیلی است. در روش تحلیلی مجموعه‌ای از معیارهای ارزیابی به کار می‌روند که برای هر یک از آن‌ها نمره جداگانه‌ای در نظر گرفته می‌شود. در روش کلی معیارهای مختلف با هم درنظر گرفته می‌شود و یک نمره واحد گزارش می‌شود. روش تحلیلی برای ارزیابی تکوینی کارپوشه و روش ترکیبی برای ارزیابی تراکمی مناسب است. در ارزیابی هر یک از آیتم‌های کارپوشه، روش تحلیلی و ارزیابی کل آن، روش کلی مناسب است. لازم است برای هر یک از مقیاس‌ها در مقیاس نمره‌دهی، توصیف دقیق و اختصاصی از رفتار مناسب آن مقیاس به عمل آید تا از ذهنی بودن نمره‌دهی کاسته شود. علاوه بر این، به منظور تأمین ثبات نمره‌دهی بهتر است از نظر بیش از یک ارزیاب استفاده گردد.
۱۲	ارزشیابی کارپوشه	به نظر می‌رسد ترکیبی از روش‌های کمی و کیفی به منظور ارزشیابی کیفیت کارپوشه مفید باشند و تنها رویکرد سنتی بررسی ویژگی‌های روانسنجی روایی و پایایی کافی نباشد.

تشکیل کمیته مرکزی

لازم است طراحی، اجرا و نظارت بر اجرای کارپوشه به صورت نظام‌مند تحت سرپرستی یک کمیته مرکزی باشد. این کمیته همچنین وظیفه تربیت متخصصان کارپوشه را به عهده دارد که آن‌ها نیز افراد دیگر در سیستم را آموزش می‌دهند تا یکپارچگی سیستم حفظ شود. همچنین ضروری است از نظرات، تجربه و تخصص افراد صاحب‌نظر در طراحی و اجرای این ابزار پیچیده استفاده شود.

تعیین هدف کارپوشه

همان‌طور که اشاره شد کارپوشه با اهداف رشد و ارتقای فردی و حرفه‌ای، یادگیری (تکوینی) و ارزیابی (تراکمی) به کار می‌رود. هدف کارپوشه بر بخش‌های دیگر آن از جمله محتوا و ساختار آن تأثیرگذار است. به عنوان مثال کارپوشه‌ای که با هدف ارزیابی تراکمی به کار می‌رود، بسیار ساختارمند و دارای چارچوبی از پیش‌تعیین‌شده برای ورود محتوا است. در حالی که در کارپوشه یادگیری، نقش فراگیر در انتخاب محتوا و ساختار کارپوشه بسیار برجسته است. در حال حاضر در آموزش پزشکی معمولاً کارپوشه با ترکیبی از اهداف فوق به کار می‌رود. در هر صورت به منظور روشن شدن مسیر باید از همان ابتدا هدف کارپوشه مشخص شود.

برقراری ارتباط بین پیامدهای دوره و محتوای کارپوشه

مزیت کارپوشه آن است که می‌تواند میزان دستیابی به انواع اهداف و پیامدهای یادگیری را در ساختار خود جای دهد. در این خصوص به تفصیل در بخش کارپوشه ساختارمند اشاره شد. تدوین بلوپرینت، بین اهداف یادگیری و محتوای کارپوشه ارتباط برقرار می‌کند و از درج محتوای غیرضروری و حجیم شدن کارپوشه پیشگیری می‌کند. علاوه بر این، تضمین می‌کند دانشجو محتوای لازم را در کارپوشه قرار داده است و از نشان دادن عملکرد کمتر از واقعیت جلوگیری می‌کند.

تعیین محتوا و ساختار کارپوشه

تصمیم‌گیری در مورد محتوا و ساختار کارپوشه ارتباط نزدیکی با هدف کارپوشه دارد. در مورد محتوا و ساختار کارپوشه به طور مفصل در قسمت‌های قبلی اشاره شد. لازم است در مورد میزان استانداردسازی و ساختارمند بودن کارپوشه، فرمت آن (کاغذی یا الکترونیکی) و ظاهر، بخش‌های مختلف و جزئیات آن تصمیم‌گیری شود.

ایجاد فرهنگ حمایت از فراگیران به صورت نظام‌مند

پیاده‌سازی کارپوشه به دلیل مسائلی مانند در پیش گرفتن یک رویکرد طولی، ارائه بازخورد تکوینی، واگذاری مسؤولیت به دانشجو به منظور مستندسازی دستیابی به توانمندی‌ها و ... چالش‌هایی را در مقابل رویکرد سنتی یادگیری و ارزیابی معلم‌محور به وجود می‌آورد. در نتیجه لازم است یک فرهنگ حمایتی در دانشکده به ویژه در ابتدای استقرار سیستم ایجاد شود.

فراگیران در تهیه کارپوشه، مشارکت در آن و بازاندیشی بر یادگیری خود نیاز به حمایت دارند. این موضوع به ویژه برای فراگیرانی که برای اولین بار درگیر این فعالیت می‌شوند، حائز اهمیت بیشتری است. حمایت فراگیران برای کارپوشه‌هایی که مورد ارزیابی بیرونی قرار می‌گیرد نیز اهمیت بسیاری دارد.

ماهیت حمایت متناسب با نیازهای فراگیران و هدف کارپوشه متفاوت است. در مقطع پزشکی عمومی احتمالاً یکی از استادان آموزشی حمایت فراگیر را به عهده می‌گیرد و نقش وی هدایت و مشاوره است. در مقطع دستیاری با توجه به این‌که دستیاران به نوعی پزشکان مبتدی در حال طبابت هستند، حمایت از نوع منتورینگ کاربرد بیشتری دارد. در مواردی که کارپوشه با هدف ارزیابی مورد استفاده قرار می‌گیرد، توصیه می‌شود فرد حامی متفاوت از فرد ارزیاب باشد.

منتورینگ چیست؟

- یک ارتباط قوی بین یک فرد مبتدی در یک حرفه و فرد صاحب‌نظر در آن حوزه به منظور جلب مشارکت بالا
- راهنمایی یک فرد مبتدی به منظور ارتقای حرفه‌ای و پیمودن مسیر تعالی حرفه‌ای با یکدیگر

حمایت منتور می‌تواند به یکی یا چند مورد از روش‌های زیر انجام شود:

- منتورینگ برنامه‌ریزی شده: در این نوع ارتباط، ملاقات‌ها بر اساس برنامه زمانی از پیش‌تعیین شده انجام می‌شود. در نتیجه دو طرف نسبت به فرایند متعهد هستند و پیامدها مورد بحث قرار می‌گیرند.
- منتورینگ در راهرو[1]: ملاقات‌ها به صورت غیررسمی با توجه به شرایط پیش‌آمده شکل می‌گیرد.
- منتورینگ متقابل[2]: گروه‌های دونفره یا با تعداد بیشتر از فراگیران برای حمایت یکدیگر یا با هدف مطالعه شکل می‌گیرد. • منتورینگ تلفنی: به ویژه زمانی که منتور و متنی از یکدیگر دور هستند کاربرد دارد.

به عنوان مثال، انجمن سلطنتی پزشکان عمومی انگلیس مدلی از منتورینگ گروهی را پیاده می‌کند که در آن دو یا سه پزشک عمومی در یک گروه به عنوان فراگیران کارپوشه، نیازهای یکدیگر را برآورده می‌کنند و نوعی از حمایت، تسهیل و تشویق را برای همکاران خود فراهم می‌آورند. نکته قابل توجه این است که حمایت منتور نباید قضاوتی باشد. منتور عقاید و مسائل را بر اساس تجربیات فردی خود بیان می‌کند، اما مسؤولیت نهایی اعمال فراگیر در تدوین و تکمیل کارپوشه را به عهده ندارد. منتورینگ با ارزیابی و قضاوت متفاوت است (چالیس ۱۹۹۹).

1. Corridor
2. Mutual

تدوین راهنماها و دستورالعمل‌های شفاف در این زمینه موجب ثبات در پایش پیشرفت دانشجو و اطمینان از صحت فرایند می‌شود. در مجموع موارد زیر باید مشخص شوند:

- چه کسی مسؤول حمایت دانشجو در فرایند ارزیابی با کارپوشه است؟
- هر فرد حمایت‌کننده (منتور یا استاد مشاور) مسؤولیت حمایت از چند فراگیر را به عهده دارد؟
- نحوه ارتباط فراگیر و منتور در حالت رسمی (جلسات یادگیری)، غیررسمی (در موقعیت‌های امکان‌پذیر) یا ترکیبی از این‌دو چگونه است؟
- ترتیب و توالی جلسات مرور و زمان اختصاص یافته به چه میزان است؟
- آخرین مهلت تحویل کارپوشه چه زمانی است و در صورت دیرکرد چه تبعاتی به دنبال خواهد داشت؟ (زمان تحویل کارپوشه باید از ابتدا مشخص باشد و فراگیران بدانند چه قدر برای تکمیل آن وقت دارند).
- فرایند مرور کارپوشه متناسب با هدف آن (تکوینی یا تراکمی) چگونه خواهد بود؟
- زمان مورد نظر برای بررسی و مرور کارپوشه برای اعضای هیأت علمی چقدر است؟ (معمولاً زمانی در حد ۲۰ دقیقه در هفته کفایت می‌کند).

تجربه کالج پزشکی CCLCM

در کارپوشه مورد استفاده در این دانشکده سیستمی برای مشاوره راه‌اندازی شده است. به این صورت که در همان ابتدا (در بدو ورود به دوره پزشکی عمومی) هر دانشجوی پزشکی به یک استاد مشاور (مدرسان بالینی) معرفی می‌شود. هر استاد مشاور مسؤولیت حمایت از ۱۰ دانشجو را به عهده دارد. برای ایفای این مسؤولیت زمان خاصی برای مشاوره جدا از زمان آموزش در بالین در نظر گرفته شده است. در یک سیستم کارپوشه الکترونیکی استادان مشاور داده‌های مربوط به ارزیابی‌های دانشجو، کارهای تولید شده توسط دانشجو و تکالیف درسی را مرور و نظارت می‌کنند. در این زمینه دستورالعمل‌های مشخصی تدوین شده است، برای استادان مشاور کارگاه‌های آموزشی برگزار شده است تا مرور کارپوشه دارای ثبات باشد. علاوه بر آن ایشان در جلسات هفتگی شرکت می‌کنند تا فهم مشترکی از نقش استاد مشاور و روش‌های ایفای این نقش داشته باشند.

توجیه و حمایت فراگیران در تکمیل کارپوشه

تکمیل‌کننده اصلی کارپوشه فراگیران هستند و دیگران از جمله منتورها نقش حمایتی را ایفا می‌کنند. در ابتدا باید در مورد کارپوشه، چگونگی تهیه آن و اهداف آن به فراگیران توضیح داده شود سپس در طول سال تحصیلی یا دوره، مطالب با آن‌ها مرور و در مورد فعالیت‌های آینده برنامه‌ریزی شود. انتخاب محتوا با فراگیران است اما لازم است منتور یا استاد مشاور در انتخاب و ورود فعالیت‌ها به فراگیران کمک کند. باید به فراگیران توضیح داده شود که بهتر است نمونه‌هایی از کارهایشان را که معرف دستیابی آنان به توانمندی‌های موردنظر است، انتخاب کنند و شواهدی از یادگیری خودراهبر و پیشرفت را در کارپوشه بگنجانند. ضروری است تاریخ انجام هر بخش از کار و توضیح مختصری در مورد علت انتخاب آن توسط فراگیران وارد شود.

همان‌طور که قبلاً توضیح داده شد، لازم است طیف وسیعی از اطلاعات، از منابع مختلف و در طول زمان جمع‌آوری و در کارپوشه قرار داده شود. به دلیل زمان‌بر بودن فرایند تکمیل باید مطمئن شویم که کارپوشه به راحتی و به طور مؤثری تکمیل می‌شود. یکی از راه‌ها می‌تواند درج شواهد و مستندات کمتر اما گویاتر باشد. به عنوان مثال، برای پیشگیری از افزایش کاذب حجم کارپوشه، می‌توان مواردی مانند محدودیت واژگان را لحاظ کرد. علاوه بر این، به منظور عدم تعویق تکمیل کارپوشه به انتهای دوره بهتر است به طور مرتب محتوای آن کامل شود و منتورها به ارائه بازخورد به محتوای به روز شده بپردازند.

آموزش دست‌اندرکاران

یکی از مهمترین عوامل موفقیت در اجرای کارپوشه اطلاع‌رسانی و آموزش دست‌اندرکاران در خصوص نحوه اجرای کارپوشه است. این آموزش می‌تواند از طریق دستورالعمل‌های کتبی، جلسات توجیهی حضوری و برگزاری کارگاه باشد. فراگیران، منتورها و ارزیابان باید آموزش ببینند. توجه به دغدغه‌ها و بیم و هراس‌های این افراد و تلاش برای رفع آنها، احتمال استفاده از کارپوشه یادگیری را افزایش می‌دهد. حضور فراگیران و معلمان با هم در یک جلسه و بحث، یکی از اقدامات اساسی در رفع موانع است. کارپوشه یادگیری در صورتی اجرا خواهد شد که فراگیران از آن نفعی ببرند و استفاده از آن برایشان خوشایند باشد. در کارگاه آموزشی دغدغه‌ها رفع می‌شود، دستورالعمل‌ها تدوین می‌شوند، یادداشت‌ها اصلاح می‌شوند و شرکت‌کنندگان مراحل مربوط به مرور کارپوشه و برنامه‌ریزی عملیاتی را تمرین می‌کنند. یکی از مهم‌ترین بخش‌های کارگاه، حضور افرادی است که قبلاً از کارپوشه استفاده کرده‌اند و تجربیات خود را در اختیار دیگران قرار می‌دهند.

یکی از بهترین منابعی که می‌توان در کارگاه آموزشی مورد استفاده قرار داد، نمونه‌های کارپوشه واقعی است. می‌توان پس از کسب اجازه از دانشجویان و بدون فاش شدن هویت آنان نمونه‌هایی از کارپوشه‌های آن‌ها را برای آموزش در کارگاه استفاده کرد. یکی از بخش‌های کارگاه‌ها باید آموزش نحوه ارائه بازخورد باشد. ضروری است هم دانشجویان و هم منتورها با فرایند ارائه بازخورد توصیفی غیر قضاوتی و بر اساس عملکرد مشاهده شده آشنا شوند.

تعیین سطح دسترسی به کارپوشه

با توجه به ماهیت فردی و خصوصی برخی از قسمت‌های محتوای کارپوشه، به ویژه یادداشت‌های بازاندیشی و توصیه‌ها، باید تصمیم گرفته شود که چه کسانی و چگونه این بخش‌ها را ببینند. مؤثرترین یادداشت‌های بازاندیشی آن‌هایی هستند که هم فراگیر و هم منتورها هر دو به طور منظم آن را ملاحظه می‌کنند. به غیر از این، مستندات نباید بدون کسب اجازه از فراگیر توسط دیگران دیده شود.

طراحی رویکرد نظام‌مند برای ارتقای بازاندیشی فراگیران بر تجربیات خود

یکی از بخش‌های ضروری هر کارپوشه تأمل و بازاندیشی بر تجربیات است. بازاندیشی باید به چهار سؤال زیر پاسخ دهد:

- چه چیزی یاد گرفته‌ام؟
- چه چیزی باید یاد بگیرم؟
- چه منابعی برای یادگیری بیشتر لازم دارم؟
- چه چیز بیشتری یاد گرفته‌ام؟

شواهد نشان می‌دهند بازاندیشی بر تجربیات در برخی از موارد از کیفیت لازم برخوردار نیست. در نتیجه باید در دستورالعمل‌های تدوین‌شده، در مورد نحوه بازاندیشی با جزئیات پرداخته شود و مربیان در مورد برخی از بازاندیشی‌ها به ویژه وقتی دانشجویان برای اولین بار درگیر این کار هستند، بازخورد ارائه دهند. طراحی یک رویکرد نظام‌مند به ارتقای توانایی بازاندیشی دانشجویان در طول زمان و تشویق دستیابی به سطوح بالاتر بازاندیشی با بالاتر رفتن سطح دانشجو مفید به نظر می‌رسد.

تعیین ملاک‌های مرور کارپوشه و فرایند مرور

هدف کارپوشه هر چه باشد، ملاک‌های مرور و ارزیابی باید مشخص و واضح باشند. این ملاک‌ها بر اساس هدف کارپوشه، اهداف آموزشی و توانمندی‌ها تعیین می‌شوند. ملاک‌های مرور و ارزیابی باید در اختیار فراگیران قرار گیرد.

ضروری است طبق دستورالعمل‌هایی که از قبل تعیین شده است، جلساتی برای مرور کارپوشه تنظیم شود. این جلسات می‌تواند کاملاً برنامه‌ریزی شده یا فی‌البداهه در جلسات درسی یا موقعیت‌های دیگر باشد. منتور ممکن است پس از مرور کارپوشه نظرات خود را به صورت نقاط ضعف و قوت در کنار هر یک از کارها به صورت کتبی بنویسد. اما توصیه می‌شود جلسات به صورت بحث و تبادل نظر با دانشجو باشد و در انتها دانشجو برنامه یادگیری خود را برای آینده تنظیم کند. ممکن است جلسات مرور کارپوشه با ارائه شفاهی دانشجو همراه باشد. در مجموع، منتورها باید پیشرفت دانشجو را در دستیابی به اهداف برنامه درسی در طول زمان بررسی و مرور کنند. در زیر به نمونه‌ای از معیارها و فرایند مرور کارپوشه اشاره شده است:

فرایند مرور کارپوشه در کالج پزشکی CCLCM (ارزیابی تکوینی)

به طور مرتب (یک تا سه بار در سال)، دانشجویان جلسه‌ای با استاد مشاور خود جهت مرور کارپوشه تکوینی دارند. برنامه زمانی این جلسات از قبل بر اساس برنامه درسی مشخص است. علاوه بر این، دستورالعمل‌هایی در اختیار دانشجویان قرار می‌گیرد که در آن شواهد مورد نیاز برای هر توانمندی در کارپوشه تکوینی بر اساس منابع، روش‌ها و شرایط مشخص شده است. در سال اول کارپوشه تکوینی کاملاً ساختارمند است و دانشجویان ملزم به بازاندیشی بر ۴ توانمندی (پژوهش، دانش پزشکی، ارتباطات و تعهد حرفه‌ای) در ارتباط بسیار نزدیک با عملکرد خود در بلوک پژوهش برگزار شده در سال اول و نشان دادن سیر تکاملی خود به عنوان یک پژوهشگر هستند. در مقابل، در سال دوم در اولین کارپوشه تکوینی باید تکامل فردی و حرفه‌ای خود به عنوان یک پژوهشگر و پزشک را با بحث در مورد موفقیت‌ها و چالش‌ها و تعیین توانمندی‌های که در آن مشکل دارند، تحلیل کنند. جلسات مرور کار پوشه تکوینی با کسب مهارت ایشان در تدوین برنامه یادگیری و تعیین مسائل مربوط به رشد حرفه‌ای کمتر می‌شود. ماهیت مداوم خود ارزیابی و بازاندیشی با ارتباط نزدیک با استادان مشاور باعث شکل‌گیری عادت طبابت بازاندیشانه در دانشجویان می‌شود. مسؤولیت اصلی نشان دادن دستیابی به توانمندی‌ها به عهده دانشجویان است. برای هر جلسه مرور، دانشجویان مستندات خود را تحلیل می‌کنند و یک متن بازاندیشی می‌نویسند که تسلط در توانمندی‌ها و رشد و تکامل وی به عنوان پژوهشگر یا پزشک را نشان می‌دهد. آن‌ها یک برنامه یادگیری ساختارمند با پیامدهای قابل اندازه‌گیری و همچنین بازاندیشی بر پیشرفت خود در اهداف یادگیری جلسات قبلی تهیه می‌کنند. در زمان مرور کارپوشه استادان سه معیار را در نظر می‌گیرند:

- آیا دانشجو الگوی عملکردی خود را که در مستندات موجود در کارپوشه الکترونیکی نشان داده، تشخیص داده است؟
- سطح بازاندیشی دانشجو
- برنامه یادگیری دانشجو

جلسه به صورت گفتگو با دانشجویان در مورد مستندات، خود ارزیابی آن‌ها، و انطباق برنامه یادگیری آن‌ها با اهداف یادگیری تعیین‌شده توسط خودشان و اطمینان از قابل دستیابی بودن اهداف است.

مثال: یکی از استانداردهای توانمندی برقراری ارتباط در سال اول در دانشکده پزشکی کلیولند «نشان دادن ارتباط موثر شفاهی و کتبی در تحقیقات علوم پایه» است. مستندات مود نیاز برای جلسه مرور در مورد استاندارد مذکور، ارزیابی دانشجو توسط مربی و همکاران در آزمایشگاه، ارزیابی استاد و همگنان از ارائه در ژورنال کلاب و بازخورد تسهیل‌گر از عملکرد دانشجو در جلسات بحث در گروه‌های کوچک است. نمونه کارهای دانشجو، شامل پروپوزال، دو مورد ارائه (اسلایدهای آن) و مستنداتی که خودش انتخاب کرده است، می‌شود. مرور مستندات توسط منتور نشان می‌دهد اعضای هیأت علمی و همگنان در ارزیابی‌هایشان از فراگیر، عبارت «ساکت بودن» را ذکر کرده‌اند. این در حالی است که ارتباطات کتبی دانشجو عالی گزارش شده است. دانشجو ممکن است این الگو را تشخیص دهد یا تشخیص ندهد (معیار اول ذکر شده در بالا). دانشجو ممکن است الگو را تشخیص دهد اما کیفیت سطح بازاندیشی وی متغیر باشد (معیار دوم ذکر شده در بالا). به عنوان مثال می‌تواند بازاندیشی وی توصیفی باشد «من همیشه خجالتی بوده‌ام». در حالی که ممکن است بازاندیشی وی همراه با ایجاد ارتباط باشد و توضیح داده باشد که ساکت بودن وی ریشه فرهنگی دارد. بسته به سطح بازاندیشی ممکن است استاد مشاور به دانشجو کمک کند تا اهمیت این موضوع را تشخیص دهد، آن را تحلیل کند و این که دانشجو شروع به بررسی کند که صحبت کردن، در مراقبت از بیمار در یک تیم پزشکی چه کاربردی دارد. هر چند این موضوع از بازخوردهای ارتباط برقرار کردن در علوم پایه در آمده است اما می‌تواند در دوره‌های بعدی مثلا در جلسات حل مسأله نیز (در ادامه دوره) دیده شود. بنابراین دانشجو باید برنامه‌ای برای پاسخ به این مشکل بریزد و در کارپوشه تکوینی بعدی شواهدی ارائه دهد که به اهداف یادگیری خود در این ارتباط رسیده است. این مثال نشان می‌دهد چگونه بازخورد تکوینی در ارتباط با یک توانمندی وقتی از منابع و محیط‌های مختلف جمع‌آوری شود می‌تواند برای تعیین الگوی عملکرد استفاده شود. همچنین این مثال بر نقش استاد مشاور به عنوان مربی دانشجو و نه ارزیاب وی تأکید دارد.

ارزیابی کارپوشه

همان کیفیت‌هایی که کارپوشه را برای ارزیابی تکوینی جذاب می‌کند، کاربرد آن را برای ارزیابی تراکمی با چالش مواجه می‌کند. مهمترین مسأله در استفاده از کارپوشه برای اهداف تراکمی، ارزیابان خوب آموزش‌دیده برای قضاوت‌های محکم و مستدل است.

روش‌ها و فنون نمره‌دهی کارپوشه شامل چک‌لیست، مقیاس درجه‌بندی، روش کلی و روش تحلیلی است. در روش

تحلیلی مجموعه‌ای از معیارهای ارزیابی به کار می‌روند که برای هر یک از آن‌ها نمره جداگانه‌ای در نظر گرفته می‌شود. در روش کلی معیارهای مختلف با هم درنظر گرفته می‌شود و یک نمره واحد گزارش می‌شود. گاهی از یک روش ترکیبی استفاده می‌شود مثلاً روش تحلیلی و مقیاس درجه‌بندی در هم ادغام می‌شود و در نتیجه یک سری معیارها تدوین می‌شود و در مقیاس لیکرت نمره‌دهی صورت می‌گیرد. روش تحلیلی برای ارزیابی تکوینی کارپوشه و روش ترکیبی برای ارزیابی تراکمی مناسب است. در ارزیابی هر یک از آیتم‌های کارپوشه، روش تحلیلی و ارزیابی کل آن، روش کلی مناسب است.

لازم است برای هر یک از مقیاس‌ها در مقیاس نمره‌دهی، توصیف دقیق و اختصاصی از رفتار مناسب آن مقیاس به عمل آید تا از ذهنی بودن نمره‌دهی کاسته شود. علاوه بر این، به منظور تأمین ثبات نمره‌دهی بهتر است از نظر بیش از یک ارزیاب استفاده گردد. در صورت ارزیابی تراکمی بهتر است نام فراگیر برای ارزیاب پوشیده باشد اما چون این امکان در اکثر موارد وجود ندارد، از جمله راهبردها این است که پس از گذشت مدت زمانی از تصحیح اولیه دوباره کارپوشه را تصحیح کنیم و نتایج را با هم مقایسه کنیم یا از استادی که همان دوره را در همان مقطع ارائه می‌دهد، بخواهیم که کارپوشه را تصحیح کند. این عمل پایایی نمره‌دهی را افزایش می‌دهد. در صورتی که کارپوشه به عنوان یک سیستم ارزیابی منطقه‌ای یا ملی استفاده می‌شود، نمره‌دهی آن باید پیرو یک شیوه نظام‌مند و دقیق‌تر باشد. در این موارد معمولاً اساتیدی را آموزش می‌دهند تا با ملاک‌های یکسان کارپوشه‌ها را نمره‌دهی کنند هر چند این موضوع نیازمند صرف وقت و منابع است.

توصیه می‌شود پس از ارزیابی کارپوشه نتیجه به اطلاع دانشجو رسانده شود تا دانشجو قبل از نهایی شدن نتیجه فرصت دفاع از مستندات و عملکرد خود را داشته باشد. در برخی از برنامه‌های آموزشی که کارپوشه استفاده می‌کنند این اتفاق تنها در مورد دانشجویان مرزی و ردی رخ می‌دهد. هر چند توصیه بر این است که همه دانشجویان این فرصت را در اختیار داشته باشند. پس از این مرحله نتیجه ارزیابی به صورت اجماع نظر دو یا تعداد بیشتری ارزیاب یا یک کمیته، نهایی می‌شود.

در برخی از موارد، ارائه شفاهی دانشجو در جلسات ارزیابی یا مرور جزء الزامات آن در نظر گرفته می‌شود. اگر برنامه‌ای می‌خواهد دفاع شفاهی دانشجو از کارپوشه را پیاده کند، باید به یک سری از عوامل بالقوه توجه کند که شامل ذهنی بودن قضاوت ارزیاب در خصوص ارائه شفاهی، فقدان معیارهای استاندارد برای ارائه شفاهی، تفاسیر متفاوت ارزیابان از ارائه فراگیر و محتوای آن و همچنین انتظارات نابرابر ارزیابان است. از جمله راهبردهای مؤثر در مقابله با این عوامل می‌توان به موارد زیر اشاره کرد:

- فراگیر ارائه‌دهنده مطالب باشد و نه این‌که از کارپوشه دفاع کند.
- انتظارات ارزیابان از جلسه ارائه مانند مدت زمان ارائه، نحوه ارائه و ... به اطلاع فراگیر رسانده شود.
- ارائه شفاهی تنها تصحیح یا تثبیت‌کننده پیش قضاوت‌های ارزیابان (بر اساس مستندات و شواهد موجود در کارپوشه) باشد و نمره‌دهی بر اساس ارائه شفاهی انجام نشود.

در زیر به نمونه‌هایی از معیارها و فرایند ارزیابی کارپوشه اشاره شده است:

تجربه دانشکده پزشکی ماستریخ[1] (دریسون و همکاران ۲۰۰۵)

در این دانشکده از کارپوشه به منظور ارتقای توانمندی‌های تحصیلی و حرفه‌ای در دوره شش ساله پزشکی عمومی استفاده می‌شود. در انتهای سال اول کارپوشه با یک روند ارزیابی جامع تراکمی بر اساس معیارهای زیر مورد قضاوت قرار می‌گیرد:

- آیا تحلیل دانشجویان از نقاط قوت و ضعف خود در مورد عملکرد در نقش‌های مختلف مناسب است؟
- آیا دانشجویان اهداف یادگیری شفاف و قابل دستیابی تعیین کرده‌اند؟
- آیا دانشجویان مستندات مناسبی برای حمایت از نقاط قوت و ضعف خود ارائه داده‌اند؟
- آیا کارپوشه شامل همه موارد مورد نیاز است و به موقع حاضر شده است؟

قبل از انتهای سال منتورها برای قضاوت در مورد عملکرد، آموزش‌هایی دریافت می‌کنند که شامل بحث و بررسی کارپوشه‌های سال گذشته است. ارزیابی تراکمی به صورت نمره‌دهی توانمندی بازاندیشی به صورتِ «ضعیف»، «رضایت‌بخش» و «خوب» است. منتور نمره را در اختیار دانشجو قرار می دهد تا توافق و عدم توافق خود را اعلام کند. متعاقباً، هر کارپوشه توسط منتور دیگری نیز ارزیابی می‌شود و نمره نهایی توسط کمیته ارزیابی تعیین می‌شود که متشکل از همه منتورهاست. کمیته ارزیابی تنها کارپوشه‌هایی که بین نظر منتور، دانشجو و منتور دوم اختلاف نظر وجود دارد را بررسی می‌کند.

1. Maastricht

فرایند ارزیابی کارپوشه در کالج پزشکی CCLCM (ارزیابی تراکمی)

در دانشکده پزشکی کلیولند دانشجویان باید در انتهای سال اول، دوم و در ابتدا و انتهای سال پنجم (در یک کوریکولوم پنج ساله) کارپوشه تراکمی خود را ارائه دهند. دستورالعمل‌های دقیق برای تهیه این نوع از کارپوشه (محتوا و ساختار) در اختیار دانشجویان و استادان مشاور قرار می‌گیرد و در کارگاهی در مورد آن توضیح داده می‌شود. این دستورالعمل‌ها قطعی و غیرقابل انعطاف
هستند (برخلاف کارپوشه تکوینی) تا فرایند ارزیابی را تسهیل کنند و عدالت را تضمین کنند. بر اساس دستورالعمل‌ها، دانشجویان باید دستیابی به استانداردها را با نوشتن متنی که به نمونه‌های از مستندات ارجاع داده شده است نشان دهند. این مستندات باید متعادل باشند و از موقعیت‌های مختلف مربوط به کوریکولوم نشأت گرفته باشند. استادان مشاور کارپوشه تراکمی را بر اساس معیارهای زیر بررسی می‌کنند:

- کارپوشه مربوط به کارهای خود دانشجو است
- شواهد و مستندات انتخاب شده نمایان‌گر عملکرد کلی دانشجو است (روایی).

دانشجویان کارپوشه خود را به یک کمیته ارتقا و مرور تحویل می‌دهند که بر اساس توانمندی‌های برنامه درسی، شایستگی دانشجو برای ارتقا به مرحله بعد کوریکولوم یا فارغ‌التحصیلی را تعیین می‌کند. کمیته متشکل از اعضای هیأت علمی علوم پایه و بالینی است. مدیران گروه و مسؤولان برنامه دستیاری که تجربه ارزیابی عملکرد پزشکان و پژوهشگران را دارند یا افرادی که تجربه زیاد در نظارت بر فارغ‌التحصیلان پزشکی دارند عضو کمیته هستند. تجربه این افراد کمک می‌کند تا اعضای کمیته بتوانند قضاوت بهتری در مورد پیشرفت مناسب فراگیران در پیمودن مسیر حرفه‌ایشان داشته باشند. آموزش اعضای کمیته شامل آشنایی با کوریکولوم، روش‌های ارزیابی مورد استفاده در آن و سیستم کارپوشه است. در آموزش ارزیابان از دانشجویان فرضی استفاده می‌شود تا اعضا، درک یکسانی از رویکردهای قضاوت در مورد عملکرد دانشجو داشته باشند. کمیته ۲۱ نفر عضو دارد که در هر جلسه (پانل) ۱۶ تا ۱۸ نفر حضور دارند. هر پانل در طی دو روز به بررسی کارپوشه‌های تدوین شده توسط یک کلاس (۳۲ دانشجو) می‌پردازد.
فرایند کار به این صورت است که ابتدا هر عضو پانل چهار کارپوشه را به صورت فردی بررسی می‌کند و سپس به صورت دونفری در مورد کارپوشه‌ها بحث می‌کنند تا به اجماع و تصمیم‌گیری نهایی برسند. اگر بین دو نفر در مورد هر یک از توانمندی‌ها یا تصمیم‌گیری نهایی عدم توافق وجود داشته باشد، عضو دیگری از پانل کارپوشه را بررسی می‌کند و در صورت عدم توافق، در این مرحله کل پانل کارپوشه را بررسی و تصمیم‌گیری می‌کند. مسؤول کمیته به هر دانشجو نامه‌ای در خصوص جزئیات قضاوت کمیته در هر یک از توانمندی و عملکرد کلی ارسال می‌کند. یک کپی از نامه در پرونده دانشجو نگهداری می‌شود. قضاوت نهایی بر اساس پنج نوع تصمیم‌گیری است:

- قبول
- قبول با برخی مشکلات[1]
- قبول به شرط برخی دروه‌های اصلاحی[2]
- تکرار دوره (سال)
- اخراج از دانشکده پزشکی

در صورت «قبول با برخی مشکلات»، دانشجو باید یک طرح یادگیری با کمک استادان خود بریزد. در صورت کسب نتیجه «قبول به شرط برخی دوره‌های اصلاحی» به یک طرح رسمی که در کمیته ارتقا تصویب می‌شود، نیاز است.

1. Pass with concerns
2. Pass with remediation

نکته مورد توجه این است که به منظور منتورینگ از استادان با تجربه در این کار استفاده شود و برای ارزیابی تراکمی استادان باتجربه در ارزیابی انتخاب شوند. این دو لزوماً با هم همپوشانی ندارند. در مجموع، با توجه به آن‌چه در بالا بیان شد، توصیه می‌شود سیستمی برای طراحی و استقرار کارپوشه طراحی گردد که در آن مسیر دستیابی به توانمندی‌های مورد نظر در طول دوره آموزشی مشخص است. شکل ۳–۲۵ رویکرد کلی دانشکده پزشکی کلیولند را به ارزیابی مبتنی بر توانمندی با استفاده از کارپوشه نشان می‌دهد. پایه و اساس سیستم را مستنداتی تشکیل می‌دهند که شامل بازخوردهای تکوینی به دانشجو از منابع مختلف و از محیط‌های مختلف و نیز کارهای تولید شده توسط دانشجو است. مستندات در کارپوشه الکترونیکی به طور مرتب وارد می‌شود که قابل دسترسی توسط استادان مشاور است و استادان مشاور به طور مرتب بازخوردهای غیررسمی به دانشجو ارائه می‌دهند. در فواصل برنامه‌ریزی شده، دانشجویان گزارش‌هایی را به همراه مستندات جهت ارزیابی تکوینی و تراکمی فراهم می‌کنند و در نهایت گزارش نهایی توانمندی ارائه می‌شود.

شکل ۳-۲۵: سیستم ارزیابی مبتنی بر توانمندی در دانشکده پزشکی کلیولند با استفاده از کارپوشه (دانافر و همکاران ۲۰۰۷)

ارزشیابی کارپوشه

به طور سنتی برای بررسی کیفیت یک ابزار ارزشیابی از ویژگی‌های روان‌سنجی مانند روایی و پایایی استفاده می‌کنیم اما برخی از خصوصیات کارپوشه مانند جمع‌آوری مجموعه‌ای از مستندات توصیفی و یادداشت‌های بازاندیشی فراگیران باعث می‌شود در استفاده از این دو معیار برای ارزشیابی کارپوشه محدودیت ایجاد کند. برخی از پژوهش‌ها از کاربرد معیارهای تحقیقات کیفی در ارزشیابی کارپوشه بحث می‌کنند (دریسون و همکاران ۲۰۰۵). به نظر می‌رسد به زودی ترکیبی از روش‌های کمی و کیفی در ارزشیابی کارپوشه به وجود آید. در مجموع، ارزشیابی کارپوشه از الگوهای پیشنهاد شده برای ارزشیابی برنامه ارزیابی تبعیت می‌کند که در بخش نهم کتاب حاضر تحت عنوان «نظام ارزیابی» به آن اشاره شده است.

ارزشیابی کیفی کارپوشه (دریسون و همکاران ۲۰۰۵)

پژوهشگران با توجه به چالش‌های بررسی سنتی روان‌سنجی کارپوشه با استفاده از معیارهای پژوهش کیفی، سیستمی را برای ارزشیابی اعتبار[۱] و قابلیت اطمینان[۲] کارپوشه دانشجویان سال اول پزشکی پیشنهاد دادند. معیارهای مورد استفاده عبارت بودند از:
اعتبار:

- تلفیق: استفاده از اطلاعات از منابع مختلف در مورد یک سازه خاص.
- درگیری طولانی مدت[۳] : تعامل کافی، در طول زمان با منتورها.
- چک کردن توسط کاربران[۴]: بررسی داده‌ها با اعضای گروه به عنوان مثال فراگیران.

قابلیت اطمینان:
- ممیزی: مستند نمودن فرایند ارزیابی به منظور ارزیابی ارزیابان بیرونی.
- ممیزی قابلیت اطمینان[5] : پروسیجرهای ارزیابی کیفیت انجام شده توسط ارزیابان بیرونی.

با استفاده از رویکرد مورد استفاده در این مطالعه ۹۶ درصد کارپوشه‌ها لزومی به بررسی توسط کل کمیته نداشت. لازم است کار این پژوهشگران با فراگیران در سطوح بالاتر و در محیط‌های دیگر مورد بررسی قرار گیرد.

1. Credibility
2. Dependability
3. Prolonged engagement
4. Member checking
5. Dependability audit

تجربه دانشکده پزشکی ماستریخ در به کارگیری کارپوشه

در دانشکده پزشکی ماستریخ در کشور هلند یادگیری مبتنی بر موارد واقعی از سال اول اجرا می‌شود. به دنبال آن کارپوشه نیز در همان سال اول به منظور شکل گیری توانایی بازاندیشی در دانشجویان اجرا می‌شود. در ابتدای کوریکولوم تاکید بر مهارت بازاندیشی است زیرا این مهارت پیش نیاز یادگیری مؤثر از تجارب است. کارپوشه شامل سه بخش است:
- خودارزیابی کتبی دانشجویان از پیشرفت در چهار نقش حرفه‌ای
- مستندات مربوط به خودارزیابی
- بازخورد (کتبی) منتورهای اختصاصی هر دانشجو بر عملکرد ایشان در دو مورد قبل

دانشجویان دو بار در سال با منتور خود به صورت فردی ملاقات می‌کنند. کارپوشه تا حدود زیادی ساختارمند است و از دانشجویان انتظار می‌رود تا بر چهار نقش حرفه‌ای خود به عنوان «متخصص پزشکی»، «عضوی از حرفه پزشکی»، «دانش پژوه» و به عنوان «یک فرد» بازاندیشی داشته باشند. بازخورد منتورها به دانشجویان به صورت کتبی در کارپوشه ثبت می‌شود. مهارت بازاندیشی دانشجویان به صورت سالانه در کمیته کارپوشه بررسی می‌شود و به صورت ردی یا قبولی ارزیابی می‌شود. همه منتورها در کمیته حضور دارند اما هر منتور در مورد کارپوشه دانشجوی خود رأی نمی‌دهد.

در مطالعه‌ای که دریسون و همکاران (۲۰۰۷) به صورت مصاحبه با ۱۳ منتور این برنامه انجام دادند، شرایط موفقیت کارپوشه با هدف بازاندیشی عبارت بودند از:
- منتورینگ مناسب
- ساختار کارپوشه و دستورالعمل‌ها
- تجارب و منابع جدید کافی برای بازاندیشی
- ارزیابی تراکمی

منتورینگ و حمایت نقش بسیار مهمی را در شکل‌گیری مهارت بازاندیشی به ویژه در این مرحله از تحصیل ایفا می‌کنند. اکثر دانشجویان این مهارت را به صورت ذاتی ندارند و باید به آن‌ها کمک کرد تا بدانند زمانی که بر عملکرد خود تأمل می‌کنند چه سؤالاتی را بپرسند. موضوع دیگر کمک به دانشجویان در تشخیص نیازهای یادگیری و طراحی پلان یادگیری است. اکثر دانشجویان در زمان طراحی پلان یادگیری فهرستی از کارهایی را که باید انجام دهند، بدون ارجاع و ارتباط با نیازهای یادگیری و اهداف یادگیری می‌نویسند. از دیگر نقش‌هایی که منتورها ایفا می‌کنند افزایش انگیزه دانشجویان در انجام فعالیت‌های بازاندیشی است. نقش منتورها در شکل‌گیری توانایی بازاندیشی انقدر اهمیت دارد که برخی از پژوهشگران (پیرسون و هیود ۲۰۰۴) معتقدند بدون حمایت و مربیگری منتورها و استادان مشاور عملاً کارپوشه نقشی در ارتقای مهارت بازاندیشی ندارد.

منتورها معتقد بودند باید درخصوص انجام بازاندیشی ساختار مشخص و راهنماهایی برای دانشجویان وجود داشته باشد. این نیاز به ویژه در اوایل دوره بیشتر حس می‌شود و با کسب توانایی در انجام بازاندیشی ساختار انعطاف‌پذیرتر توصیه می‌شود. علاوه بر این، ساختارمند کردن بیش از حد فعالیت بازاندیشی از خلاقیت دانشجویان به ویژه آن‌هایی که از توانایی مناسبی در این زمینه برخوردارند می‌کاهد. به بیانی دیگر باید بین آزادی کامل و ساختار کاملاً تجویزی تعادل برقرار کرد. به طوری که دانشجویان حس مالکیت نسبت به کارپوشه داشته باشند.

باید تجارب دانشجویان از تنوع و کمیت لازم برای بازاندیشی برخوردار باشد. منتورها معتقد بودند تکراری بودن تجارب از شکل‌گیری مهارت بازاندیشی در دانشجویان پیشگیری می‌کند. در نهایت اکثر منتورها معتقد بودند در صورت عدم استفاده از فعالیت‌های بازاندیشی دانشجویان به عنوان ارزیابی تراکمی دانشجویان آن را جدی نمی‌گیرند و از موفقیت برنامه کاسته می‌شود. هر چند تعدادی نیز معتقد بودند استفاده از گزینه ارزیابی صداقت دانشجویان را کاهش می‌دهد. این دو مورد اخیر در مطالعات دیگر گزارش نشده است.

نکات مربوط به طراحی و اجرای کارپوشه الکترونیک

هر چند طراحی و اجرای کارپوشه الکترونیکی شباهت زیادی با دیگر انواع کارپوشه دارد اما مواردی وجود دارند که به طور خاص در استقرار یک سیستم کارپوشه الکترونیکی کاربرد دارند. در زیر توصیه‌هایی به منظور استقرار موفقیت آمیز کارپوشه الکترونیکی ارائه شده است (مورس و پارک[1] ۲۰۱۰):

1. Moores & Parks

- **مشخص کردن ارزش افزوده استفاده از کارپوشه الکترونیکی:** کارپوشه الکترونیکی تنها به معنی کامپیوتری کردن کارپوشه کاغذی نیست. کارپوشه الکترونیکی گزینه‌های بسیاری دارد که از جمله آن می‌توان به ایجاد ارتباط با سایت‌های اینترنتی مختلف و ویرایش و به اشتراک گذاشتن آسان‌تر اطلاعات اشاره کرد. بنابراین قبل از اقتباس یک سیستم کارپوشه الکترونیکی این سؤال باید پاسخ داده شود که این سیستم چه ارزش افزوده‌ای به ارمغان می‌آورد. در صورت مشخص کردن پاسخ این سؤال نحوه معرفی و اجرای کارپوشه تحت تأثیر قرار می‌گیرد.
- **در نظر گرفتن استفاده کوتاه‌مدت و طولانی کارپوشه الکترونیکی:** کارپوشه الکترونیکی مانند فرم کاغذی آن می‌تواند به طور موفقیت آمیزی برای یادگیری و ارزیابی در یک ماژول و در یک چارچوب زمانی مشخص یا در سطح وسیعی مانند آموزش مداوم حرفه‌ای در یک حرفه و به مدت طولانی به کار رود. بنابراین باید تلاش شود دانشجویان مهارت لازم برای استفاده از کارپوشه الکترونیکی را در طولانی مدت و در محیط کار نیز به دست آورند.
- **در نظر گرفتن زمان و نحوه معرفی کارپوشه الکترونیکی:** مطالعات نشان داده‌اند معرفی کارپوشه در مقطع دستیاری نیازمند در نظر گرفتن زمان قابل ملاحظه در مرحله معرفی ابزار است. این زمان، صرف درک فراگیران از پیامدهای مورد انتظار و کسب مهارت در استفاده از سیستم می‌شود. برای معرفی کارپوشه در مقاطع کارشناسی و پزشکی عمومی مطالعات توصیه می‌کنند بهتر است معرفی و استقرار کارپوشه الکترونیکی هر چه زودتر شروع شود. هر چند، در طول معرفی ابزار، دانشجویان باید با موارد دیگری مانند محیط‌های یادگیری مجازی و فن‌آوری‌هایی مانند پست الکترونیکی آشنا شوند و استفاده از آن را یاد بگیرند. در نتیجه اگر احساس شد که دانشجویان غرق اطلاعات شده‌اند می‌توان معرفی کارپوشه الکترونیکی را به تعویق انداخت. مهارت‌های تکنیکی ممکن است به دلیل طبیعت بصری بسیاری از سیستم‌های کارپوشه نیاز به آموزش رسمی نداشته باشند. به طور خلاصه، اگر شروع استفاده از کارپوشه در مراحل ابتدایی تحصیل و در سطح پزشکی عمومی رخ دهد، اجرای آن سریع‌تر اتفاق می‌افتد اما اگر در دوره تحصیلات تکمیلی و دستیاری شروع شود مرحله معرفی و آمادگی قبل از اجرا طولانی‌تر است.
- **ایجاد فضای یادگیری فردی برای دانشجویان:** یکی از فواید کارپوشه الکترونیکی نسبت به دیگر روش‌های یادگیری مانند محیط‌های یادگیری مجازی این است که با فراهم نمودن یک فضای یادگیری فردی حس مالکیت را برای فراگیران به ارمغان می‌آورد. تا زمانی که فراگیر اجازه دسترسی افراد دیگر را صادر نکند، کارپوشه الکترونیکی خصوصی باقی می‌ماند. فراگیران می‌توانند تعریف کنند چه کسی و به چه محتوایی دسترسی داشته باشند. هر چند در مواردی که کارپوشه ساختارمند می‌شود، مواردی مانند خلاقیت دانشجویان در محتوا و ساختار کارپوشه تحت تأثیر قرار می‌گیرد.
- **در نظر گرفتن کارپوشه به عنوان بخشی از ارزیابی دانشجویان به منظور افزایش انگیزه:** این موضوع هرچند که مختص کارپوشه الکترونیکی نیست، علاوه بر افزایش انگیزه دانشجویان موجب آشنایی دانشجویان با کاربردهای مختلف آن می‌شود
- **ارائه دستورالعمل‌های ارزیابی شفاف ولی نه بیش از حد تجویزی:** لازم است پیامدهای مورد ارزیابی و معیارهای قضاوت مشخص باشند و ساختار کارپوشه و میزان و نوع محتوایی که برای ارزیابی در کارپوشه الکترونیکی قرار می‌گیرد، برای فراگیران واضح باشد. در نوشتن معیارهای ارزیابی باید مواردی ذکر شوند که متناسب با کارپوشه الکترونیکی و نه آزمون‌های کتبی باشند.
- **تدوین راهنما درخصوص محرمانه ماندن اطلاعات بیماران و استفاده از رسانه‌های دیجیتال:** دانشجویان در مقطع بالینی ممکن است از مستنداتی مانند عکس یا رسانه‌های الکترونیکی استفاده کنند که در صورتی که رضایت بیمار را جلب نکرده باشند، می‌تواند به محرمانه ماندن اطلاعات بیماران آسیب برساند چرا که ممکن است دانشجویان این محتوا را به منظور ارزیابی عملکرد یا ارائه مستندات مربوط به آموزش مداوم در اختیار دیگران قرار دهند یا حتی به دلیل جالب بودن

در شبکه‌های اجتماعی قرار دهند. بنابراین لازم است راهنماهایی درخصوص محرمانه ماندن اطلاعات بیماران و استفاده از رسانه‌های دیجیتال تدوین شود و به دانشجویان توصیه شود مستندات را در اختیار نفر سومی حتی اگر یکی از اعضای هیأت علمی باشد قرار ندهند.

- **توجه به این امر که کارپوشه الکترونیکی به خودی خود مهارت بازاندیشی را آموزش نمی‌دهد:** به منظور سود بردن از امکانات کارپوشه الکترونیکی در زمینه آموزش مهارت مذکور، دانشجویان باید درک درستی از توانمندی بازاندیشی داشته باشند و مهارت‌های لازم در دستیابی به توانمندی بازاندیشی بر طبابت از طریق کارپوشه را کسب کنند. به بیان دیگر، کارپوشه الکترونیکی به خودی خود این مهارت را آموزش نمی‌دهد بلکه فرصتی فراهم می‌کند تا از طریق آن فراگیران مهارت‌هایی مانند بازاندیشی، خود ارزیابی و تعیین هدف و ... را به اجرا در آورند.
- **ارائه بازخورد با استفاده از کارپوشه الکترونیکی:** این امکان در کارپوشه الکترونیکی وجود دارد که بازخوردهای ارائه شده ذخیره شود و توسط دانشجویان در زمان بازاندیشی بر پیشرفت خود استفاده شود.
- **اطمینان یافتن از دسترسی دانشجویان به کارپوشه الکترونیکی خود** در برخی از محیطها به ویژه محیط بالینی دسترسی دانشجویان به کامپیوتر یا اینترنت امکان‌پذیر نیست و در مواردی مستند نمودن یک تجربه و بازاندیشی بر آن بلافاصله پس از رخداد آن مفیدتر است. در این صورت ممکن است دانشجویان همان روش‌های سنتی کتبی را ترجیح دهند. یکی از راهکارها در این زمینه استفاده از امکانات گوشی‌های همراه جهت دسترسی به کارپوشه الکترونیکی است.
- **استفاده از امکانات و حمایت‌های موجود در مؤسسه:** گاهی مسائل تکنیکی مرتبط با استقرار کارپوشه، دست‌اندکاران آن را از پرداختن به اصول آموزشی آن باز می‌دارد. بنابراین توصیه می‌شود از خدماتی که در مؤسسه آموزشی وجود دارد مانند گروه فن‌آوری اطلاعات، متخصصان فن‌آوری یادگیری و خدمات حمایت دانشجویی استفاده شود تا خدمات مناسب را به دانشجویان ارائه دهند و اعضای هیأت علمی به استفاده از اصول آموزشی در کارپوشه الکترونیکی بپردازند.
- **پرهیز از دوباره کاری:** هر چند تعداد آن زیاد نیست، اما مؤسساتی در آموزش عالی وجود دارند که از کارپوشه الکترونیکی به صورت موفقیت آمیزی استفاده کرده‌اند. تجربیات و توصیه‌های آن‌ها در گروه‌های در دسترس است و دیگران نیز می‌توانند از تجربیات آن‌ها استفاده کنند و تجربیات خود را با آن‌ها به اشتراک بگذارند. برخی منابع مفید در این زمینه در زیر آمده است:
  - E-Portfolio.ac.uk: http://www.eportfolios.ac.uk
  - HEA Health Network Group: http://www.health.ac.uk
  - JIS C: http://www.jisc.ac.uk/whatwedo/themes/elearning/
  - eportfolios.aspx
  - Pebble Learning: http://www.pebblepad.co.uk/default.asp

سودمندی کارپوشه

سودمندی کارپوشه به ویژه روایی و پایایی آن به دو عامل بستگی دارد. عامل اول ویژگی‌های روان‌سنجی آزمون‌هایی است که به عنوان محتوا در کارپوشه قرار می‌گیرند و عامل دوم نحوه قضاوت در مورد خود کارپوشه است که نیازمند حداقل ویژگی‌های روان‌سنجی است. این فرایند یک قضاوت انسانی است که باید به طور پایا تعیین کند آیا کارپوشه به عنوان یک کل می‌تواند نشان دهد که فراگیران (به ویژه برای ارزیابی تراکمی) به پیامدهای آموزشی دست یافته‌اند یا خیر. این فرایند در مورد کارپوشه با هدف تکوینی شامل قضاوت در مورد فرایندهای بازاندیشی فراگیران است.

روایی کارپوشه

در مورد روایی کارپوشه اطلاعات زیادی در اختیار نیست. به نظر می‌رسد این ابزار از روایی صوری مناسبی برخوردار است زیرا همان‌طور که پیشتر بیان شد یکی از مهمترین مزیت‌های کارپوشه فراهم نمودن اطلاعات غنی و واقعی از عملکرد روزمره، از منابع مختلف، از محیط‌های یادگیری متنوع و در طول زمان است. علاوه بر این، فرایند کارپوشه و رویکرد فردی آن به ارزیابی، مورد استقبال مدرسان و فراگیران واقع شده است.

در صورتی که بلوپرینت مناسبی تهیه شود که نمونه مناسبی از اهداف برنامه درسی یا توانمندی‌های مورد انتظار را پوشش دهد، روایی محتوایی آن افزایش می‌یابد. وسعت محتوایی که در کارپوشه میتوان جای داد وسیع است و پوشش محتوایی آن خوب است. ادعا می‌شود می‌توان هر شش حیطه توانمندی‌های ACGME را در کارپوشه مورد ارزیابی قرار داد.

جنبه واقعی بودن کارپوشه روایی پیش‌بین آن را افزایش می‌دهد. البته به این دلیل که کارپوشه در سال‌های اخیر در آموزش پزشکی رواج یافته است، مطالعات طولانی‌مدت که ارتباط کارپوشه با عملکرد حرفه‌ای آتی را نشان دهند، وجود ندارند. بررسی روایی همزمان کارپوشه نیز مشکل است به این دلیل که آزمون‌های دیگر معمولاً با این وسعت عملکرد دانشجویان را مورد ارزیابی قرار نمی‌دهند.

با توجه به این‌که کارپوشه توانمندی بازاندیشی را مورد ارزیابی قرار می‌دهد، از روایی سازه خوبی برخوردار است (هولمبو و هاوکینر ۲۰۰۸). هر چند، زمانی که کارپوشه با هدف تراکمی به کار می‌رود، به نظر می‌رسد فراگیران تنها به تکمیل مستندات و شواهد مربوط می‌پردازند و از ارائه یادداشت‌های بازاندیشی پرهیز می‌کنند. بنابراین، این چالش مطرح می‌شود که آیا واقعا کارپوشه تراکمی توانمندی بازاندیشی را مورد ارزیابی قرار می‌دهد؟

بخشی از پیچیده بودن روایی کارپوشه به دلیل ساختار باز آن است. ارزیابان باید کارپوشه‌هایی را داوری کنند که از نظر محتوا، حجم و در بسیاری موارد ساختار متفاوت هستند. یکی از منابع بالقوه سوگیری در قضاوت ارزیابان این است که ارزیاب ممکن است تحت تأثیر کیفیت‌های نامربوط مانند کیفیت نوشتن، ساختار، صفحه آرایی، خصوصیات فردی دانشجو و شناخت قبلی از دانشجو و در مورد ارائه شفاهی، تحت تأثیر نحوه ارائه دانشجو قرار گیرد. بنابراین به منظور تولید نمرات معتبر در کارپوشه ضروری است که ارزیابان به طور آگاهانه بین عوامل مرتبط و غیرمرتبط تأثیرگذار بر ارزیابی تمایز قائل شوند.

دریسون و همکاران (۲۰۰۶) به منظور بررسی میزان تأثیر عوامل نامربوط بر نمره‌دهی کارپوشه، ابزاری ۱۵ آیتمی تحت عنوان Portfolio Quality Analysis Scoring Inventory طراحی کردند و بر اساس آن دو نفر به صورت مستقل کارپوشه‌ها را نمره‌دهی کردند. توافق بین ارزیابان ۰/۴۶ تا ۰/۸۷ بود. نتایج نشان داد کیفیت بازاندیشی تنها آیتمی بود که سهم مشخصی (۶۴ درصد) در توضیح واریانس نمره‌دهی کارپوشه داشت که مطلوب است. عوامل دیگر مانند صفحه‌آرایی، نکات دستوری و ساختار کارپوشه تأثیر غیربارزی در نمره‌دهی داشتند. این موضوع می‌تواند به دلیل آموزش منتورها در نمره‌دهی کارپوشه باشد. یافته‌های این مطالعه از روایی مناسب کارپوشه در زمان قضاوت کلی آن حکایت دارد. هر چند مطالعات بیشتری مورد نیاز است تا عوامل مختلف تاثیرگذار بر قضاوت ارزیابان را در هنگام ارزیابی کارپوشه مورد بررسی قرار دهد.

یکی از سؤالات دیگری که در مورد روایی کارپوشه مطرح است این است که آیا اگر این روش برای تصمیم‌گیری در مورد ارتقای دانشجویان به سال (سطح) بالاتر و فارغ‌التحصیلی به کار رود، عادلانه است؟ رعایت عدالت در آزمون مفهومی ذهنی است و اندازه‌گیری آن در یک آزمون ساختارمند مشکل است چه رسد به آزمونی مانند کارپوشه که ساختار آن بسیار متغیر است. معیارهای رعایت عدالت در آزمون به صورت کلی عبارت هستند از:

- رفتار عادلانه آزمون‌گران
- پیامدهای برابر برای زیرگروه‌های فراگیران

- نبود آیتم‌ها و ساختارهای همراه با سوگیری
- فرصت‌های برابر برای یادگیری

تاکنون تنها یک مطالعه به بررسی عادلانه بودن کارپوشه به صورت تجربی پرداخته است که صرفاً معیار دوم یعنی پیامدهای یادگیری برای زیرگروه‌ها را بررسی کرده است. در این مطالعه ارتباط بین تصمیمات اتخاذشده بر مبنای کارپوشه با جنس، وضعیت شهروندی و توانایی زبانی (مانند سلیس بودن در زبان انگلیسی) بررسی شده است. بر اساس نتایج این مطالعه، ارتباط معنی‌داری بین تصمیمات نهایی در مورد عملکرد و ویژگی‌های زیرگروه‌های دانشجویان مشاهده نشد. البته به طور قطع نمی‌توان با تکیه بر نتایج این مطالعه، عادلانه بودن کارپوشه را تأیید کرد زیرا تنها یک جنبه از رعایت عدالت بررسی شده است (بیرر و دانفر[1] ۲۰۱۱).

یکی از چالش‌های مربوط به روایی کارپوشه این است که چگونه می‌توانیم مطمئن شویم که مستنداتی که دانشجو در کارپوشه قرار داده است، ماحصل فعالیت‌ها و تلاش خود دانشجو است. هر چند در متون به سرقت ادبی[2] به عنوان مسأله جدی در روایی کارپوشه پرداخته نشده است اما برخورد با این مسأله همچنان با ابهام همراه است.

پایایی کارپوشه

همان‌طور که در مورد بسیاری از ارزیابی‌های مبتنی بر محل کار مصداق دارد، کارپوشه نیز نیازمند قضاوت انسانی است که بدون شک این قضاوت عاری از خطا نیست. این امر می‌تواند تعمیم‌پذیری ارزیابی عملکرد فراگیر را تهدید کند. مسائل خاص مربوط به نمره‌دهی ارزیابان در کارپوشه شامل ثبات قضاوت بین ارزیابان[3] و درون ارزیابان[4]، ثبات قضاوت در طول زمان و قابلیت تکرار تصمیم‌گیری‌های رد و قبول است.

در زمینه بررسی پایایی کارپوشه مطالعات محدودی در آموزش پزشکی انجام شده است و اغلب آن‌ها به بررسی پایایی بین ارزیابان و درون ارزیابان در نمره‌دهی کارپوشه پرداخته‌اند. نتایج این مطالعات متغیر بوده است که به نظر می‌رسد به دلیل متفاوت بودن معیارهای مود استفاده برای نمره‌دهی بوده است. به عنوان مثال، پژوهش‌های پیتس و همکاران[5] از پایایی پایین کارپوشه حکایت دارد در حالی که نتایج مطالعه دریسون و همکاران و نیز مطالعه ریس و شرد[6] مثبت بوده است.

پیتس و همکاران (۱۹۹۹، ۲۰۰۱، ۲۰۰۲)

مطالعه فوق از اولین مطالعاتی است که در حوزه آموزش پزشکی به بررسی پایایی کارپوشه پرداخته است. ۱۲ کارپوشه دانشجوی پزشکی عمومی در مقابل شش معیار از پیش تعیین شده توسط هشت ارزیاب (دارای تجربه مکفی در آموزش پزشکی عمومی) مورد ارزیابی قرار گرفتند. چارچوب راهنمای نمره‌دهی در اختیار ارزیابان قرار گرفت. در نهایت ارزیابان در مورد قبولی یا نیازمند ارجاع بودن فراگیران تصمیم‌گیری کردند. هر هشت ارزیاب هر یک از کارپوشه‌ها را در دو جلسه به فاصله یک ماه نمره‌دهی کردند. ضریب توافق (کاپا) بین ارزیابان برای شش معیار از ۰/۱ تا ۰/۴۱ متغیر بود. ضریب توافق درون ارزیابان از ۰/۳۸ تا ۰/۵۴ متغیر بود. این ضرایب برای تصمیم‌گیری نهایی در خصوص سرنوشت فراگیر کافی و قابل قبول نبود.
پژوهشگران مجدداً مطالعه را با همان شرایط بر روی ۱۳ نفر از ورودی دیگری از دانشجویان در دوره مشابه انجام دادند و به نتایج مشابه دست یافتند. در مطالعه‌ای جدیدتر (۲۰۰۲) پژوهشگران اثر بحث بین دو ارزیاب را در زمان نمره‌دهی بر توافق بین ارزیابان بررسی کردند و به این نتیجه رسیدند که بحث و گفتگو بین ارزیابان ضریب توافق را تا ۰/۵ بهبود می‌بخشد.

1. Bierer & Dannefer
2. Plagiarism
3. Inter-rater
4. Intra-rater
5. Pitts et al.
6. Rees & Sheard

دریسون و همکاران (۲۰۰۷)

دریسون و همکاران در یک مقاله مروری نظام‌مند که به منظور بررسی کارایی کارپوشه انجام شد نشان دادند که شش مطالعه پایایی نسبتا خوبی را برای کارپوشه گزارش کرده‌اند. به طوری‌که به ترتیب برای یک، دو، سه و چهار ارزیاب پایایی ۰/۶۳، ۰/۷۷، ۰/۸۴ و ۰/۸۷ به دست آمد.

ریس و شرد (۲۰۰۴)

در این مطالعه ۱۰۰ کارپوشه دانشجویان سال دوم پزشکی در دوره آموزشی مهارت‌های ارتباطی توسط دو ارزیاب بر اساس پنج معیار مشخص مورد قضاوت قرار گرفت. ارزیابان ابتدا به طور مستقل و مجدداً پس از بحث و گفتگو نمره‌دهی کردند. میزان توافق (ICC) بین ارزیابان برای نمره کل معیارها ۰/۷۷ و برای هر یک از معیارها بین ۰/۳۶ تا ۰/۶۹ متغیر بود.

هر چند مطالعات اخیر در حوزه آموزش پزشکی از توافق بین ارزیابان مطلوب کارپوشه حکایت دارد اما نتایج پژوهش‌ها در حوزه‌های دیگر تا این حد خوش‌بینانه نبوده است. دو مقاله مروری انجام‌شده در پرستاری نشان داد مطالعات این حوزه نتایج بسیار متنوعی را از پایایی کارپوشه (استفاده شده با هدف ارزیابی تراکمی) گزارش کردند. پژوهشگران نتیجه گرفتند احتمالاً دیدگاه سنتی در مورد پایایی برای محاسبه پایایی کارپوشه با اطلاعات جامع آن کارآیی نداشته باشد و روش‌های کمی به تنهایی پاسخگو نیست (مک‌کردی[1] ۲۰۰۷، مک‌مولان و همکاران[2] ۲۰۰۳).

اقداماتی که به منظور بهبود توافق بین و میان ارزیابان توصیه می‌شود عبارت هستند از:

- استفاده از ارزیابان متعدد
- آموزش ارزیابان
- تدوین معیارهای کلی[3] واضح و یکسان و تدوین روبریک‌های ارزیابی
- اطلاع رسانی معیارها و ابزارهای ارزیابی به ارزیابان
- ایجاد فهم و درک مشترک بین اعضای هیأت علمی از اهداف ارزیابی و معیارهای ارزیابی
- بحث و گفتگوی ارزیابان قبل از شروع ارزیابی و پس از ارزیابی بخشی از کارپوشه

با توجه به پیچیدگی ارزیابی عملکرد از طریق کارپوشه، زمان طولانی که صرف تکمیل آن می‌شود و دانش ارزیابان در مورد فراگیران که در طول زمان شکل می‌گیرد، مطالعه ثبات قضاوت ارزیابان در یک مدت زمان کوتاه امکان‌پذیر نیست. ثبات ارزیابان باید در ثبات تفاسیر ایشان از الگوهای عملکرد فراگیران در طول زمان و در مورد تکالیف مختلف باشد و نه توافق بر یک واحد از عملکرد یا بر یک نمره واحد. ارزیابان باید به فهم مشترکی از پیشرفت دانشجویان در مسیر پیامدهای یادگیری برنامه درسی برسند.

مهمترین نگرانی در مورد عدم توافق بین ارزیابان، اشتباه در مورد رد و قبولی دانشجویان یا نوع ارجاع ایشان به دلیل خطای ارزیابان است. توافق بین ارزیابان برای بخش‌های کارپوشه ممکن است بالا باشد اما آن‌چه مهم است تکرارپذیری تصمیمات رد و قبول ارزیابان است.

تجربه دانشگاه داندی [1]

کارپوشه دانشجویان پزشکی دانشگاه داندی شامل نمونه‌گیری وسیعی از فعالیت‌ها و دستاوردهای دانشجویان است. این کارپوشه شامل چندین ارائه بیمار و مراقبت از بیمار، لاگ‌بوک پروسیجرها، ارزیابی انتهای چرخش بالینی، گزارشات بازاندیشی ساختارمند، قراردادهای آموزشی و ماژول‌های انتخابی (SSM) است. در این سیستم، معیارها و انتظارات از دانشجو مشخص است. در ارزشیابی کارپوشه دانشگاه داندی توافق تصمیمات ردی و قبولی بین دو ارزیاب ۹۸ درصد گزارش شده است. در این برنامه ارزیابان آموزش قابل ملاحظه‌ای دریافت می‌کنند و قضاوت در مورد ردی و قبولی بر اساس یک رویکرد رسیدن به اجماع است.

1. Dundee

1. McCready
2. McMullan et al.
3. Global

توصیه می‌شود در ابتدا برای دانشجویان و برای افراد حرفه‌ای که تازه شروع به کار کرده‌اند از کارپوشه‌های ساختارمند استفاده شود. با رشد فرد به نظر می‌رسد روش‌های کیفی‌تر برای ارزیابی کارپوشه مناسب‌تر است تا بتواند توانمندی‌های کمتر ملموس مانند قضاوت‌ها و ارزش‌های حرفه‌ای را ارزیابی کند.

تأثیر آموزشی کارپوشه

به نظر می‌رسد اگر کارپوشه به خوبی پیاده شود، موجب بهبود یادگیری فردی و حرفه‌ای و افزایش مسؤولیت فراگیران در برابر یادگیری خود شود. این باور عمومی وجود دارد که کارپوشه هم بر فرایند و هم بر پیامد یادگیری تأثیرگذار است. در مجموع، کارپوشه موجب بهبود بازخورد و خودآگاهی، ارتقای دانش و درک (شامل ادغام تئوری و عمل) و آمادگی برای تحصیلات تکمیلی می‌شود. ارتباط بین دانشجو و مربی را بهبود می‌بخشد و توانایی مقابله با شرایط احساسی مشکل‌دار را بهبود می‌دهد.

اکثر مطالعاتی که با هدف بررسی پیامدهای کارپوشه بر یادگیری انجام شده‌اند، مطالعه موردی، مقطعی و در سطح بررسی واکنش دانشجویان نسبت به یادگیری خود بوده‌اند. در زیر به برخی از پیامدهای کارپوشه بر یادگیری اشاره می‌شود:

- **بهبود یادگیری (حیطه دانشی):** اکثر مطالعات نتایج مثبتی از تاثیر کارپوشه بر یادگیری دانشجویان و تعداد اندکی نتایج خنثی یا منفی گزارش کرده‌اند. استفاده از کارپوشه توانایی دانشجویان را در تشخیص نیازهای یادگیری، رسیدن به اهداف دوره و نیز پایش دستیابی به اهداف یادگیری بهبود می‌بخشد و توانایی آن‌ها را برای ادغام بهتر یادگیری از منابع مختلف و به ویژه ادغام موضوعات تئوری و عملی بهبود می‌دهد. نتایج نمونه‌ای از مطالعات انجام شده در این‌جا ارائه می‌شود:
  - از نظر فراگیران، کارپوشه موجب یک رویکرد عمیق‌تر به مطالعه موضوعات و فرایند یادگیری می‌شود (تیواری و تنگ[1] ۲۰۰۳، وب[2] ۲۰۰۶).
  - در پیگیری از داروسازانی که از کارپوشه استفاده کردند، دیده شد آن‌ها برای خود اهداف یادگیری می‌نوشتند و در دوسوم موارد به اهداف تعیین شده دست یافتند (استین و همکاران[3] ۲۰۰۵).
  - هیچ تفاوتی بین نمرات آزمون‌های دانشجویانی که از کارپوشه استفاده کرده‌اند، با دانشجویان دیگر که از کارپوشه استفاده نکرده‌اند مشاهده نشد. در حالی‌که در مورد دانشجویانی که از کارپوشه استفاده کردند، بین میزان استفاده از کارپوشه (میزان متن نوشته شده) و نمره انتهای بخش ارتباط وجود داشت (گرانت و همکاران[4] ۲۰۰۶).
  - دانشجویان فعال‌تر در استفاده از کارپوشه، در مشاهده پروسیجر فعال‌تر بودند که ممکن است به دلیل تعهد بیشتر این دانشجویان باشد و نه به دلیل استفاده از کارپوشه (لونکا و همکاران[5] ۲۰۰۱).
  - دانشجویانی که به تکمیل کارپوشه پرداختند، نسبت به هم‌کلاسی‌های خود، دانش بیشتری از انکولوژی را نشان دادند به طوری‌که بعد از هشت هفته استفاده از کارپوشه در فرمول‌بندی پاسخ مناسب به یک تمرین بالینی نسبت به گروه مداخله بهتر عمل کردند (مطالعه فینلی و همکاران ۱۹۹۸)
  - فراگیران پرستاری که از کارپوشه استفاده کردند، درک بهتری از تئوری پرستاری داشتند، به یادگیری عمیق‌تر و کاربرد یادگیری در عمل رغبت بیشتری نشان دادند (تیواری و تنگ ۲۰۰۳)
  - رویکرد استفاده‌کنندگان از کارپوشه در مقایسه با پزشکانی که از سیستم سنتی آموزش مداوم استفاده کردند، به میزان بیشتری منطبق بر چرخه یادگیری تجربی کلب[6] بود. همچنین کاربران اظهار داشتند که در کارپوشه، آن‌ها با

1. Tiwari & Tang
2. Webb
3. Austin et al.
4. Grant et al.
5. Lonka et al.
6. Kolb experiential learning cycle

موضوعات بیشتری برای مطالعه و نیز فعالیت‌های یادگیری بیشتری سر و کار داشتند (مطالعه مترز و همکاران[1] ۱۹۹۹).

- متخصصان تغذیه که در طرح کارپوشه شرکت کرده بودند، به طور معنی‌داری سریع‌تر از گروه کنترل، ارزیابی نیازهای یادگیری خود را انجام دادند (P=۰/۰۰۲) و همچنین در تدوین برنامه یادگیری نیز سریع‌تر عمل کردند (P=۰/۰۱۸) (کیم و همکاران[2] ۲۰۰۱).

- **ارتقای مهارت**: استفاده از کارپوشه موجب بهبود خودآگاهی، افزایش مشارکت دانشجویان در بازاندیشی و ترغیب به آن، بهبود توانایی یادگیری مستقل، مهارت تصمیم‌گیری و تفکر نقاد و همچنین بهبود مهارت‌های ارتباطی می‌شود. استفاده از کارپوشه موجب تسریع نگرش نقادانه نسبت به عملکرد خود و کمک به مدیریت پیشرفت خود می‌شود. علاوه بر موارد مذکور، کارپوشه موجب ارتقای مهارت‌های فناوری اطلاعات و سازماندهی دانشجویان می‌شود.

- در مطالعه کمبل و همکاران[3] (۱۹۹۶)، دو سوم شرکت‌کنندگان گزارش کردند کارپوشه باعث شد آن‌ها بر مراقبت از بیمار بازاندیشی کنند و یادداشت کنند کدام فعالیت آموزشی خبرگی آن‌ها را افزایش داده است.

- مطالعه دیگری نیز ارتقای تصمیم‌گیری بالینی در دانشجویان پرستاری به دنبال استفاده از کارپوشه را گزارش کرد (باکلی و همکاران[4] ۲۰۰۹).

- این احتمال وجود داد که ارتقای بازاندیشی توسط کارپوشه تحت تأثیر دو مسأله ترجیحات یادگیری فرد و برخی جنبه‌های خود کارپوشه قرار گیرد. اسالیوان و همکاران[5] (۲۰۱۲) در مطالعه خود به بررسی درک دانشجویان در مورد نقش کارپوشه در ارتقای توانمندی‌هایی مانند بازاندیشی بر طبابت، درک اخلاقیات و یادگیری خودراهبر در دو دانشکده پزشکی با دو کارپوشه با ساختار متفاوت پرداختند. نتایج مطالعه نشان داد کارپوشه جدا از ساختار آن، باعث بهبود این مهارت‌ها می‌شود. در این مطالعه تأثیر کارپوشه بر کار گروهی پایین گزارش شد و نیز تأثیر بارزی بر مهارت‌های ارتباطی گزارش نشد که پژوهشگران علت احتمالی آن را عدم ارائه شفاهی کارپوشه عنوان کردند.

- **نگرش و رفتار**: استفاده از کارپوشه موجب افزایش اعتماد به نفس فراگیران، بهبود در نگرش آن‌ها شامل تمایل برای به عهده گرفتن مسؤولیت یادگیری و ارتقای تعهد حرفه‌ای، بهبود در دیدگاه دانشجویان نسبت یادگیری و رضایت آن‌ها نسبت به فرایند یادگیری می‌شود:

- با استفاده از کارپوشه، اعتماد به نفس دانشجویان از طریق به اشتراک گذاشتن تجاربشان با دیگر دانشجویان افزایش پیدا می‌کند. این افزایش اعتماد به نفس، دانشجویان را برای دوره دستیاری که استفاده از کارپوشه در آن بیشتر است، آماده می‌کند (تاچل و همکاران[6] ۲۰۰۹).

- استفاده از کارپوشه در دستیاران بخش زنان و زایمان نسبت به دستیاران در دانشکده‌های دیگر که با کارپوشه مواجه نبودند، آگاهی از یادگیری خودراهبر[7] (P< ۰/۰۵)، نگرش مثبت نسبت به یادگیری مادام‌العمر (P< ۰/۰۰۱) و علاقه نسبت به یادگیری جدید (p< ۰/۰۱۸) را افزایش داد (فانگ و همکاران[8] ۲۰۰۰).

- نگرش فراگیران نسبت به رشد حرفه‌ای و کارآمدی نسبت به انجام خود ارزیابی در ابتدای استفاده از کارپوشه بالا بود و پس از دو سال پیگیری تغییری نکرد (کیم و همکاران ۲۰۰۱)

- فراگیرانی که از کارپوشه استفاده کردند بهتر توانستند از عهده یک موقعیت ابهام برانگیز برآیند و در طول یکسال، درصد گزارش آن‌ها برای مقابله با موقعیت‌های مشکل‌دار کاهش یافت (دالوف و همکاران[9] ۲۰۰۴).

1. Mathers et al.
2. Keim et al.
3. Campbell et al.
4. Buckley et al.
5. O'Sullivan et al.
6. Tochel et al.
7. Self-directed learning
8. Fung et al.
9. Dahllof et al.

- **تاثیر بر دیدگاه و رفتار مربیان**: علاوه بر تأثیر بر دانشجویان، کارپوشه می‌تواند بر دیدگاه و رفتار مربیان نیز تأثیر داشته باشد. مطالعات گزارش کردند که منتورها و مدرسان در فرایند مرور و ارائه بازخورد به کار دانشجویان از نیازهای دانشجویان بیشتر آگاه می‌شوند و در نتیجه این موضوع رویکرد آن‌ها به تدریس را تحت تاثیر قرار می‌دهد. اجرای کارپوشه اثراتی در فعالیت‌های کلاسی معلمان و نیز برنامه درسی داشته است، به طوری‌که معلمان پس از آن از روش‌های تدریس حل مسأله و کار در گروه‌های کوچک بیشتری استفاده کرده‌اند. همچنین کارپوشه باعث بازنگری برنامه درسی در بخش محتوا و راهبردهای تدریس شده است. کارپوشه فضای تدریس را با ادغام ارزیابی در تدریس تغییر می‌دهد (اسنادن و توماس[1] ۱۹۹۸). پژوهش‌ها از طیف وسیعی از حرفه‌ها گزارش کرده است که استفاده از کارپوشه باعث بهبود ارتباط بین فراگیر و مربی می‌شود. کارپوشه با به اشتراک گذاشتن احساسات در محیطی امن، تکمیل کننده بحث‌هایی است که در کلاس امکان آن وجود ندارد. مربیان می‌توانند برای دانشجویانی که در موقعیت‌های مشکل‌دار قرار می‌گیرند حمایت احساسی فراهم کنند، به دانشجویانی که با موقعیت‌های مشکل‌دار مانند مرگ یک بیمار مواجه می‌شوند کمک کنند و حتی ممکن است دانشجو موارد مشکل‌دار مربوط به زندگی فردیش را نیز پررنگ کند. هر چند در بسیاری از موارد نیز دانشجویان اذعان دارند که نمی‌خواهند احساساتشان را بیان کنند.
- **اثرات ناخواسته (جانبی)**: اجرای کارپوشه ممکن است منجر به برخی پیامدهای ناخواسته شود. به عنوان مثال، مطالعات بسیاری گزارش کردند که کامل کردن کارپوشه زمان‌بر و استرس‌زا است و تاکنون دو مطالعه گزارش کردند که فراگیران معتقدند کامل کردن کارپوشه به دلیل زمان‌بر بودن باعث شده که آن‌ها از کار بالینی غافل شوند. مسأله زمان‌بر بودن می‌تواند برای منتورها و مربیان هم پیامدهای نامطلوبی را ایجاد کند (دیویس و همکاران ۲۰۰۱). به طوری که در مطالعه ریس و همکاران (۲۰۰۵) یکی از شرکت‌کنندگان ذکر کرده است: «در حالی که شما در حال پر کردن کاغذها هستید، بیماران شما در تخت بستری از دست می‌روند».

پیامدهای ناخواسته کارپوشه در محیط بالینی (نگلر و همکاران[1] ۲۰۰۹)

استفاده از کارپوشه در آموزش پزشکی به ویژه برنامه‌های دستیاری پذیرفته شده است، در حالی‌که هنوز در مورد محتوا، اهداف، مفید بودن و مهم‌تر از همه رعایت حریم خصوصی و محرمانه ماندن اطلاعات دغدغه‌هایی وجود دارد.

پزشکان، مسؤولان بیمارستان‌ها و اعضای هیأت علمی باید از عواقب کاربرد کارپوشه در محیط بالینی آگاه باشند. کاربرد آن در محیط بالینی مسائلی را رقم می‌زند که در حوزه‌های دیگر کمتر رخ می‌دهد. از جمله این مسائل می‌توان به حفظ حریم خصوصی بیماران، افشای اطلاعات بالینی و ایجاد برخی مسؤولیت‌های بالینی برای پزشکان اشاره کرد.

به عنوان مثال، مستندسازی عملکردهای ضعیف برای برنامه‌ریزی یادگیری فردی و توانمندسازی حرفه‌ای بسیار مفید است، اما این مستندات می‌توانند به صورت بالقوه برای حرفه پزشکی آسیب‌رسان باشند اگر برای دعاوی قصور پزشکی استفاده شوند. از نظر آموزشی ادغام و مستند کردن بازاندیشی در کارپوشه کار شایسته‌ای است. اما هنوز نحوه حفاظت از این اطلاعات مشخص نیست. اگر نحوه استفاده از کارپوشه به عنوان یک ابتکار در محیط بالینی به درستی تعیین نشود و با خوش بینی افراطی و بدون تفکر در مورد جزئیات اجرا باشد، می‌تواند عواقب قانونی ناخواسته به همراه داشته باشد. در نتیجه باید قوانین محکم، آیین‌نامه‌ها و کدهایی تعریف شود تا از این داده‌ها حفاظت کنند. در غیر این‌صورت دستیاران و موسسه را در معرض خطر قرار می‌دهد.

منتقدان کارپوشه معتقد هستند هنوز ارزش و پیامدهای آن مشخص نیست و شواهد محکمی برای روایی و پایایی آن به عنوان یک ابزار ارزیابی وجود ندارد، مضاف بر این‌که کاربرد آن به عنوان یک ابزار ارزیابی مسؤولیت‌های جدیدی را برای پزشکان و مؤسسات به ارمغان می‌آورد. هر چند، ACGME ادعا می‌کند که از محتوای کارپوشه حفاظت می‌کند اما هیچ سیستم قانونی این موضوع را بررسی نکرده است. کارپوشه ابزاری فردی است توسط دستیار تهیه می‌شود و دستیار صاحب آن است. اما بیمارستان یا دانشکده پزشکی مربوط این اطلاعات را در اختیار ACGME قرار می‌دهد. بنابراین هر سه – دستیار، مؤسسه آموزشی و ACGME- می‌توانند به عنوان نگهداران مشروع این اطلاعات باشند و بنابراین در مواردی برای در اختیار دیگران قرار دادن این اطلاعات مورد درخواست قرار می‌گیرند. آیا در مواردی که کارپوشه در اختیار کمیته مرور و ارزیابی قرار می‌گیرد اطلاعات آن محرمانه می‌ماند؟ آیا خود فرایند ارزیابی و تصمیمات اتخاذ شده در کمیته محرمانه می‌ماند؟ بنابراین چه انتظاری داریم پزشکان خطاها و قصور پزشکی و پیامدهای نامناسب خود را در کارپوشه قرار دهند؟ در حالی که از عدم استفاده از این اطلاعات برای شکایات بر علیه پزشکان مطمئن نیستیم. این‌ها مواردی است که می‌تواند موجب عدم استفاده برنامه‌های دستیاری از کارپوشه به عنوان یک ابزار ارزیابی با درج موارد مربوط به بازاندیشی و پیگیری رشد حرفه‌ای شود. بنابراین، مؤسسات آموزشی و اعتباربخشی باید با دقت سیاست‌های ضبط مستندات را تدوین کنند و آن را اجرا نمایند.

1. Nagler et al

1. Snadden & Thomas

هزینه، قابلیت اجرا و مقبولیت کارپوشه

کارپوشه در صورتی مقبولیت و قابلیت اجرای مناسبی دارد که استفاده از آن آسان بوده و از نظر صرف زمان و هزینه برای داوطلب، ارزیاب و تیم اجرا کننده به صرفه باشد. مهمترین عوامل مؤثر در استفاده از کارپوشه عبارت هستند از:

- **نگرش کاربران**: نگرش کاربران نسبت به کارپوشه در استفاده از آن اهمیت بسیار دارد. در صورتی که کاربران نگرش مثبتی نسبت به این روش و کاربرد آن داشته باشند، بالتبع میزان استفاده از آن افزایش خواهد یافت. نگرش کاربران به استفاده از کارپوشه به منظور ارتقای بازاندیشی ضد و نقیض است. در برخی مطالعات اکثر مشارکت‌کنندگان معتقد بودند کارپوشه باید برای ارتقای بازاندیشی استفاده شود اما در برخی دیگر کاربران مطمئن نبودند مواردی که در کارپوشه قرار می‌دهند، علیه آن‌ها استفاده نشود. برخی دیگر اظهار داشتند آن‌ها از مهارت بازاندیشی برخوردارند و اجبار به استفاده از یک ابزار برای این‌کار با رویکرد آن‌ها برای رشد حرفه‌ای تداخل ایجاد می‌کند.
- **جنسیت کاربران**: مطالعات نشان داده‌اند پزشکان تازه فارغ‌التحصیل‌شده زن بیشتر کاربر کارپوشه بودند و درک مثبت‌تری از تأثیر آموزشی آن داشتند. اسالیوان و همکاران (۲۰۱۲) مقبولیت خوبی را از سوی دانشجویان پزشکی گزارش کردند و دانشجویان دختر از این نظر نمره بالاتری به کارپوشه دادند.
- **سطوح متفاوت حمایت سازمانی در پیاده‌سازی کارپوشه**: به نظر می‌رسد حمایت موسسه در سطوح مختلف در پیاده سازی کارپوشه موثر باشد. وب و همکاران (۲۰۰۶) گزارش کردند کارپوشه‌ای که برای جراحان در آمریکا اجرا شد و شامل خود ارزیابی و بازاندیشی نیز بود، با میزان مشارکت اولیه کمی (کمتر از ۵۰ درصد) همراه بود اما پس از انجام اصلاحاتی در فرایند اجرا شامل ارائه بازخورد ماهانه به پزشکان، اجرای بحث‌های سازمان‌دهی‌شده، امکان تماس پزشکان استفاده‌کننده با ناظران کارپوشه از طریق پست الکترونیکی و اطلاع‌رسانی مرتب (به صورت فصلی و به صورت یادآوری مشارکت در کارپوشه) میزان مشارکت به ۱۰۰ درصد رسید. نویسندگان دو عامل را از عوامل موفقیت دانستند: مرور کارپوشه توسط اعضای هیأت علمی اختصاص یافته به هر فرد و درک اهمیت پروژه توسط دست‌اندرکاران. بنابراین می‌توان نتیجه گرفت فرایند پیاده‌سازی کارپوشه از ساختار کارپوشه مهمتر است. از دیگر حمایت‌های سازمانی در اجرای صحیح کارپوشه می‌توان به اختصاص زمانی جداگانه برای کاربران در برنامه درسی به منظور تکمیل (فراگیران) یا مرور (منتورها و ارزیابان) کارپوشه، ارائه اطلاعات درباره چگونگی استفاده از کارپوشه به استفاده‌کنندگان و استفاده از تکنیک‌های عملی اشاره نمود. از زمانی که فرد با کارپوشه تماس اولیه پیدا می‌کند تا درگیر شدن کامل با آن در ارتباط بودن با وی بسیار مهم است.
- **حمایت و منتورینگ اولیه یا مداوم**: منتورینگ یکی از عوامل مهم در موفقیت کارپوشه است. همان‌طور که اشاره شد، بازخورد ماهانه از منتورهای اختصاص یافته به هر فرد میزان مشارکت را از حدود ۵۰ به ۱۰۰ درصد افزایش داد (وب و همکاران ۲۰۰۶). منتورینگ علاوه بر افزایش میزان مشارکت در تکمیل کارپوشه، میزان بازاندیشی را هم افزایش می‌دهد. در مورد تأثیر منتورینگ بر تداوم استفاده از کارپوشه مطالعات کمی انجام شده است. دیده شده است بین میزان ارائه بازخورد توسط منتور و احتمال استفاده مداوم از کارپوشه ارتباط وجود دارد. علاوه بر این، برخی از فراگیران از مهارت‌های شناختی و بازاندیشی کافی برخوردار نیستند و نمی‌توانند از کارپوشه استفاده کافی کنند. در نتیجه حمایت باید با نیازها منطبق باشد و فراگیران با سطح مهارت‌های بازاندیشی ناکافی مورد حمایت بیشتری قرار گیرند. همچنین دغدغه‌هایی در مورد دانش ناکافی و عدم درک منتورها از کارپوشه وجود دارد. بنابراین باید نحوه استفاده کارپوشه برای ارزیابان نیز آموزش داده شود و تأثیر و اهمیت آموزشی آن برای هر دو گروه به طور دائم نشان داده شود. نتایج مطالعه در برنامه پیش‌دستیاری انگلیس نشان داد که بیش از نیمی از فراگیران معتقد بودند منتورها دانش کافی برای استفاده از کارپوشه ندارند (ریسوس و همکاران[1] ۲۰۰۸). در مجموع، مطالعات کمی و کیفی انجام

1. Hrisos

شده در مورد استفاده از کارپوشه توسط پزشکان و پرستاران بر این نکته تاکید دارد که پایش، حمایت و هدایت مداوم و منظم فراگیران در مرحله راه‌اندازی و اجرای کارپوشه لازم است.

- **حمایت همگنان:** در این خصوص مطالعات کمی منتشر شده است. استین و همکاران (۲۰۰۵) در مطالعه‌ای با مشارکت ۱۴۱۵ داروساز در کانادا اجازه دادند شرکت‌کنندگان تجربیات خود را در استفاده از کارپوشه با هم به اشتراک بگذارند. پس از برگزاری این جلسه، بازخوردهای شرکت‌کنندگان نشان داد که آگاهی و حمایت آن‌ها نسبت به کارپوشه افزایش یافت. در مطالعه تیواری و تنگ (۲۰۰۳) نیز پرستاران پس از مدتی در نتیجه درگیر شدن در فرایند کارپوشه به صورت خود به خود راهبرد یادگیری همکارانه را در پیش گرفتند و یکدیگر را حمایت کردند.
- **چالش‌های مربوط به وقت و هزینه:** یکی از عناصر محدود کننده استفاده از کارپوشه توسط اعضای حرفه پزشکی وقت‌گیر بودن آن با توجه به برنامه زمان‌بندی شلوغ این افراد است. در مطالعه کراس و وایت[1] (۲۰۰۴)، ۷۳ درصد پزشکان عمومی با این گزینه که «آن‌ها زمان کافی برای تکمیل کارپوشه داشتند»، مخالف و کاملاً مخالف بودند و ۷۴ درصد با گزینه «آن‌ها وقتی را که باید صرف خانواده کنند، برای تکمیل کارپوشه می‌گذارند» موافق و کاملاً موافق بودند. در هر صورت زمان صرف شده برای کارپوشه باید در مقایسه با دستاوردهای آن بررسی شود.
- **هدف کارپوشه:** اکثر کارپوشه‌ها هدف ارزیابی تکوینی را دارند و نیمی از آن‌ها با هدف ارزیابی تراکمی به کار می‌روند که ممکن است همراه یا بدون ارزیابی تکوینی باشد. مطالعات نشان داده‌اند زمانی که کارپوشه با هدف ارزیابی تراکمی به کار می‌رود، میزان همکاری استفاده‌کنندگان بیشتر است و زمانی که با این هدف به کار نرود احتمال قطع استفاده از آن وجود دارد. فراگیران رشته پزشکی معتقدند اگر کارپوشه با هدف ارزیابی تراکمی به کار نرود، برای تکمیل آن وقت صرف نمی‌کنند. کاربران برخی از مطالعات گزارش کردند که کارپوشه برای ارزیابی تکوینی مفید است بنابراین بخشی از آن باید به این منظور اختصاص یابد. در هر صورت گاهی بین نقش کارپوشه در ارزیابی (تراکمی) و یادگیری تناقض ایجاد می‌شود. به عنوان مثال، نتایج سه مطالعه کیفی و دو مطالعه مروری نشان دادند ارزیابی تراکمی ارزش یادگیری کارپوشه را کم می‌کند، محتوا را محدود می‌کند، فراگیران تقلب می‌کنند و از درج موارد ضعف و احساسات منفی خود اجتناب می‌ورزند. هرچند مطالعاتی هم وجود دارند که نشان می‌دهند بین این دو کاربرد کارپوشه در تحصیلات تکمیلی تناقضی وجود ندارد و می‌توان این دو را با هم ادغام کرد (تاچل و همکاران ۲۰۰۹).
- **اختیاری یا اجباری بودن کارپوشه:** به طور قطع، اختیاری یا اجباری بودن کارپوشه بر استفاده از آن تأثیرگذار است. به طوری‌که در دو مطالعه‌ای که تکمیل کارپوشه اختیاری بود، میزان مشارکت بسیار کم و به میزان ۲۳ درصد و ۴۳ درصد بود (تاچل و همکاران ۲۰۰۹).

بنابراین، اولین قدم در موفقیت در استقرار کارپوشه، برنامه‌ریزی و سپس حمایت در اجرا است. حمایت از طرف یک منتور مطلع در شروع بسیار اهمیت دارد. این موضوع همچنین بر وسعت استفاده از کارپوشه تاثیرگذار است. به ویژه زمانی که بازخورد اختصاصی مداوم ارائه می‌شود. البته با وجود این مسأله هم نمی‌توان از استفاده درازمدت کارپوشه مطمئن شد. از جمله عوامل مزاحم مسأله کمبود وقت است و کاربران معتقد هستند به حمایت بیشتری از طرف منتورها نیاز دارند. عوامل دیگری که در شروع و تداوم استفاده از کارپوشه تاثیرگذار هستند، می‌توان به ویژگی‌ها، نگرش، تجربه و ترجیحات یادگیری کاربران اشاره کرد. عوامل تأثیرگذار دیگری نیز وجود دارند که به صورت عینی مورد آزمایش قرار نگرفته‌اند. از جمله این عوامل می‌توان به در دسترس بودن و انعطاف‌پذیری زمان کاربران، دسترسی به کامپیوتر، ارتباط و کیفیت بخش فردی کارپوشه اشاره کرد. مطالعه‌ای که ویژگی‌های کارپوشه (اجزا، عملکرد، ارتباطات و اهداف آن) را در برابر چگونگی استفاده از آن بررسی کند، وجود ندارد. این در حالی است که اندازه‌گیری استفاده از کارپوشه با تغییر ویژگی‌های آن کار دشواری نیست و حتی می‌تواند به صورت گذشته‌نگر انجام شود. یکی از مواردی که بر نگرش، میزان استفاده و زمانی که

1. Cross & White

کاربران صرف کارپوشه می‌کنند، تأثیرگذار است اختیاری یا اجباری بودن کارپوشه است. جمع‌بندی مطالعات مختلف در پزشکی نشان می‌دهد که کارپوشه در اکثر مواقع به صورت اجباری اجرا شده است، بازاندیشی جزء الزامات آن بوده است و با دیگران به اشتراک گذاشته شده است و کاربرد ارزیابی با هر دو هدف تکوینی و تراکمی درنظر گرفته شده است.

منابع

1. Austin Z, Marini A, Desroches B. Use of a learning portfolio for continuous professional development: A study of pharmacists in Ontario (Canada). Pharm Educ. 2005;5(3-4):175–81.
2. Bierer SB, Dannefer EF. Does students' gender, citizenship, or verbal ability affect fairness of portfolio-based promotion decisions? Results from one medical school. Acad med. 2011;86(6):773–7.
3. Buckley S, Coleman J, Davison I, Khan KS, Zamora J, Malick S, et al. The educational effects of portfolios on undergraduate student learning: a best evidence medical education (BEME) review. BEME guide No. 11. Med Teach. 2009;31(4):282–98.
4. Burch VC, Seggie JL. Use of a structured interview to assess portfoliobased learning. Med Educ. 2008;42(9):894–900.
5. Campbell CM, Parboosingh JT, Gondocz ST, Babitskaya G, Lindsay E,De Guzman RC, Klein LM. Study of physicians' use of a software program to create a portfolio of their self-directed learning. Acad Med. 1996;71(10):S49–S51.
6. Cantillon P, Wood D. ABC of Learning and Teaching in Medicine. 2nd ed. West Sussex: John Wiely & Sons; 2010.
7. Carraccio C, Englander R. Evaluating competence using a portfolio: a literature review and web-based application to the ACGME competencies. Teach Learn Med. 2004;16(4):381–7.
8. Challis M. AMEE medical education guide no. 11 (revised): portfolio-based learning and assessment in medical education. Med Teach. 1999;21(4):370–86.
9. Cole G. The definition of "portfolio". Med Educ. 2005;39(11):1142.
10. Cross M, White P. Personal development plans: the Wessex experience. Educ Prim Care. 2004;15(2):205–12.
11. Dahllof G, Tsilingaridis G, Hindbeck H. A logbook for continuous self-assessment during 1 year in paediatric dentistry. Eur J Paediatr Dent. 2004;5:163–9.
12. Dannefer EF, Henson LC. The portfolio approach to competency-based assessment at the Cleveland Clinic Lerner College of Medicine. Acad Med. 2007;82(5):493–502.
13. Davis DA, Mazmanian PE, Fordis M, van Harrison R, Thorpe KE, Perrier L. Accuracy of physician self-assessment compared to observed measures of competence: A systematic review. JAMA. 2006;296(9):1094–102.
14. Davis MH, Ponnamperuma GG. Portfolio assessment. JVME. 2005;32(3):279–84.
15. Davis MH, Ponnamperuma GG, Ker JS. Student perceptions of a portfolio assessment process. Med Educ. 2009;43(1):89–98.
16. Dornan T, Carroll C, Parboosingh J. An electronic learning portfolio for reflective continuing professional development. Med Educ. 2002;36:767–9.

17. Dornan T, Maredia N, Hosie L, Lee C, Stopford A. A web-based presentation of an undergraduate clinical skills curriculum. Med Educ. 2003;37(6):500–8.
18. Dreissen EW, Overjeem K, van Tartwijk J, van der Vleuten CPM, Muijtjens AMM. Validity of portfolio assessment: which qualities determine ratings? Med Educ. 2006;40(9):862–6.
19. Driessen EW, van der Vleuten C, Shuwirth L, van Tartwijk J, Vermunt J. The use of qualitative research criteria for portfolio assessment as an alternative to reliability evaluation: a case study. Med Educ. 2005;39(2):214–20.
20. Driessen EW, Van Tartwijk J, Overeem K, Vermunt JD, Van Der Vleuten CP. Conditions for successful reflective use of portfolios in undergraduate medical education. Med educ. 2005;39(12):1230–5.
21. Driessen EW, Van Tartwijk J, van der Vleuten C, Wass V. Portfolios in medical education: why do they meet with mixed success? A systematic review. Med Educ. 2007;41(12):1224–33.
22. Driessen EW, van Tartwijk J, Vermunt JD, van der Vleuten CPM. Use of portfolios in early undergraduate medical training. Med Teach. 2003;25(1):14–9.
23. Eva KW, Regehr G: Self-assessments in the health professions: A reformulation and research agenda. Acad Med. 2005;80(10):S46–S54.
24. Finlay IG, Maughan TS, Webster DJ. T. A randomized controlled study of portfolio learning in undergraduate cancer education. Med Educ. 1998;32:172–6.
25. Friedman BDM, Davis MH, Harden RM, Howie PW, Ker J, Pippard M. AMEE edical education guide no. 24: portfolios as a method of student assessment. Med Teach. 2001;23(6):535–51.
26. Holmboe ES, Hawkins RE. Practical guide to the evaluation of clinical competence. Philadelphia: Mosby/Elsevier; 2008.
27. Grant A, Kinnersley P, Metcalf E, Pill R, Houston H. Students' views of reflective learning techniques: an efficacy study at a UK medical school. Med Educ. 2006;40(4):379–88.
28. Goldie J, Dowie A, Cotton P, Morrison J. Teaching professionalism in the early years of a medical curriculum: a qualitative study. Med educ. 2007;41(6):610–7.
29. Gómez SS, Ostos EM, Solano JM, Salado TF. An electronic portfolio for quantitative assessment of surgical skills in undergraduate medical education. BMC Med Educ. 2013;13(1):65.
30. Gordon J. Assessing students' personal and professional development using portfolios and interviews. Med Educ. 2003;37(4):335–40.
31. Haffling AC, Beckman A, Pahlmblad A, Edgren, G. Students› reflections in a portfolio pilot: Highlighting professional issues. Med Teach. 2010;32(12):e532–e40.
32. Hays RB. Reflecting on learning portfolios. Med Educ. 2004;38(8):801–3.
33. Howe A, Barrett A, Leinster S. How medical students demonstrate their professionalism when reflecting on experience. Med Educ. 2009;43(10):942–51.
34. Hrisos S, Illing J, Burk J. Portfolio learning for foundation doctors: early feedback on its use in the clinical workplace. Med Educ. 2008;42(2):214–23.

35. Jarvis RM, O'Sullivan PS, McClain T, Clardy JA. Can one portfolio measure the six ACGME general competencies? Acad Psychiatry. 2004;28(3):190–6.
36. Jasper MA. The potential of the professional portfolio for nursing, J Clin Nurs. 1995;4(4):249–55.
37. Jasper MA, Fulton J. Marking criteria for assessing practice-based portfolios at masters' level. Nurse Educ Today. 2005;25(5):377–89.
38. Keim KS, Gates GE, Johnson CA. Dietetics professionals have a positive perception of professional development. J Am Diet Assoc. 2001;101(7):820–4.
39. Kjaer NK, Maagaard R, Wied S. Using an online portfolio in postgraduate training. Med Teach. 2006;28(8):708–12.
40. Kjaer NK, Maagaard R, Wied S. Designing an online portfolio for postgraduate training of GPs in Denmark: Stepwise development in collaboration with users. Scand J Prim Health. 2008;26(2):70–3.
41. Lonka K, Slotte V, Halttunen M, Kurki T, Tiitinen A, Vaara L, Paavonen J. Portfolios as a learning tool in obstetrics and gynaecology undergraduate training. Med Educ. 2001;35(12):1125–30.
42. Lynch DC. Assessing practice-based learning and improvement. Teach Learn Med. 2004;16(1):85–92.
43. Maidment YG, Rennie JS, Thomas M. Revalidation of general dental practitioners in Scotland: the results of a pilot study. Part 1 – Feasibility of operation. Brit Den J. 2006a;200(7):399–402.
44. Maidment YG, Rennie JS, Thomas M. 'Revalidation of general dental practitioners in Scotland: the results of a pilot study. Part 2 – acceptability to practitioners'. Brit Dent J. 2006b;200(8):455–8.
45. Mathers NJ, Challis MC, Howe AC, Field NJ. Portfolios in continuing medical education– Effective and efficient? Med Educ. 1999;33(7):521–30.
46. McCready T. Portfolios and the assessment of competence in nursing: a literature review. Int J Nurs Stud. 2007;44(1):143–51.
47. McMullan M, Endacott R, Gray MA, Jasper M, Miller CML, Scholes J, Webb C. Portfolios and assessment of competence: a review of the literature. J Adv Nurs. 2003;41(3):283–294.
48. Melville C, Rees M, Brookfield D, Anderson J. Portfolios for assessment of paediatric specialist registrars. Med Educ. 2004;38(10):1117–25.
49.
50. Moores A, Parks M. Twelve tips for introducing E-Portfolios with undergraduate students. Med Teach. 2010;32(1):46–9.
51.
52. Nagler A, Andolsek K, Padmore JS. The unintended consequences of portfolios in graduate medical education. Acad Med. 2009;84(11):1522–6.
53. O'Sullivan AJ, Harris P, Hughes CS, Toohey S, Balasooriya C, Velan G, Kumar RK, Mcneil HP. Linking assessment to undergraduate student capabilities through portfolio examination. Assess Eval

High Educ. 2012;37(3):379–91.

54. O›sullivan AJ, Howe AC, Miles S, Harris P, Hughes CS, Jones P, et al. Does a summative portfolio foster the development of capabilities such as reflective practice and understanding ethics? An evaluation from two medical schools. Med Teach. 2012;34(1):e21–e8.
55. O'Sullivan PS. Demonstration of portfolios to assess competency of residents. Adv Health Sci Educ. 2004;9:309–23.
56. Pearson DJ, Heywood P. Portfolio use in general practice vocational training: a survey of GP registrars. Med Educ. 2004;38(1):87–95.
57. Pitts J, Coles C, Thomas P. Educational portfolios in the assessment of general practice trainers: reliability of assessors. Med Educ. 1999;33(7):515–20.
58. Pitts J, Coles C, Thomas P. Enhancing reliability in portfolio assessment: shaping the portfolio. Med Teach. 2001;23(4):351–6.
59. Pitts J, Coles C, Thomas P, Smith F. Enhancing reliability in portfolio assessment: discussions between assessors. Med Teach. 2002;24(2):197–201.
60. Rees C. 'Portfolio' definitions: do we need a wider debate?. Med Educ. 2005;39(11):1142.
61. Rees C. The use (and abuse) of the term ‹portfolio›. Med Educ. 2005;39(4):436–6.
62. Snadden D, Thomas M. AMEE Guide No 11: The use of portfolio learning in medical education. Med Teach. 1998;20(3):192–9.
63. Sturmberg JP, Farmer L. Educating capable doctors-A portfolio approach. Linking learning and assessment. Med Teach. 2009;31(3):e85-e9.
64. Swanwick T. Uderstanding Medical Education: Evidence, Theory and Practice. West Sussex: John Wiely & Sons; 2010.
65. Tiwari A, Tang C. From process to outcome: the effect of portfolio assessment on student learning. (Qualitative and quantitative research in Hong Kong). Nurse Educ Today. 2003(4);23:269–77.
66. Tochel C, Haig A, Cadzow A, Beggs K, Colthart I, Peacock H. The effectiveness of portfolios for postgraduate assessment and education: BEME Guide No 12. Med Teach. 2009;31(4):320–39.
67. Tulinius C, Hølge-Hazelton B. Continuing professional development for general practitioners: Supporting the development of professionalism. Med Educ. 2010;44(4):412–20.
68. Van Tartwijk J, Driessen EW: Portfolios for assessment and learning: AMEE medical education guide no. 45. Med Teach. 2009;31(9):790–801.
69. Wearne S, Dornan T, Teunissen PW, Skinner T. General practitioners as supervisors in postgraduate clinical education: an integrative review. Med Educ. 2012; 46(12):1161–73.
70. Webb TP, Aprahamian C, Weigelt JA, Brasel KJ. The surgical learning and instructional portfolio (SLIP) as a self-assessment educational tool demonstrating practice-based learning. Curr Surg. 2006;63(6):444–7.
71. Webb C, Endacott R, Gray M, Jasper M, Miller C, Mcmullan M, et al. Models of portfolios. Med

Educ. 2002;36(10):897–8.

72. Wilkinson TJ, Challis M, Hobma SO, Newble DI, Parboosingh JT, Sibbald RG, Wakeford R. The use of portfolios for assessment of the competence and performance of doctors in practice. Medi Educ. 2002;36(10):918–24.

فصل ۲۶

ارزیابی ۳۶۰ درجه

ساختار ارزیابی ۳۶۰ درجه

بازخورد از منابع متعدد (MSF)، که اغلب ارزیابی ۳۶۰ درجه نامیده می‌شود، فرایندها و ابزارهای خاصی را توصیف می‌کند که هدف آن جمع‌آوری اطلاعات، ارزیابی و ارائه بازخورد در محیط کار است. به این منظور پرسشنامه‌هایی جهت جمع‌آوری اطلاعات درباره رفتارها یا سازه‌های حرفه‌ای خاص، به نمایندگی از جانب فرد مورد ارزیابی توزیع می‌شود. در این روش، قضاوت بر اساس رفتاری که مستقیماً مورد مشاهده قرار گرفته است صورت می‌گیرد با این تفاوت که این قضاوت در طول زمان و مشاهده عملکرد فرد مورد ارزیابی در مواجهه‌های متعدد شکل می‌گیرد در حالی که در روش‌های مشاهده مستقیم (مانند mini-CEX و DOPS)، قضاوت بر اساس یک مواجهه خاص صورت می‌گیرد.

مشاهده‌گران (ارزیابان) می‌توانند شامل همکاران پزشکی (همگنان، دیگر پزشکان و فراگیران)، همکاران غیر پزشک (از جمله پرستاران، داروسازان و روانشناسان)، بیماران و اعضای خانواده ایشان و خود فرد باشند. ارزیابان که از گروه‌های مختلفی تشکیل شده‌اند، دیدگاه‌ها و مشاهدات متفاوتی در مورد عملکرد پزشکان، متناسب با نحوه تعامل با ایشان، دارند. در مجموع، مشاهده‌گران باید از بین کسانی باشند که از نزدیک با فرد مورد ارزیابی کار می‌کنند تا بتوانند عملکرد روزانه و تعاملات وی را مورد مشاهد قرار دهند. هر چند، ارزیابی ۳۶۰ درجه در طیف وسیعی از آموزش پزشکی استفاده شده است، متون منتشر شده بیشتر بر ارزیابی پزشکان در حال طبابت و همچنین فراگیران مقطع دستیاری متمرکز است.

ابزار ارزیابی ۳۶۰ درجه شامل آیتم‌هایی است که می‌توانند به روش‌های مختلفی تدوین شوند. به طور معمول، آیتم‌ها عبارات کوتاه، جملات استفهامی یا طبقاتی از سازه‌ها هستند که در مقیاس لیکرت (به طور مثال، از ۱ تا ۵ یا از ۱ تا ۹) یا نمره‌دهی رفتاری (برای مثال، معمولاً دیر تا سروقت) ارزیابی می‌شوند. این روش‌ها باعث می‌شوند اطلاعات بر اساس یک رفتار خاص یا سازه‌ها یا طبقات کلی‌تر، یا به هر دو صورت به دست آید. نمی‌توان گفت مقیاس‌ها با چه تعداد یا از چه نوعی صحیح‌ترین هستند اما عواملی وجود دارند که بر انتخاب مقیاس تأثیرگذار هستند.

پرسشنامه بیمار در ارزیابی ۳۶۰ درجه، شامل اطلاعاتی در مورد تجربه بیمار با پزشک، اغلب متمرکز بر جنبه‌های ارتباطی، مراقبت از بیمار و تعهد حرفه‌ای است. ارزیابی بیمار همچنین می‌تواند شامل مشاهدات بیمار درباره همکاری بین‌حرفه‌ای و مطب‌داری (برای مثال، سیستم پاسخگویی تلفنی، زمان انتظار در مطب و ...) باشد. پرستاران و کارکنان دیگر حرف علوم پزشکی می‌توانند اطلاعاتی در مورد مراقبت از بیمار، همکاری بین حرفه‌ای، تعهد حرفه‌ای و مهارت‌های ارتباطی فراهم کنند. همکاران پزشک (به عنوان مثال، همگنان) نیز اغلب در مورد همین حوزه‌ها مورد سؤال واقع می‌شوند اما می‌توانند اطلاعاتی را درباره دانش پزشکی، مهارت‌های تکنیکی، طرفداری از حقوق افراد و استفاده از منابع فراهم کنند.

نکته قابل توجه آن است که در سیستم سنتی ارزیابی، مانند ارزیابی انتهای دوره (چرخش بالینی)، بازخورد از طرف

افرادی ارائه می‌شود که در موقعیت سلسله مراتبی یا سازمانی بالاتری از فرد مورد ارزیابی قرار دارند. در ارزیابی ۳۶۰ درجه، بازخورد از گروه وسیع‌تری از ارزیابان ارائه می‌شود که از نزدیک با دانشجو یا پزشک مورد ارزیابی کار می‌کنند.

نوع ارائه بازخورد می‌تواند متنوع باشد اما اکثر اوقات داده‌های مربوط به هر منبع (برای مثال، گروه همگنان)، به تفکیک سازه و آیتم تحلیل می‌شود. معمولاً بازخورد شامل داده‌های مقایسه‌ای بین افراد مورد ارزیابی و نیز نتایج ارزیابی منابع مختلف با خود ارزیابی است. گزارش نتایج غالباً هم شامل داده‌های گرافیکی و هم عدد و رقمی است. این رویکرد به بازخورد (اغلب به همراه منتورینگ و حمایت در تدوین برنامه عملیاتی به منظور ارتقای عملکرد)، فرد مورد ارزیابی را قادر می‌سازد تا عملکرد فردی خود را از جنبه‌های مختلف مورد بررسی قرار دهد. مانند دیگر روش‌های ارزیابی مبتنی بر محل کار، منطق زیر بنایی ارزیابی ۳۶۰ درجه در برنامه‌های آموزشی با اهداف تکوینی است؛ به این صورت که فراگیران نیازهای یادگیری خود را تشخیص دهند برای آن برنامه بریزند و آن را ارتقا دهند.

همان‌طور که در فصل اول بیان شد، در سال‌های اخیر استفاده از این روش ارزیابی با کاربرد تراکمی یا بخشی از قضاوت نهایی از عملکرد پزشکان بیشتر شده است. به عنوان نمونه در آمریکا، ارزیابی ۳۶۰ درجه یکی از اجزای تمدید پروانه طبابت توسط بورد متخصصان آمریکا و یکی از الزامات ارزیابی دستیاران توسط ACGME معرفی شده است. در هر صورت توصیه می‌شود حتی در صورت استفاده از ارزیابی ۳۶۰ درجه با اهداف تکوینی، در خصوص نحوه برخورد با پزشکان دارای ضعف عملکرد، ساز و کاری تعیین شود تا سلامت و ایمنی بیماران در معرض خطر قرار نگیرد. همچنین در صورت کاربرد با اهداف تراکمی، ارزیابی ۳۶۰ درجه نباید تنها منبع قضاوت در مورد عملکرد پزشکان باشد.

در حوزه پزشکی، برای سال‌ها ارزیابی عملکرد دانشجویان و دانش‌آموختگان توسط همکاران[1] (همگنان) مطرح بوده است. به عنوان مثال، در دانشگاه میسوری–کانزاس دانشجویان پزشکی در دوره کارآموزی زنان و زایمان و داخلی به قضاوت در مورد مهارت‌های تعهد حرفه‌ای همکلاسی‌های خود می‌پردازند. این روش ارزیابی به مدت بیش از ۳۰ سال در این دانشگاه مورد استفاده قرار گرفته است و در طول سال‌ها تکامل یافته است. در دانشگاه فلوریدا سیستم ارزیابی همگنان برای تعیین آن دسته از دانشجویان پزشکی که رفتار حرفه‌ای عالی دارند، به کار می‌رود. اطلاعات مربوط به این ارزیابی در توصیه نامه‌ای که به رئیس دانشکده برای ادامه تحصیل در مقطع دستیاری فرستاده می‌شود، درج می‌شود (سوان‌ویک ۲۰۱۰). تا قبل از سال ۲۰۰۰ میلادی، کالج پزشکان و جراحان[2] در ایالت آلبرتای کانادا، مرور همگنان در مطب پزشکان را اجرا می‌کرد که در آن دو پزشک از مطب همکار خود بازدید می‌کردند و ارزیابی مفصلی از ثبت اطلاعات پزشکی، مراقبت بالینی، امکانات مطب و رویه مطب‌داری انجام می‌دادند. البته هرچند این فرایند بسیار دقیق بود، هزینه‌بر و ناکارآمد بود به طوری که تنها ۴۰ پزشک در سال مورد ارزیابی قرار می‌گرفتند و بازخوردی به پزشکان ارائه نمی‌شد. در راستای تغییر دیدگاه‌های فلسفی در خصوص ارزیابی پزشکان، کالج مذکور تصمیم گرفت رویکرد تکوینی‌تری به ارزیابی، با تکیه بر اصول ارتقای کیفیت اختیار کند (رمزی و همکاران[3] ۱۹۹۳ و ۱۹۹۶، ونریخ و همکاران[4] ۱۹۹۳).

در سال‌های اخیر، این قضاوت‌ها به روش‌های سیستماتیک و از منابع متعدد جمع‌آوری شده و مبنای ارزیابی عملکرد افراد قرار گرفته است. به عنوان اولین تلاش در زمینه ارزیابی ۳۶۰ درجه می‌توان به برنامه مرور دستاوردهای پزشکان[5] (PAR) اشاره کرد که توسط کالج پزشکان و جراحان آلبرتا برای پزشکان خانواده طراحی و اجرا شد. کالج مذکور با همکاری دانشگاه کالگری و آلبرتا در یک فرایند نظام‌مند و مشورتی با مشارکت وسیع پزشکان، افرادی از دیگر حِرَف پزشکی و بیماران، ابزارها را طراحی و سپس در یک فرایند ارزیابی روان‌سنجی، پرسشنامه‌هایی استاندارد شامل حوزه‌هایی متنوع از عملکرد پزشکان و آیتم‌هایی مربوط به هر حوزه تدوین کرد. اخیراً کالج پزشکان و جراحان آلبرتا فرایند مشابهی را برای تدوین ابزارهای ارزیابی ۳۶۰ درجه به منظور ارزیابی جراحان، متخصصان اطفال، بیهوشی و فارغ التحصیلان پزشکی که در خارج کشور فارغ التحصیل شده اند انجام داده است.

در سال ۲۰۰۰ میلادی، کالج پزشکان و جراحان نوااسکوشیا نیز برنامه ارزیابی ۳۶۰ درجه آلبرتا را برای ارزیابی پزشکان خانواده در آن ایالت به کار برد. قبل از اجرا، روش ارزیابی به طور وسیعی مورد آزمایش قرار گرفت و پس از اجرا نیز پژوهش‌های متنوعی در

1. Peers

۲. در کانادا، کالج پزشکان و جراحان هر ایالت نهاد نظارتی مسؤول پایش عملکرد پزشکان و تضمین سلامت بیمار در آن ایالت است.

3. Ramsey et al.

4. Wenrich et al.

5. Physician Achievement Review (PAR)

مورد پیامدها، مقبولیت و قابلیت اجرای برنامه انجام شد که در بخش‌های مختلف این فصل به پژوهش‌های مذکور اشاره می‌شود.

در انگلیس، در ابتدا ابزار مرور همگنان شفیلد[1] (SPRAT) به منظور ارزیابی عملکرد پزشکان متخصص، پزشکان خانواده و دستیاران تخصصی و فوق تخصصی طراحی شد. ابزار مذکور در ارزیابی عملکرد دستیاران و متخصصان اطفال به کار رفت و بررسی‌ها نشان داد از پایایی و قابلیت اجرای مناسبی برخوردار است (آرچر و همکاران[2] ۲۰۰۵). ابزار SPRAT با انجام اصلاحاتی تحت عنوان ابزار ارزیابی کوتاه توسط همگنان[3] (mini-PAT) در ارزیابی فراگیران در برنامه پیش‌دستیاری کشور انگلیس به کار رفت.

علاوه بر این، GMC که مسؤول اعطای پروانه طبابت پزشکان در انگلیس است، از سال ۲۰۱۲ فرایند جدیدی را برای تمدید پروانه طبابت شروع کرده است که بخش مهمی از این فرایند جمع‌آوری بازخورد از بیماران و همکاران با استفاده از پرسشنامه‌های ساختارمند است. این بازخوردها در وهله اول با هدف تکوینی جمع‌آوری می‌شود و انتظار می‌رود هر پنج سال یک بار پزشکان این اطلاعات را جمع‌آوری کنند، بر بازخوردهای کسب شده بازاندیشی داشته باشند و از آن برای ارتقای حرفه‌ای در آینده در جای مناسب آن استفاده کنند (رایت و همکاران[4] ۲۰۱۲).

انواع ارزیابی ۳۶۰ درجه

یک نوع خاص از ارزیابی ۳۶۰ درجه وجود دارد که به آزمون mini-PAT معروف است. در این ارزیابی، فراگیران هشت نفر را از میان اعضای هیأت علمی، دستیاران تخصصی با تجربه، مربیان و پزشکان غیر هیأت علمی، پزشکان عمومی، پرستاران و دیگر افراد مرتبط با حرفه پزشکی انتخاب می‌کنند. فرم mini-PAT برای ارزیابی توسط بیماران و کارکنان اداری مناسب نیست. فراگیران خود ارزیاب را انتخاب می‌کنند و اسامی آنان را به مرکز اجرای ارزیابی ۳۶۰ درجه ارائه می‌دهند و هماهنگی‌های لازم از طرف مرکز آزمون با ارزیابان انجام می‌شود. فرم‌ها به همراه یک برگه راهنمای آزمون و یک پاکت تمبردار تحویل ارزیابان می‌شود. هر یک از ارزیابان انتخاب شده پرسشنامه ساختارمند مربوط به ارزیابی را دریافت می‌کنند و پس از تکمیل به مرکز آزمون تحویل می‌دهند. این اقدام باعث می‌شود فراگیر از نظرات ارزیابان مطلع نشود. هر داوطلب نیز پرسشنامه خودارزیابی را که با پرسشنامه ارزیابان یکسان است، کامل می‌کند و به مرکز آزمون تحویل می‌دهد. در شکل ۱-۲۶ نمونه‌ای از پرسشنامه mini-PAT مورد استفاده در برنامه پیش‌دستیاری آمده است. این پرسشنامه شامل ۱۶ سؤال است که حیطه‌های مورد ارزیابی در آن عبارت هستند از:

- مراقبت مناسب از بیمار
- طبابت مناسب
- یاددهی و یادگیری
- ارتباط با بیماران
- کار با همکاران
- ارزیابی کلی

در این روش قضاوت ارزیاب بر اساس عملکرد روتین فراگیر است و در مورد کیفیت و تناسب عملکرد قضاوت می‌کند. سپس پرسشنامه‌ها بررسی و مقایسه می‌شوند و به صورت فردی به داوطلب بازخورد داده می‌شود. در برنامه پیش‌دستیاری این فرایند به صورت الکترونیکی انجام می‌شود. بخشی از بازخورد به صورت نمودار است که در آن نمره خود ارزیابی، میانگین نمره ارزیابان و نمره میانگین کشوری به تفکیک هر سؤال ترسیم شده است. تمامی توصیه‌ها و پیشنهادها ارزیابان به صورت کلمه به کلمه و به صورت بدون نام درج می‌شود. داوطلب این بازخوردها را با استاد راهنمای خود به صورت حضوری مرور می‌کند و بر اساس آن برنامه عملیاتی‌اش را تدوین می‌کند. هم فراگیران و هم استادان راهنما در خصوص فرایند بازخورد و تفسیر آن آموزش می‌بینند. این فرایند دو بار در هر سال تحصیلی تکرار می‌شود.

1. Sheffield Peer Review Assessment Tool (SPRAT)
2. Archer et al.
3. Mini-peer assessment tool
4. Wright et al.

Mini-PAT (Peer Assessment Tool) – F_1 Version

لطفا با گذاشتن علامت ضربدر مقابل سؤالات فرم را کامل کنید: ☒ لطفا از قلم مشکی برای تکمیل فرم استفاده کنید.

نام خانوادگی پزشک

نام پزشک

شماره GMC (نظام پزشکی)

شما این پزشک را در موارد زیر چگونه ارزیابی می‌کنید؟	زیر حد انتظار برای F_1		مرزی	در حد انتظار برای F_1	بالاتر از حد انتظار برای F_1		بدون نمره*
	۱	۲	۳	٤	٥	٦	
مراقبت بالینی از بیمار							
۱. توانایی تشخیص مشکل بیمار	☐	☐	☐	☐	☐	☐	☐
۲. توانایی جمع‌بندی مناسب برنامه درمانی	☐	☐	☐	☐	☐	☐	☐
۳. آگاهی از محدودیت‌هایش	☐	☐	☐	☐	☐	☐	☐
٤. توانایی پاسخ به جنبه‌های روانی ناخوشی	☐	☐	☐	☐	☐	☐	☐
٥. استفاده مناسب از منابع مانند درخواست تست‌های تشخیصی	☐	☐	☐	☐	☐	☐	☐
انجام طبابت به طور مناسب							
٦. توانایی مدیریت مؤثر زمان/اولویت‌بندی	☐	☐	☐	☐	☐	☐	☐
۷. مهارت‌های تکنیکی (متناسب با طبابت روز)	☐	☐	☐	☐	☐	☐	☐
یاددهی و یادگیری، نقد و ارزیابی							
۸. اشتیاق و کارایی در آموزش همکاران	☐	☐	☐	☐	☐	☐	☐
ارتباط با بیماران							
۹. ارتباط با بیماران	☐	☐	☐	☐	☐	☐	☐
۱۰. ارتباط با همراهان یا خانواده بیمار	☐	☐	☐	☐	☐	☐	☐
۱۱. احترام به بیمار، حقوق و حریم خصوصی وی	☐	☐	☐	☐	☐	☐	☐
کار با همکاران							
۱۲. ارتباط کلامی با همکاران	☐	☐	☐	☐	☐	☐	☐
۱۳. ارتباط نوشتاری با همکاران	☐	☐	☐	☐	☐	☐	☐
١٤. توانایی تشخیص و بهادادن به سهم دیگران	☐	☐	☐	☐	☐	☐	☐
١٥. در دسترس/قابل اطمینان	☐	☐	☐	☐	☐	☐	☐
١٦. در کل، این پزشک را نسبت به فردی که دوره F_1 را با موفقیت پشت سر گذرانده چگونه ارزیابی می‌کنید؟	☐	☐	☐	☐	☐	☐	☐

آیا در مورد پاکدامنی یا سلامت این پزشک دغدغه‌ای دارید؟ بلی ☐ خیر ☐

در صورت پاسخ مثبت موارد را ذکر نمایید:

* بدون نمره: لطفا این مورد را در صورتی علامت بزنید که رفتار مورد نظر مشاهده نشده است و در نتیجه قادر به نمره‌دهی نیستید.

موارد مثبت عملکرد	موارد پیشنهادی برای ارتقای عملکرد

در صورت لزوم لطفا پیشنهادتان را در برگه دیگری مرقوم فرمایید

شکل ۱-۲۶: نمونه‌ای از پرسشنامه mini-PAT مورد استفاده در برنامه پیش‌دستیاری انگلیس (برای اطلاعات بیشتر در مورد این فرم و جزئیات توانمندی‌های مورد انتظار برای F1 و F2 به سایت www.mmc.nhs.uk مراجعه کنید.)

جنسیت شما	☐ مذکر	☐ مؤنث
نژاد شما	☐ انگلیسی	☐ بنگلادش
	☐ ایرلندی	☐ دیگر نژاد آسیایی
	☐ دیگر نژاد سفید	☐ سفید و سیاه کارائیب
	☐ کارائیب	☐ سفید و سیاه آفریقایی
	☐ آفریقایی	☐ سفید و آسیایی
	☐ دیگر نژاد سیاه	☐ دیگر نژادهای دورگه
	☐ هندی	☐ چینی
	☐ پاکستانی	☐ نژادهای دیگر
شما در اکثر مواقع پزشک را در کدام موقعیت مشاهده نموده‌اید؟	☐ بخش ☐ درمانگاه ☐ بخش و درمانگاه ☐ پذیرش بیمار ☐ عرصه اجتماع ☐ آزمایشگاه/مرکز تحقیقات	☐ بخش مراقبت‌های ویژه ☐ مطب ☐ پزشک عمومی ☐ غیره (لطفا مشخص نمائید)
موقعیت شما:	☐ بخش ☐ پرستار ☐ پزشک عمومی ☐ مشاغل دیگر سلامتی	☐ بخش مراقبت‌های ☐ مطب	☐ بخش مراقبت‌های ویژه ☐ غیره (لطفا مشخص نمائید)
تجربه کاری (پرستاران و مشاغل دیگر سلامتی)	سال	مدت ارتباط کاری	ماه
آیا در مورد این روش تاکنون آموزش دیده‌اید؟	چهره به چهره ☐	مطالعه دستورالعمل‌ها ☐	اینترنت/CD ☐
امضای ارزیاب	تاریخ	زمان صرف شده برای تکمیل فرم (دقیقه)	
نام خانوادگی ارزیاب			
شماره نظام پزشکی ارزیاب (پزشکان)			

ادامه شکل ۱-۲۶

مزایا و محدودیت‌های ارزیابی ۳۶۰ درجه

مزایای ارزیابی ۳۶۰ درجه

- ارزیابی ۳۶۰ درجه اطلاعات منحصر به فردی راجع به عملکرد فعلی و احتمالاً آینده افراد در اختیار خود فرد، استادان و مسؤولان دوره آموزشی قرار می‌دهد.
- عملکرد واقعی فراگیر و نه توانایی بالقوه انجام کار توسط وی مورد ارزیابی قرار می‌گیرد.
- توانایی ارزیابی توانمندی‌هایی مانند تعهد حرفه‌ای و مهارت‌های ارتباطی که با روش‌های سنتی قابل ارزیابی نیستند را دارد.
- در آن از ارزیابان متعدد و موقعیت‌های متنوع استفاده می‌شود که علاوه بر ارتقای روایی آزمون موجب افزایش احتمال پذیرش نتایج توسط فراگیر می‌شود.
- موجب ارتقای مهارت‌های خودآگاهی و بازاندیشی در ارزیابی شوندگان می‌شود و زمینه را برای ایجاد تغییر در ایشان فراهم می‌آورد.
- این روش نسبتاً ارزان و انعطاف‌پذیر است و از قابلیت اجرای مناسبی برخوردار است.

- امکان اصلاح آن وجود دارد. به ویژه زمانی که ارزیابی ۳۶۰ درجه با هدف تکوینی برگزار می‌شود و از حساسیت کمتری برخوردار است، به راحتی می‌توان اصلاحاتی را در محتوای ابزار انجام داد.
- امکان توزیع آن از طریق رسانه‌های مختلف وجود دارد. البته هرچند این روش ارزیابی می‌تواند از طریق رسانه‌های مختلفی انجام شود، باید به تناسب ابزار با مخاطبان توجه داشت. به عنوان مثال، پرسشنامه کاغذی یا تلفنی برای بیمار بر روش‌های کامپیوتری ارجح است.
- طراحی ابزار در ارزیابی ۳۶۰ درجه در یک فرایند گروهی با مشارکت مسؤولان، برنامه‌ریزان، مشاهده‌گران و ارزیابی شوندگان شکل می‌گیرد. بحث و تبادل نظر بین گروه‌های مذکور به منظور تدوین آیتم‌های ابزارها منجر به شفاف شدن قوانین موجود در فرهنگ و محیط مؤسسه و نیز ارزش‌ها و انتظارات حرفه‌ای می‌شود.

محدودیت‌های ارزیابی ۳۶۰ درجه

- کیفیت فرایند جمع‌آوری اطلاعات متغیر است. ماهیت غیراستاندارد ارزیابی ۳۶۰ درجه چالش‌هایی را در تفسیر داده‌ها و تحلیل پایایی آن ایجاد می‌کند. گاهی استاندارد نبودن ساختار آزمون و تغییر جزئی در روند اجرا منجر به تغییرات بزرگ در اطلاعات جمع‌آوری شده شود. به عنوان مثال، عدم اعتماد به این‌که فقط اطلاعات جنبه تکوینی دارند، باعث می‌شود مشاهده‌گران با ارفاق نمره‌دهی کنند یا توصیه‌های کمتری در بخش پیشنهادها وارد کنند. بدیهی است که آموزش مشاهده‌گران در خصوص وظیفه مشاهده، ابزارها و مفهوم آیتم‌ها نقش اساسی در موفقیت ارزیابی ۳۶۰ درجه دارد. هرچند مطالعات بسیار ناچیزی در آموزش پزشکی در مورد اثرات آموزش ارزیابان وجود دارد، مطالعات خارج از حوزه آموزش پزشکی اثرات سودمندی را گزارش کرده است. تدوین دستورالعمل‌های اجرای ارزیابی ۳۶۰ درجه، توزیع و اطلاع رسانی آن و در نهایت آموزش آن ضروری است.
- در برخی از مؤسسات یا شرایط، عدم موفقیت گزارش شده است. ارزیابی ۳۶۰ درجه ممکن است در برخی محیط‌ها موفقیت‌آمیز نباشد. در وهله اول، این روش ارزیابی نیازمند فرهنگ سازمانی است که اعتماد در آن رسوخ کرده است و مشارکت جدی در بازخورد را از انتظارات حرفه‌ای و نه یک بار اضافی تحمیل شده از طرف مدیران می‌بینند. با وجود این، دغدغه‌های مشروع افراد باید کاملاً درک شود. به عنوان مثال، ضروری است اطلاعات ارزیابان محرمانه باقی بماند، با موارد استفاده نابه‌جا از اطلاعات برخورد قانونی صورت پذیرد و روند اجرای ارزیابی ۳۶۰ درجه و ارائه اطلاعات به گونه‌ای باشد که منجر به مخدوش شدن روابط حرفه‌ای نشود. علاوه بر فرهنگ سازمانی قابل اعتماد، مشارکت و حمایت جدی افراد کلیدی، تصمیم‌گیرندگان اصلی و رهبران مؤسسه، جهت پذیرش و مشارکت وسیع روش ارزیابی در سطح سازمان، و اطمینان از هم‌راستا بودن اطلاعات با اهداف و دیگر مداخلات در سازمان ضروری است. حتی در صورت فرهنگ سازمانی مناسب و جلب مشارکت مدیران کلیدی، در برخی از محیط‌های کاری امکان اجرای ارزیابی ۳۶۰ درجه وجود ندارد. به عنوان مثال، این روش ارزیابی بر دریافت بازخورد از منابع متعدد استوار است. در برخی از مناطق روستایی کوچک، محیط‌هایی که فرد به صورت مجزا یا در ارتباط با افراد بسیار کمی کار می‌کند، تعداد مشاهده‌گران ممکن است برای تولید نتایج پایا و باثبات کافی نباشد. در برنامه‌های دستیاری کوچک نیز اگر ساز و کاری برای مشارکت پرستاران و مثلاً دانشجویان پزشکی وجود نداشته باشد، ممکن است اجرای آن مشکل باشد.
- ارائه بازخورد سلیقه‌ای است. برخلاف ارزیابی دانش و مهارت، بازخوردهای متمرکز بر ارزش‌ها و رفتار می‌تواند کاملاً سلیقه‌ای باشد که به نوبه خود اثرات فوری و طولانی مدت خواهد داشت.
- ابهام در مفهوم بازخورد وجود دارد. ارزیابان باید در مورد تفاوت ارزیابی با بازخورد توجیه شوند و راهنماهایی در مورد ارائه بازخورد دریافت کنند. همان‌طور که پیش‌تر بیان شد، بازخورد در ارزیابی ۳۶۰ درجه، به دلیل ماهیت آزمون بسیار متفاوت از بازخورد در مورد ضعف دانش یا مهارت‌های بالینی است. این پیش‌فرض وجود دارد که دانش و مهارت بالینی قابل اصلاح هستند اما بازخورد در مورد ارزش‌ها و رفتارهای فرد با ویژگی‌های شخصیتی و ذاتی وی تداخل دارد. در نتیجه عدم آمادگی و دریافت آموزش توسط مشاهده‌گران می‌تواند نتایج ناخوشایندی به بار آورد.

- ارائه اطلاعاتی که شنیدن آن برای فرد مشکل است. گاهی نتایج ارزیابی ۳۶۰ درجه پیام‌هایی دارند که شنیدن آن برای فرد بسیار مشکل است. به عنوان مثال، همکار فرد در مقابل آیتم «این پزشک به طور مناسبی مشارکت در کارها را قبول می‌کند»، گزینه کاملاً مخالفم را انتخاب کرده است. تنها در صورتی شنیدن این پیام قابل تحمل است که در بخش پیشنهادها توضیح داده شود که تحت چه شرایطی می‌توان این ضعف را برطرف کرد.

گام‌های طراحی و اجرای آزمون ارزیابی ۳۶۰ درجه مطلوب

معرفی و شروع ارزیابی ۳۶۰ درجه در یک دانشکده کار عظیمی است که نیازمند حمایت مسؤولان و مشارکت همه سطوح سازمانی برای تداوم آن است. اولین گام ارزیابی، آمادگی سازمان برای پیاده‌سازی ارزیابی ۳۶۰ درجه است. به ویژه، توجه خاص به این موضوع لازم است که این روش ارزیابی چگونه به نیازهای سازمان پاسخ می‌دهد و چگونه ابزارها، سیاست‌ها و فرایندهای موجود را تکمیل می‌کند یا جایگزین آن می‌شود. ارزیابی ۳۶۰ درجه باید به عنوان یک برنامه با ارزش افزوده دیده شود که اطلاعاتی را جهت ارتقای افراد فراهم می‌کند. این مفهوم باید در فرهنگ سازمانی، برنامه درسی و سیستم ارزیابی تزریق شود. به این منظور لازم است مواردی در طراحی و اجرای ارزیابی ۳۶۰ درجه رعایت شود که خلاصه آنها در جدول ۱-۲۶ آمده است.

تشکیل کمیته مرکزی و تعیین تیم هدایت‌کننده ارزیابی

به طور مشخص باید یک تیم رهبری برای هدایت و پایش فرایند ارزیابی ۳۶۰ درجه تشکیل شود و با ذی‌نفعان در ارتباط دائم باشد. این تیم در صورتی موفق است که همه مشارکت‌کنندگانِ در فرایند، مانند مسؤولان، ارزیابان و ارزیابی‌شوندگان عضو آن باشند.

تعیین هدف از کاربرد ارزیابی ۳۶۰ درجه

لازم است مشخص شود کاربرد این ارزیابی تکوینی یا تراکمی است. آیا برای ارتقای فردی و پایش پیشرفت به کار می‌رود یا به منظور تصمیم‌گیری برای ارتقای فراگیر و یا فارغ‌التحصیلی. از همان ابتدا این تصمیمات باید شفاف شوند به این دلیل که بر ادامه فرایند تأثیرگذار هستند. به عنوان مثال، اگر دستیاری حس کند نتایج ارزیابی قرار است کاربرد تراکمی داشته باشد، مشاهده‌گری را انتخاب می‌کند که سخت‌گیری کمتری دارد.

تعیین سازه‌ها، حوزه‌ها و محتوای مورد ارزیابی

ارزیابی ۳۶۰ درجه شامل مجموعه‌ای از آیتم‌ها است. ابزاری انعطاف‌پذیر است و می‌تواند برای ارزیابی تقریباً هر نوع رفتاری در محیط کار استفاده شود. بنابراین یکی از مراحل مهم، تصمیم‌گیری در مورد حیطه‌ها و سازه‌های مورد ارزیابی است. ارزیابی ۳۶۰ درجه می‌تواند تعهد حرفه‌ای، رفتارهای مبتنی بر تیم، مهارت‌های ارتباطی، قضاوت بالینی یا تقریباً هر جنبه‌ای از طبابت پزشکی را مورد ارزیابی قرار دهد. اگر تصمیم بر آن است که ابزاری برای ارزیابی تعهد حرفه‌ای طراحی شود باید اطمینان حاصل شود که ابزار تنها این سازه و نه چیز دیگری را مورد ارزیابی قرار می‌دهد. حوزه‌های مورد ارزیابی و آیتم‌های زیرمجموعه آن متناسب با هر گروه از مشاهده‌گران تدوین می‌شود. همان‌طور که در بخش معرفی ابزار بیان شد، پرسشنامه بیمار شامل اطلاعاتی در مورد تجربه بیمار با پزشک، اغلب متمرکز بر جنبه‌های ارتباطی، مراقبت از بیمار، تعهد حرفه‌ای، همکاری بین حرفه‌ای و مطب‌داری است. پرستاران و کارکنان دیگر حرف علوم پزشکی می‌توانند اطلاعاتی در مورد مراقبت از بیمار، همکاری بین حرفه‌ای، تعهد حرفه‌ای و مهارت‌های ارتباطی فراهم کنند. همکاران پزشک می‌توانند علاوه بر این حوزه‌ها، اطلاعاتی درباره دانش پزشکی، مهارت‌های تکنیکی، طرفداری از حقوق افراد و استفاده از منابع فراهم کنند. ضروری است رفتارهای مورد مشاهده ارزیابی شوند.

جدول ۱-۲۶ : خلاصه مراحل طراحی و اجرای ارزیابی ۳۶۰ درجه

ردیف	عنوان	توضیح
۱	تشکیل کمیته مرکزی و تعیین تیم هدایت‌کننده ارزیابی	ضروری است یک تیم رهبری برای هدایت و پایش فرایند ارزیابی ۳۶۰ درجه تشکیل شود و با ذی‌نفعان در ارتباط دائم باشد.
۲	تعیین هدف ارزیابی ۳۶۰ درجه	ارزیابی ۳۶۰ درجه می‌تواند برای ارتقای فردی و پایش پیشرفت به کار می‌رود یا به منظور تصمیم‌گیری برای ارتقای فراگیر و یا فارغ‌التحصیلی. از همان ابتدا این تصمیمات باید شفاف شوند به این دلیل که بر ادامه فرایند تأثیرگذار هستند.
۳	تعیین سازه‌ها، حوزه‌ها و محتوای مورد ارزیابی	ارزیابی ۳۶۰ درجه شامل مجموعه‌ای از آیتم‌ها است. یکی از مراحل مهم، تصمیم‌گیری در مورد حیطه‌ها و سازه‌های مورد ارزیابی است. ارزیابی ۳۶۰ درجه می‌تواند تعهد حرفه‌ای، رفتارهای مبتنی بر تیم، مهارت‌های ارتباطی، قضاوت بالینی یا تقریبا هر جنبه‌ای از طبابت پزشکی را مورد ارزیابی قرار دهد.
۴	طراحی، اقتباس یا خریداری ابزارهای ارزیابی	ممکن است با توجه به منابع در دسترس در سازمان تصمیم گرفته شود که ابزاری جدید طراحی گردد یا از ابزارهای تهیه شده در مؤسسات دیگر استفاده شود.
۵	تعیین گروه مشاهده‌گران و ارزیابی شوندگان	تیم مسؤول استقرار ارزیابی ۳۶۰ درجه نیازمند تصمیم‌گیری در مورد مشاهده‌گران و افراد مورد ارزیابی است. تصمیم‌گیری در مورد مشاهده‌گران به این موضوع بستگی دارد که چه رفتارهایی را این گروه از مشاهده‌گران می‌توانند مورد مشاهده قرار دهند و قابلیت اجرا و مقبولیت ارزیابی پزشکان توسط این گروه به چه میزان است.
۶	تعیین تناوب توزیع ابزارها	دفعات تکمیل پرسشنامه بر اساس قابلیت اجرا و حجم کار تنظیم می‌شود.
۷	تعیین روش‌های توزیع ابزارها	ابزارها می‌توانند به صورت کاغذی، آنلاین، از طریق تلفن یا ترکیبی از این روش‌ها با توجه به شرایط و امکانات توزیع شوند.
۸	تهیه پروتکل ارائه بازخورد، منتورینگ و تدوین برنامه عملیاتی	قبل از اجرا باید در مورد شکل ارائه بازخوردها و محتوای آن تصمیم‌گیری شود. گزارش‌ها می‌تواند به صورت مقایسه عملکرد فرد با دیگران به همراه پیشنهادهای مشاهده‌گران برای بهبود عملکرد باشد. در کل فرایند، باید از بدون نام بودن و محرمانه ماندن هویت مشاهده‌گر مطمئن شد. می‌توان بازخوردها را از طریق پست الکترونیکی برای فرد ارسال کرد یا یک منتور می‌تواند به صورت فردی بازخورد را ارائه دهد. لازم است از همان ابتدا نحوه برخورد با اطلاعات و نتایج به دست آمده، چه ضعیف و چه عالی، و نحوه اعلام نتایج منفی و اجرای اقدامات اصلاحی مشخص شود.
۹	اجرای آزمایشی ابزارها	ضروری است ابزارها قبل از اجرا در گروه کوچکی از فراگیران آزمایش شوند و اطلاعات حاصل از آن‌ها از جمله دیدگاه استفاده‌کنندگان، در اجرا و استقرار کامل برنامه لحاظ شود.
۱۰	مشاهده عملکرد فرد مورد ارزیابی	به منظور تأمین اعتبار فرایند ارزیابی لازم است مشاهده‌گران (به ویژه همگنان) از میان ارزیابانی که می‌توانند رفتار مورد سؤال را مشاهده کنند، انتخاب شوند.
۱۱	ارزشیابی و اصلاح برنامه ارزیابی	از جمله جنبه‌هایی که باید مورد ارزشیابی قرار گیرد، نحوه اجرای ارزیابی، دیدگاه مشارکت‌کنندگان و کیفیت روان‌سنجی اطلاعات تولید شده است.

طراحی، اقتباس یا خریداری ابزارهای ارزیابی

ممکن است با توجه به منابع در دسترس در سازمان تصمیم گرفته شود که ابزاری جدید طراحی گردد یا از ابزارهای تهیه شده در مؤسسات دیگر استفاده شود. تصمیم‌گیری و انتخاب بین این دو گزینه نیازمند کسب اطلاع و اطمینان از روایی و پایایی ابزارهای موجود، میزان همگرایی بین نیازهای مؤسسه و ابزارهای موجود و میزان فعالیت مورد نیاز جهت تهیه ابزاری منطبق با نیازهای مؤسسه است.

تعیین گروه مشاهده‌گران و ارزیابی شوندگان

تیم مسؤول استقرار ارزیابی ۳۶۰ درجه نیازمند تصمیم‌گیری در مورد مشاهده‌گران و افراد مورد ارزیابی است. تصمیم‌گیری در مورد مشاهده‌گران به دو عامل بستگی دارد:

- چه رفتارهایی را این گروه از مشاهده‌گران می‌توانند مورد مشاهده قرار دهند؟
- قابلیت اجرا و مقبولیت ارزیابی پزشکان توسط این گروه به چه میزان است؟

به عنوان مثال، جمع‌آوری اطلاعات از بیماران بخش اورژانس، روانپزشکی یا بیماران در مراحل انتهایی زندگی دشوار است. همین‌طور بیماران با محدودیت سواد و مشکلات زبانی در این دسته قرار می‌گیرند. باید مشخص شود ارزیابی‌شوندگان خود در مورد انتخاب مشاهده‌گران تصمیم می‌گیرند یا از ابتدا مشاهده‌گران مشخص هستند. تعداد مشاهده‌گران باید با توجه به قابلیت اجرا، هدف ارزیابی (تشخیص پزشکان دارای مشکل عملکرد یا ارتقای عملکرد) و مسایل مربوط به استحکام روان‌سنجی آزمون مشخص شود.

تعیین تناوب توزیع ابزارها

دفعات تکمیل پرسشنامه بر اساس قابلیت اجرا، حجم کار و فاصله زمانیِ در دسترس بین توزیع فرم‌ها تنظیم می‌شود. مورد آخر به این دلیل است که فرصت کافی در اختیار ارزیابی شوندگان باشد تا مشکلات و ضعف‌های تعیین شده را رفع کنند.

تعیین روش‌های توزیع ابزارها

ابزارها می‌توانند به صورت کاغذی، آنلاین، از طریق تلفن یا ترکیبی از این روش‌ها توزیع شوند. هر کدام از این روش‌ها فواید و مشکلات خود را دارد و با توجه به شرایط و امکانات مانند دسترسی به اینترنت پرسرعت، طول پرسشنامه (پرسشنامه‌های کوتاه می‌توانند از طریق تلفن پرسیده شوند) و تسهیلات فن‌آوری اطلاعات در مورد آن تصمیم‌گیری می‌شود.

تهیه پروتکل ارائه بازخورد، منتورینگ و تدوین برنامه عملیاتی

قبل از اجرا باید در مورد شکل ارائه بازخوردها و محتوای آن تصمیم‌گیری شود. برخی گزارش‌ها به صورت مقایسه عملکرد فرد با دیگران ارائه می‌شود. اگر قرار است گزارش شامل پیشنهادهای مشاهده‌گر باشد و قبل از ارائه به ارزیابی‌شوندگان غربال نمی‌شود، باید از بدون نام بودن و محرمانه ماندن هویت مشاهده‌گر مطمئن شد. می‌توان بازخوردها را از طریق پست الکترونیکی برای فرد ارسال کرد یا یک منتور می‌تواند به صورت فردی بازخورد را ارائه دهد. نقش منتور یا تسهیل‌گر در راستای افزایش احتمال پذیرش بازخورد و استفاده از آن در جهت افزایش توانایی خودارزیابی و بازاندیشی و رشد فردی و حرفه‌ای ارزیابی‌شوندگان است. اگر هدف ارزیابی تکوینی است، یکی از اجزای ضروری بازخورد، تعیین هدف و تدوین برنامه عملیاتی است.

لازم است از همان ابتدا نحوه برخورد با اطلاعات و نتایج به دست آمده، چه ضعیف و چه عالی، و نحوه اعلام نتایج منفی و اجرای اقدامات اصلاحی مشخص شود. تعیین تکلیف در مورد اطلاعاتی که می‌تواند بر دیگر فرایندهای مؤسسه تأثیرگذار باشد، مانند رفتارهای غیرحرفه‌ای نامناسب اهمیت بیشتری دارد. حتی برنامه‌های ارزیابی با هدف تکوینی به دلیل مسؤولیت اخلاقی در برابر ایمنی بیمار و محیط کار باید ارائه و برخورد با چنین اطلاعاتی (مانند رفتارهای غیرحرفه‌ای نامناسب) را به روش مناسبی مدیریت کند.

اجرای آزمایشی ابزارها

ضروری است ابزارها قبل از اجرا در گروه کوچکی از فراگیران آزمایش شوند و اطلاعات حاصل از آنها از جمله دیدگاه استفاده‌کنندگان، در اجرا و استقرار کامل برنامه لحاظ شود. به عنوان مثال، GMC در فاصله سال‌های ۲۰۰۴ تا ۲۰۰۵ ابزارها را به صورت آزمایشی مورد استفاده قرار داد و سپس با انجام اصلاحاتی ابزارها را به کار برد.

مشاهده عملکرد فرد مورد ارزیابی

به منظور تأمین اعتبار فرایند ارزیابی لازم است مشاهده‌گران (به ویژه همگنان) از میان ارزیابانی که می‌توانند رفتار مورد سؤال را مشاهده کنند، انتخاب شوند. علاوه بر این، لازم است ارزیابان در مورد استفاده از گزینه «عدم مشاهده» در صورت عدم مشاهده رفتار خاص توجیه شوند.

ارزشیابی و اصلاح برنامه ارزیابی

ارزشیابی و اصلاح برنامه ارزیابی نیز مهم است. از جمله جنبه‌هایی که باید مورد ارزشیابی قرار گیرد، نحوه اجرای برنامه، دیدگاه مشارکت‌کنندگان و کیفیت روان‌سنجی اطلاعات تولید شده است.

علاوه بر موارد فوق، برقراری ارتباط ضروری‌ترین جزء برنامه است و لازم است مکرر و شفاف باشد. در کل فرایند، از ابتدای برنامه که برنامه و اهداف آن اعلام می‌شود، در مرحله آزمایشی، و در مرحله استقرار و ارزشیابی آن ارتباطات نقش بسیار مهمی دارد. ارتباط از طریق پست الکترونیکی، خبرنامه، جلسات گروه‌های آموزشی، راندهای آموزشی و فعالیت‌های علمی تسهیل می‌شود. پیام باید دارای انسجام باشد و ماهیت تکوینی یا تراکمی ارزیابی را به ویژه در خصوص پیامدهای احتمالی که به فراگیران مترتب است شفاف سازد.

سودمندی ارزیابی ۳۶۰ درجه

ارزیابی ۳۶۰ درجه یک روش مبتنی بر پرسشنامه و مشابه مطالعات پیمایشی است. بنابراین، به منظور اطمینان از روایی و پایایی ارزیابی ۳۶۰ درجه، پرسشنامه‌های مورد استفاده باید بر اساس استانداردهای مورد پذیرش در مطالعات پیمایشی طراحی و بررسی شوند. تحلیل ابزار به صورت یک کل و تحلیل عوامل آن (حیطه‌ها و زیرحیطه‌ها) مهم است. علاوه بر این، هر آیتم در پرسشنامه‌ها به این دلیل که برنامه ارتقای فردی ارزیابی شوندگان را هدایت می‌کند ارزش خاص خود را دارد، بنابراین مسأله روایی باید در سطح آیتم نیز گسترش یابد.

روایی ارزیابی ۳۶۰ درجه

به منظور تضمین روایی صوری و محتوا، مشارکت جمعیت مورد هدف و افراد متخصص در روند طراحی و تدوین ابزار ضروری است. به این دلیل که هر گروه رفتار متفاوتی را ارزیابی می‌کند، آیتم‌های هر پرسشنامه باید متناسب با هر گروه طراحی شود. ارتباط پایین گزارش شده بین گروه‌های مشاهده‌گران این موضوع را تأیید می‌کند که هر گروه اطلاعات منحصر به فردی را فراهم می‌کند (سرجانت و همکاران[1] ۲۰۰۳). این یافته‌ها همچنین روایی سازه پرسشنامه‌ها را تأیید می‌کند. علاوه بر این، این روش در ارزیابی پزشکان متخصص داخلی مشغول به طبابت توانسته است پزشکان دارای گواهینامه طبابت را از پزشکان بدون گواهینامه افتراق دهد. روایی سازه ابزار SPRAT (ابزار مرور همگنان شفیلد) به این صورت تأیید شده است که پزشکان مورد ارزیابی نمرات پایین‌تری در سؤالات مربوط به مدیریت بیماران پیچیده و مهارت‌های رهبری دریافت کردند. همچنین

1. Sargeant et al.

این ابزار توانست بین پزشکان با رتبه‌های مختلف تمایز قائل شود (آرچر و همکاران ۲۰۰۵). mini-PAT توانست بین فراگیران در مقطع پیش دستیاری سال اول و دوم به طور معنی‌داری تمایز قائل شود. بر اساس نتایج این مطالعه، تعداد فراگیران مرزی در سال اول بیشتر از سال دوم بودند (۱۹/۶ درصد در مقابل ۵/۶ درصد) (آرچر و همکاران ۲۰۰۸).

در مورد روایی معیاری ارزیابی ۳۶۰ درجه، به صورت ارتباط ابزار با ابزارهای مربوط به ارزیابی سازه‌های مشابه، مطالعات اندکی در حوزه آموزش پزشکی انجام شده است. رایت و همکاران (۲۰۱۲) به بررسی روایی همگرایی پرسشنامه‌های تهیه شده توسط GMC در ارزیابی پزشکان در انگلیس پرداختند. در این پژوهش، مشاهده‌گران شامل بیمار و همکاران پزشک و غیر پزشک، علاوه بر پرسشنامه‌های مربوط به ارزیابی ۳۶۰ درجه، پرسشنامه‌های دیگری که جنبه‌های مشابهی از عملکرد پزشکان را مورد سنجش قرار می‌داد، تکمیل کردند. به این صورت که بیماران پرسشنامه شش‌آیتمی توانمند کردن بیمار[1] (PEI) و پرسشنامه ۱۲ آیتمی مهارت‌های بین فردی[2] (DISQ) را تکمیل کردند و همکاران پرسشنامه ۱۸ آیتمی ارزیابی و بازخورد به همکاران[3] (CFET) را تکمیل کردند. ضریب همبستگی اسپیرمن بین پرسشنامه بیمار با DISQ (۰/۶۳) قوی‌تر از همبستگی با PEI (۰/۳۱) بود. پرسشنامه همکار با CFET همبستگی بسیار قوی (۰/۸۱) داشت. این یافته‌ها تأییدی بر روایی همگرایی ابزارهای ارزیابی ۳۶۰ درجه بود به این دلیل که DISQ و ارزیابی ۳۶۰ درجه ویژگی‌های مشابهی را ارزیابی می‌کنند و PEI سازه مربوط اما متفاوتی را ارزیابی می‌کند و CFET و پرسشنامه همکار ویژگی‌های مشابهی را ارزیابی می‌کنند.

به گزارش مطالعات انجام شده بین نتایج ارزیابی ۳۶۰ درجه با نمرات آزمون کتبی ارتباط وجود دارد (رمزی و همکاران ۱۹۸۹، ونریخ و همکاران ۱۹۹۳). در مطالعه دیگری رمزی و همکاران (۱۹۹۳) روی متخصصان داخلی که ۵ تا ۱۵ سال از اخذ مدرک آن‌ها می‌گذشت تمرکز کردند. دو گروه از همکاران شامل همکارانی از خود شرکت‌کنندگان در مطالعه و همچنین از بین سوپروایزرهایشان انتخاب شدند. سؤالات پرسشنامه به دو زیر حیطه، شامل مهارت‌های شناختی/ فنی و تعهد حرفه‌ای تقسیم شدند که عملکرد پزشکان در آزمون کتبی با حیطه اولی یعنی مهارت‌های شناختی ارتباط آماری معنی‌داری داشت اما با نمرات تعهد حرفه‌ای ارتباط نداشت. عملاً به نظر می‌رسد ارزیابی ۳۶۰ درجه به ارزیابی توانمندی‌هایی می‌پردازد که با روش‌های سنتی ارزیابی قابل ارزیابی نیستند.

در ارزیابی ۳۶۰ درجه، روایی از طریق شواهدی به دست می‌آید که توانایی ابزار را در اندازه‌گیری آن‌چه مورد نظر بوده است، ثابت می‌کند و هر زمان شواهد مربوط مورد تهدید جدیدی قرار گیرد، روایی ابزار دوباره بررسی می‌شود. بنابراین، بررسی روایی ارزیابی ۳۶۰ درجه یک فرایند مداوم و نه یک مطالعه مقطعی است. هر چه اهمیت و حساسیت آزمون بیشتر باشد، لزوم ارائه شواهد قوی در مورد روایی و پایایی آن بیشتر است.

همان‌طور که در ابتدای این بخش اشاره شد به منظور اطمینان از روایی ارزیابی ۳۶۰ درجه، پرسشنامه‌های مورد استفاده باید بر اساس استانداردهای مورد پذیرش در مطالعات پیمایشی طراحی و بررسی شود. در این راستا، هولمبو و هاوکینز (۲۰۰۸) معیارهایی را به منظور بررسی و تضمین روایی ابزار ارزیابی ۳۶۰ درجه ارائه نمودند: در اولین مرحله، ضروری است مطمئن شویم ابزارها آنچه را که مورد نظر است ارزیابی می‌کنند. بنابراین، ابتدا باید هدف ابزار شفاف شود. سپس بر اساس آن سازه‌های مورد ارزیابی تعیین شوند و در نهایت آیتم‌ها بر اساس سازه مورد نظر و مفروضات نظری زیربنایی آن تدوین شوند. به این دلیل که ارزیابی ۳۶۰ درجه مربوط به افراد و قضاوت‌های ایشان در مورد خود و دیگران است نظرخواهی از استفاده‌کنندگان نهایی ابزار از طریق گروه متمرکز یا پرسشنامه در مرحله طراحی ابزار بسیار مهم است. این اقدامات باعث می‌شود از همگونی اهداف ارزیابی و استنباط به عمل آمده از توانمندی و رفتار مطمئن شویم. این انطباق باید در مسیر طراحی ابزار به دست آید. علاوه بر این، ممکن است آیتم‌ها بر اساس بازخوردهای استفاده‌کنندگان پس از استفاده طولانی‌مدت از ابزار مورد اصلاح واقع شود.

1. Patient Enablement Instrument
2. Doctors' Interpersonal Skills Questionnaire
3. Colleague Feedback Evaluation Tool (CEFT)

سرجانت و همکاران ۲۰۰۳

۴۴ درصد همکاران پزشک و ۲۴ درصد همکاران غیر پزشک در ارزشیابی که پس از اجرای آزمایشی برنامه ارزیابی ۳۶۰ درجه به عمل آمد، اذعان داشتند سؤالات برای ارزیابی مشکل بود به این دلیل که نوع ارتباط آن‌ها با پزشک مورد ارزیابی برای پاسخ به سؤالات کافی نبود. ۳۰ درصد همکاران غیر پزشک قادر به ارزیابی آیتم «از امکانات عمومی برای بیماری‌های روانشناختی استفاده می‌کند» نبودند.

هالو و همکاران [1] ۱۹۹۹

در بررسی برنامه ارزیابی ۳۶۰ درجه آلبرتا حدود ۱۱ درصد مشاهده‌گران قادر به پاسخگویی به آیتم‌ها نبودند. این‌گونه ارزشیابی‌ها از روش ارزیابی می‌تواند در اصلاح پرسشنامه‌ها مفید باشد.

1. Hall et al

در مجموع برای اطمینان از روایی ابزار ارزیابی ۳۶۰ درجه لازم است در مرحله اول شواهد زیر فراهم شود:

- اطلاعات جمع‌آوری شده با چه کاربردی مورد استفاده قرار می‌گیرد؟
- ابزار چگونه و توسط چه کسی تولید می‌شود؟
- چه حوزه‌ها یا سازه‌هایی مورد ارزیابی قرار می‌گیرد؟
- به منظور حمایت از فرایند طراحی کدام متون مورد استناد قرار گرفته است؟
- تدوین‌کنندگان چگونه محتوا و شکل ابزار و بازخوردهای ارائه شده توسط آن را ارزیابی می‌کنند؟
- کدام گروه از متخصصان و استفاده‌کنندگان نهایی مورد مشاوره قرار گرفتند. بازخورد آن‌ها چگونه مورد استفاده قرار گرفت؟
- آیا بلوپرینتی برای نمونه‌گیری یا پوشش محتوا تهیه شده است؟
- نحوه آموزش مشاهده‌گران چگونه بوده است؟ آیا همگی مفهوم آیتم‌ها را درک کرده‌اند؟
- چه وسایل ارتباطی به منظور اشاعه اهداف ارزیابی مورد استفاده قرار گرفته است؟
- وقتی داده‌های تجربی مانند آن‌چه در بالا به آن اشاره شد، در مورد روایی ابزار تأمین شد، تمرکز به دیگر جنبه‌های ابزار مانند یکپارچگی و تأثیر ناشی از استفاده از آن منتقل می‌شود. نمونه‌ای از سؤالات مرحله دوم بررسی روایی ابزار ۳۶۰ درجه عبارت هستند از:
- نرخ پاسخ‌دهی به چه میزان است؟ آیا امکان به کارگیری تعداد کافی مشاهده‌گر (قابل پذیرش توسط ارزیابی شوندگان) جهت ارائه بازخورد وجود داشته است؟
- طیف نمرات به چه صورت است؟ نمرات توزیع نرمال دارند یا با چولگی به چپ یا راست همراه هستند؟
- میزان ارتباط با ابزارهای دیگر (به عنوان مثال، ارزیابی ۳۶۰ درجه تعهد حرفه‌ای یا مهارت‌های ارتباطی یا OSCE) به چه صورت است؟
- چه عواملی بر نمرات تأثیرگذار هستند؟ آیا تغییرپذیری و واریانس نمرات به دلیل جنس، سال‌های تحصیل، سال‌های طبابت، نژاد یا آشنایی بین مشاهده‌گر و ارزیابی‌شونده نیست؟
- اگر تحلیل عاملی ابزار انجام شده است، آیا سازه‌ها مشخص شده‌اند؟
- بازخوردها به چه نحوی توسط ارزیابی شوندگان، مسؤولان برنامه یا دانشکده استفاده شده است؟ آیا آن‌ها اعتبار داده‌ها را باور دارند؟ آیا تغییر رفتاری بر اساس بازخوردها صورت گرفته است؟ آیا این تغییر رفتار موجب ارتقای مراقبت از بیمار شده است؟
- آیا ارزیابی علاوه بر نتایج مطلوب، عواقب نامطلوبی هم به دنبال داشته است؟

در این بخش به نمونه‌هایی از نتایج مطالعاتی که به بررسی روایی ارزیابی ۳۶۰ درجه در قالب چارچوب فوق پرداخته‌اند،

به نقل از هولمبو و هاوکینز (۲۰۰۸) اشاره می‌شود:

- نرخ پاسخ‌دهی ابزارهای ارزیابی ۳۶۰ درجه بالا گزارش شده است.
- در دوره دستیاری و پزشکی عمومی، اطلاعات فراهم شده با روش ارزیابی ۳۶۰ درجه به عنوان بخشی از ارزیابی کلی دانشجو مورد استفاده قرار گرفته است.
- برای پزشکان در حال طبابت معمولاً ارزیابی تحت حمایت یک نهاد نظارت‌کننده انجام می‌شود و مشارکت در آن اجباری است (به عنوان مثال، برنامه مرور دستاوردهای پزشکان در آلبرتا).
- بورد طب داخلی آمریکا، ارزیابی ۳۶۰ درجه را به عنوان بخش انتخابی در تمدید مجوز اعلام کرده است و افرادی که آن را انتخاب می‌کنند، منافع ویژه‌ای می‌برند.
- معمولاً توزیع نمرات برای پزشکان در حال طبابت با چولگی به سمت چپ همراه است. به نظر می‌رسد زمانی که هدف برنامه (مانند کاربرد ارزیابی ۳۶۰ درجه در برنامه پیش‌دستیاری) شناسایی دستیاران ضعیف است، توزیع نمرات چولگی کمتری دارد. در مجموع، مطالعات نشان می‌دهند نتایج ارزیابی ۳۶۰ درجه به سمت ارزیابی مثبت عملکرد پزشکان چولگی دارد و در اکثر موارد میانگین نمرات بیش‌تر از ۴ (از نمره کل ۵) گزارش شده است. این احتمال وجود دارد که تغییر مقیاس نمره‌دهی یا آموزش و توجیه ارزیابان در خصوص اهداف و کاربرد پرسشنامه‌ها چولگی توزیع نمرات را کاهش دهد، هر چند این موضوع نیاز به پژوهش‌های بیشتر دارد.
- ارتباط بین نمرات ارزیابی ۳۶۰ درجه و دیگر ابزارها می‌تواند به عنوان شواهدی از روایی معیاری استفاده شود. این نوع تحلیل می‌تواند منجر به حذف ابزار یا استقرار آن شود به این دلیل که حوزه‌هایی را مورد ارزیابی قرار می‌دهد که با ابزارهای قبلی قابل ارزیابی یا غیر قابل ارزیابی است. علاوه بر این ارتباط بین داده‌های گروه‌های مشاهده‌گر متفاوت نیز بررسی می‌شود تا مشخص شود آیا انواع فرم‌ها اطلاعات مشابه یا تکمیل‌کننده فراهم می‌کنند. مطالعات در حوزه آموزش پزشکی نشان داده است نمره‌دهی خود فرد با دیگر مشاهده‌گران ارتباط ضعیف و غیرمعنی‌داری (برای مثال، $r < ۰/۲۵$) دارد، اما ارتباط بین گروه‌های مختلف مشاهده‌گر معمولاً متوسط (۰/۵۰ تا ۰/۲۵ =r) گزارش شده است. این داده‌ها شواهدی از روایی همگرا و واگرا هر دو فراهم می‌کند و اهمیت جمع‌آوری اطلاعات از گروه‌های مختلف را به طور جداگانه تأیید می‌کند.
- در ارزیابی ۳۶۰ درجه، رگرسیون خطی به منظور یافتن اهمیت عواملی که بر تغییر پذیری نمرات تأثیرگذارند به کار می‌رود. یکی از عواملی احتمالی می‌تواند انتخاب مشاهده‌گران توسط فرد ارزیابی‌شونده باشد. در مورد بیماران، مسأله با انتخاب تصادفی بیماران رفع شده است اما انتخاب همکاران، به ویژه در ارزیابی پزشکان در حال طبابت که در مناطق جغرافیایی وسیعی پراکنده هستند، اغلب توسط پزشک ارزیابی‌شونده صورت می‌گیرد. با وجود این نگرانی‌ها، مطالعه رمزی و همکاران (۱۹۹۳) نشان داد مشاهده‌گرانی که توسط ارزیابی‌شونده انتخاب شدند، ارزیابی متفاوت معنی‌داری از آن‌هایی که توسط نهاد یا فرد سومی انتخاب شدند، گزارش نکردند. به طور یقین این یافته نیازمند بررسی بیشتر در برنامه‌های ارزیابی ۳۶۰ درجه در محیط‌های مختلف است. از دیگر عوامل احتمالی تأثیرگذار، آشنایی بین مشاهده‌گر و ارزیابی شونده بیشتر مورد بررسی قرار گرفته است. برخی مطالعات این مسأله را مسؤول کمتر از ۱۰ تا ۱۵ درصد تغییرپذیری نمرات می‌دانند (سرجانت و همکاران ۲۰۰۳).

مطالعات نشان داده‌اند زمانی که بیماران پزشکان یا سیستم مراقبت سلامتی را ارزیابی می‌کنند، ویژگی‌های بیماران (مثل سن و نژاد) و نحوه توزیع پرسشنامه (مثل توزیع از طریق پست، تلفن یا در زمان ترخیص) و استفاده از نماینده به منظور تکمیل پرسشنامه بر نمره‌دهی تأثیر می‌گذارد. بر اساس نتایج مطالعه رایت و همکاران (۲۰۱۲)، ۵ متغیر تأثیرگذار بر پاسخ‌های بیمار که موجب ارزیابی مطلوب عملکرد پزشک می‌شد عبارت بودند از: بیمارانی که ویزیت پزشک را بسیار مهم دانستند، بیمارانی که پزشک معمول خود را ملاقات کرده بودند، پرسشنامه‌هایی که در زمان ترخیص بیمار و نه به

صورت ارسال پستی تکمیل شدند، بیماران سفید پوست (نژاد) و بیماران مسن‌تر. جنس بیماران و نوع تکمیل فرم (خود بیمار یا نماینده بیمار) تأثیری بر پاسخ‌های بیماران نداشتند. همچنین بر اساس نتایج مطالعه مذکور، ۳ متغیر بر پاسخ‌های همکاران تأثیرگذار بودند. مسؤولان، مدیران و همکاران غیر پزشک نمرات بالاتری نسبت به همگنان ارائه دادند، پزشکانی که تماس‌های بیشتری با پزشک مورد ارزیابی داشتند نیز نمرات بالاتری دادند و نهایتاً روش جمع‌آوری پرسشنامه‌ها با پاسخ‌های همکاران به یک سؤال پرسشنامه ارتباط داشت. سن، جنس، نژاد و تأخر یا تقدم ارتباط با پزشک، تأثیری بر پاسخ‌های همکاران نداشت. در مجموع، به منظور تقویت ارزیابی ۳۶۰ درجه لازم است تأثیر پدیده‌هایی مانند وضعیت نیمه وقت یا تمام وقت بودن هیأت علمی، نوع بیمارستان (آموزشی در مقابل عمومی)، وضعیت سلامتی بیمار، سن یا جنس بیمار، جنس پزشک یا تعداد سال‌های آموزش و دیگر متغیرها بر تغییرپذیری نمرات تعیین شود.

- در مورد دیدگاه استفاده‌کنندگان در مورد ابزار، اکثر مشاهده‌گران و ارزیابان احساس می‌کنند ابزار ۳۶۰ درجه برای ارزیابی مهم است. پزشکان با داده‌های فراهم شده به این روش احساس خوبی دارند. هرچند سؤالات باید با گروه مشاهده‌گر متناسب باشد. به طور مثال، اگر از پرستاران در مورد مهارت‌های ارتباطی پزشکان سؤال شود، آن‌ها مشکلی نخواهند داشت اما در صورتی که در مورد مهارت‌های بالینی ایشان پرسیده شود، برای پزشکان قابل پذیرش نخواهد بود.

نتایج مطالعات در خصوص ارتباط بین آشنا بودن مشاهده‌گران و پزشکان

نتایج مطالعاتی که به بررسی ارتباط بین آشنا بودن مشاهده‌گران و پزشکان مورد ارزیابی پرداخته‌اند، قطعی نیست. در برخی از مطالعات هیچ ارتباطی مشاهده نشد و در برخی دیگر ارتباط معنی‌دار و ضعیف تا متوسطی گزارش شد (رمزی و همکاران ۱۹۹۳، هال و همکاران ۱۹۹۹ و لاک‌یر[1] ۲۰۰۳). مطالعه سرجانت و همکاران نشان داد تأثیر مثبت معنی‌داری بین میزان آشنایی مشاهده‌گران (همگنان و همکاران پزشک) با میانگین نمرات هر دو گروه وجود داشت ($p < ۰/۰۰۱$). به عنوان مثال، نمرات ارزیابانی که پزشک مورد ارزیابی را «خیلی خوب» یا «خوب» می‌شناختند، به طور معنی‌داری ($p < ۰/۰۰۱$) بالاتر از نمرات آن‌هایی که «تا حدودی» پزشک را می‌شناختند، بود و نمرات ارزیابی که پزشک مورد ارزیابی را «خیلی خوب» می‌شناختند، به طور معنی‌داری ($p < ۰/۰۰۱$) بالاتر از نمرات آن‌هایی که پزشک را «خوب» می‌شناختند، بود. آشنا بودن پزشک مورد ارزیابی و مشاهده‌گران (همگنان) مسؤول ۱۲/۷ درصد واریانس نمرات متوسط آزمون و مشاهده‌گران (همکاران غیر پزشک) مسؤول ۱۱ درصد واریانس نمرات متوسط آزمون بود.

1. Lockyer

پایایی ارزیابی ۳۶۰ درجه

پس از این‌که ثابت شد ابزار آن‌چیزی را که مورد نظر است اندازه می‌گیرد، تعیین پایایی آزمون اهمیت دارد. ثبات درونی ابزار با استفاده از آلفای کرونباخ بیش از ۰/۹ گزارش شده است (وایولت و همکاران[1] ۲۰۰۳ و ۲۰۰۶ و رایت و همکاران ۲۰۱۲). تعداد پژوهش‌هایی که به بررسی پایایی ابزارها به صورت آزمون–آزمون مجدد[2] می‌پردازند، بسیار محدود است. رایت و همکاران (۲۰۱۲) به بررسی پایایی پرسشنامه‌های تهیه شده توسط GMC در ارزیابی پزشکان در انگلیس پرداختند. ارزیابان شامل بیمار و همکاران پزشک و غیر پزشک، به فاصله دو هفته مجدداً پرسشنامه‌ها را تکمیل کردند. پایایی آزمون–آزمون مجدد با محاسبه ضریب ICCبا فاصله اطمینان ۹۵ درصد محاسبه شد. این ضریب برای نمرات کلی بیماران و همکاران به ترتیب (۰/۸۷–۰/۷۹) ۰/۸۳ و (۰/۸۸–۰/۸۰) ۰/۸۵ بود. ضریب ICC در مورد آیتم‌هایی که مهارت‌های بالینی پزشک را ارزیابی می‌کردند، بالاتر و برای آیتم‌هایی که مهارت‌های تعهد حرفه‌ای پزشک را ارزیابی می‌کردند، پایین‌تر گزارش شد.

تحلیل ارزیابی ۳۶۰ درجه با استفاده از نظریه تعمیم‌پذیری به دو منظور انجام می‌شود:

در حالت اول، اثبات این مطلب که در صورت ترکیب ارزیابان و آیتم‌ها داده‌های پایایی برای ارزیابی‌شوندگان تولید

1. Violato et al.
2. Test-retest

می‌شود. نتایج این‌گونه از مطالعات نشان می‌دهد با تعداد قابل قبولی از مشاهده‌گران می‌توان پایایی بیشتر یا مساوی ۰/۷۰ به دست آورد (لیپنر و همکاران[1] ۲۰۰۲ و لاک‌یر و همکاران ۲۰۰۶). به عنوان نمونه، رمزی و همکاران (۱۹۹۳) با پرسشنامه‌ای مشتمل بر ۱۱ آیتم و با استفاده از ۱۰ تا ۱۱ ارزیاب از گروه پزشکان همکار، پایایی ۰/۷۰ را گزارش کردند. وقتی همین گروه از پزشکان توسط پرستاران مورد ارزیابی واقع شدند، ۱۰ تا ۱۵ پرستار لازم بود تا با استفاده از پرسشنامه ۱۳ آیتمی پایایی به ۰/۷۰ برسد (ونریچ و همکاران ۱۹۹۳). ویولت و همکاران[2] (۱۹۹۷) با تعداد کمتری ارزیاب اما پرسشنامه‌ای با تعداد آیتم بیشتر به پایایی ۰/۷۰ رسیدند؛ به طور مثال، با شش همکار پزشک و پرسشنامه‌ای مشتمل بر ۲۶ آیتم، شش همکار غیر پزشک با پرسشنامه ۱۷ آیتمی یا ۲۲ بیمار و پرسشنامه‌ای مشتمل بر ۴۵ آیتم به پایایی مورد نظر دست یافتند. بررسی پرسشنامه‌های مورد استفاده توسط GMC در انگلیس نشان داد تکمیل ۳۴ پرسشنامه بیمار (۹ آیتم) و ۱۵ پرسشنامه همکار پزشک و غیر پزشک (۱۸ آیتم) به منظور دستیابی به پایایی ۰/۷۰ لازم است و تغییر تعداد آیتم‌ها در مقایسه با تغییر تعداد مشاهده‌گران تأثیر کمتری بر افزایش پایایی داشت. بنابراین، یکی از عواملی که در دستیابی به پایایی مورد نظر دخالت دارد گروه مشاهده‌گران و ارتباط آن با فرد مورد ارزیابی است. به طوری که تعداد بیشتری بیمار در مقایسه با همکاران و پرستاران مورد نیاز است. عامل دیگر مؤثر بر پایایی، حیطه مورد ارزیابی است؛ به طوری که به منظور ارزیابی حیطه مهارت ارتباط با بیمار در مقایسه با توانمندی بالینی تعداد ارزیاب همکار (پزشک) بیشتری مورد نیاز است.

در حالت دوم، نظریه تعمیم‌پذیری با هدف اطمینان از تعداد کافی مشاهده‌گران به منظور تعیین پزشکان با عملکرد ضعیف یا دارای مشکلاتی در عملکرد به کار می‌رود. به عنوان مثال آرچر و همکاران (۲۰۰۵) با محاسبه خطای معیار اندازه‌گیری (فاصله اطمینان ۹۵ درصد) گزارش دادند با استفاده از پرسشنامه ۲۴ آیتمی (SPRAT) با مقیاس نمره‌دهی شش‌تایی، تنها چهار ارزیاب برای دستیابی به پایایی مناسب در تشخیص دانشجویان با عملکرد مطلوب کافی است. در مورد mini-PAT تعداد مناسب ارزیابان هشت نفر بودند. علت تفاوت عدد این دو آن است که اولی در مناطقی اجرا شد که در خصوص ارزیابی ۳۶۰ درجه فرهنگ سازی شده بود در حالی که mini-PAT در حوزه وسیعی در مدت کوتاهی اجرا شد.

علاوه بر نظریه تعمیم‌پذیری، نظریه سؤال پاسخ نیز می‌تواند در محاسبه پایایی ارزیابی ۳۶۰ درجه با تعیین ویژگی‌های هر آیتم از جمله خطای معیار هر سؤال مفید باشد. در مجموع، یافته‌های مربوط به پایایی ارزیابی ۳۶۰ درجه، از پایایی مناسب این روش با هدف تکوینی حمایت می‌کند.

تأثیر آموزشی ارزیابی ۳۶۰ درجه

ارتباط بین بازخورد و پیامدهای آن لزوماً خطی نیست و بازخود همیشه به نتایج مطلوب و موردنظر نمی‌رسد. دیده شده است که دریافت کنندگان بازخورد، ارزشیابی‌های مثبت را دقیق‌تر تلقی نموده و افرادی که بر جنبه‌های مثبت رفتارشان متمرکز می‌شوند، به احتمال بیشتری عملکرد خود را بهبود می‌دهند. متاآنالیزی که بر ۶۰۰ مداخله بازخورد بر عملکرد انجام شد، نتایج جالبی را نشان داد: در یک سوم موارد شرکت‌کنندگان ارتقای عملکرد را نشان دادند، در یک سوم موارد تغییری نکردند و در یک سوم موارد دچار افت عملکرد شدند (کلاجر و دنیزی[3] ۱۹۹۶). ارزیابی همیشه با پیامدهای مثبت همراه نیست و ممکن است موجب کاهش انگیزه، تنش روحی و تخریب عملکرد شود. پیامدهای منفی معمولاً زمانی رخ می‌دهد که شرکت‌کنندگان نسبت به فرایند ارزیابی دیدگاه‌های منفی مثل فقدان شفافیت، وجود سوگیری و عدم رعالیت عدالت دارند. این برداشت‌ها می‌تواند احتمال استفاده از بازخورد و در نتیجه بهبود عملکرد را کاهش دهد.

بنابراین، توصیه می‌شود تأثیر آموزشی ارزیابی ۳۶۰ درجه بیشتر مورد بررسی قرار گیرد تا از رخداد یادگیری مطلوب و ارتقای عملکرد مطمئن شویم. مطالعات نشان می‌دهند استفاده از داده‌های فراهم‌شده در ارزیابی ۳۶۰ درجه توسط پزشکان تحت تأثیر

1. Lipner et al.
2. Violato et al.
3. Kluger & DeNisi

برخی عوامل درونی و بیرونی نسبت به ارزیابی‌شونده قرار می‌گیرد. از جمله عوامل درونی می‌توان به موارد زیر اشاره کرد:

- درک پزشکان از عملکرد خود
- احساسات، عقاید و انتظارات فردی
- توانایی فردی ایجاد تغییر

برخی از عوامل بیرونی مؤثر عبارت هستند از:

- ماهیت بازخورد
- اعتبار فرایند ارزیابی و عادلانه بودن آن
- اعتبار بازخورد: منبع (گروه مشاهده‌گران) ارائه‌دهنده بازخورد، محتوای آن، اختصاصی بودن بازخورد و مقایسه با بازخورد از منابع دیگر
- انسجام بازخورد
- موانع تغییر

مطالعات در حوزه صنعت نشان داده‌اند زمانی که تفاوت بین خودارزیابی با داده‌های گروه‌های دیگر زیاد باشد، احتمال استفاده از نتایج توسط فرد کاهش پیدا می‌کند. به عنوان مثال، افرادی که خود را بسیار بالاتر از دیگران ارزیابی می‌کنند، در استفاده از داده‌های ارزیابی ۳۶۰ درجه انگیزه کافی ندارند (اتواتر و همکاران[1] ۱۹۹۵ و ۱۹۹۸). چنین نتایجی در مطالعات حوزه آموزش پزشکی به دست نیامده است. به نظر می‌رسد خودارزیابی در حوزه آموزش پزشکی نیاز به مطالعات بیشتر دارد. تاکنون مطالعات در حوزه آموزش پزشکی نشان داده‌اند پزشکان از توانایی لازم در زمینه خود ارزیابی برخوردار نیستند و اتفاقاً افرادی که از دانش و مهارت کمتری برخوردار هستند، توانمندی خود را بیش از حد تخمین زده‌اند (هولمبو و هاوکینز ۲۰۰۸). البته آگاه‌سازی افراد از حوزه‌هایی که در آن ضعف دارند و توانمندنمودن ایشان در حوزه‌های مورد نظر منجر به افزایش بینش این افراد و ارتقای مهارت خود ارزیابی‌شان می‌شود. بنابراین بازخورد عامل مهمی در ارتقای مهارت خود ارزیابی در افراد است (هولمبو و هاوکینز ۲۰۰۸).

معمولاً، اولین عامل تأثیرگذار بر استفاده از بازخورد، مثبت یا منفی بودن آن است. بازخوردهای مثبت به راحتی توسط دریافت‌کنندگان جذب می‌شود اما افراد در برخورد با بازخوردهای منفی ابتدا اعتبار آن را از چندین جنبه شامل منبع ارائه بازخورد، محتوا و میزان اختصاصی بودن آن و همگرایی با بازخورد از منابع دیگر مورد ارزیابی قرار می‌دهند. سرجانت و همکاران (۲۰۰۷) نشان دادند که تنها نیمی از افراد تحت مطالعه از بازخورد منفی ارائه شده در ارزیابی ۳۶۰ درجه به منظور تغییر استفاده کردند.

منبع ارائه بازخورد نیز بر استفاده از آن و ایجاد تغییر تأثیرگذار است. به نظر می‌رسد پزشکان به بازخوردهای بیماران بیشتر بها می‌دهند از این جهت که آن‌ها در برابر بیماران احساس مسؤولیت بیشتری می‌کنند. توانایی درک شده همکاران در ارزیابی، بر میزان پاسخ به بازخورد تأثیرگذار بود. به عنوان مثال، پزشکان در حال طبابت معتقدند همکاران فرصتی برای مشاهده عملکرد آن‌ها، به ویژه در مورد مشاهده مهارت‌های بالینی ندارند. بنابراین نمی‌توانند منبع خوبی برای ارزیابی این مهارت‌ها و ارائه بازخورد باشند. این در حالی است که پزشکان تنها منبعی هستند که دارای دانش کافی به منظور قضاوت در مورد مهارت‌های بالینی همکارانِ خود هستند. به نظر می‌رسد روایی ارزیابی ۳۶۰ درجه در خصوص ارزیابی مهارت‌های بالینی مورد سؤال باشد و احتمالاً روش‌های عینی‌تر برای ارزیابی این مهارت‌ها مناسب‌تر هستند. این موضوع تأییدی بر «ویژگی موارد» در ارزیابی مهارت‌های بالینی است. در مجموع، منبع ارائه بازخورد و محتوای آن بر استفاده از آن تأثیرگذار است.

یکی دیگر از عوامل تأثیرگذار بر استفاده از بازخورد اختصاصی بودن آیتم‌های پرسشنامه است. در مواردی که آیتم‌های

1. Atwater et al.

پرسشنامه بر رفتار خاصی متمرکز است، استفاده از نتایج را تسهیل می‌کند. به عنوان مثال، در مطالعه سرجانت و همکاران (۲۰۰۷) باخوردهای منفی در مورد آیتم «درمان مناسب را انتخاب می‌کند» یا «اطلاعات تشخیصی را به طور نقادانه مورد ارزیابی قرار می‌دهد» در پرسشنامه همگنان مورد استقبال پزشکان واقع نشد، در حالی که مواردی مانند «نسخه‌ها را شفاف می‌نویسد» در پرسشنامه همکاران داروساز و «پزشک شما به طور شفاف در مورد زمان و نحوه مصرف دارو توضیح می‌دهد» در پرسشنامه بیمار به دلیل اختصاصی بودن آیتم مورد توجه و استفاده پزشکان قرار گرفتند. در مجموع، در این مطالعه بیشترین تغییر مربوط به آیتم‌های مهارت‌های ارتباطی در پرسشنامه بیمار بود. دلیل این امر، باور پزشکان بر مشاهده مستقیم این مهارت‌ها توسط بیمار و اختصاصی بودن آیتم‌ها بود.

ارزیابی‌شوندگان بازخورد ارزیابی ۳۶۰ درجه را با بازخورد دریافت‌شده از منابع دیگر چه به صوت رسمی و چه غیررسمی مقایسه می‌کنند. اگر بازخورد ارزیابی ۳۶۰ درجه در راستا و هماهنگ با دیگر بازخوردها باشد احتمال پذیرش آن و تغییر رفتار افزایش پیدا می‌کند.

به این دلیل که تغییرپذیری قابل ملاحظه‌ای در میزان تاثیرگذاری اطلاعاتی که از طریق بازخورد در اختیار ارزیابی‌شوندگان قرار می‌گیرد و نیز نیاز احساس‌شده فرد به تغییر وجود دارد، لازم است انتظارات ما در خصوص تاثیر آموزشی ارزیابی ۳۶۰ درجه واقعی باشد. ولوسکی و همکاران (۲۰۰۶) در یک مطالعه مروری نشان دادند ۷۰ درصد مطالعات انجام شده در زمینه ارزیابی همگنان از عملکرد پزشکان، تأثیر مثبت بازخورد همگنان بر عملکرد بالینی پزشکان را گزارش کرده‌اند. در یک مطالعه طولی، برینکمن و همکاران[1] (۲۰۰۷) تغییر عملکرد پزشکان در نتیجه اجرای ارزیابی ۳۶۰ درجه را در طول زمان در برنامه PAR آلبرتا بررسی کردند. نتایج مطالعه نشان داد که ارزیابی ۳۶۰ درجه می‌تواند منجر به تغییر رفتار پزشکان به ویژه تعهد حرفه‌ای در طول زمان شود. در محیط صنعتی، متآنالیز انجام شده بر روی ۲۴ مطالعه که در آن‌ها عملکرد افراد در دو یا سه زمان متوالی بررسی شد، نشان داد احتمال تغییر زمانی بیشتر بود که دریافت‌کنندگان بازخورد گرایش مثبتی نسبت به بازخورد داشتند، نیاز ادراک شده‌ای نسبت به تغییر رفتار خود داشتند، واکنش مثبتی نسبت به بازخورد داشتند، باور به امکان‌پذیری تغییر داشتند، اهداف مناسبی را برای تنظیم رفتارشان تعیین کردند و اقداماتی برای بهبود مهارت و عملکرد خود انجام دادند (اسمیتر و همکاران[2] ۲۰۰۵). تحلیل نحوه استفاده ارزیابی شوندگان از داده‌ها در اثبات روایی ارزیابی ۳۶۰ درجه اهمیت دارد. اگر افراد بازخورد را در جهت تغییر به کار نبندند مشکل می‌توان روایی روش ارزیابی را ثابت کرد.

هزینه و قابلیت اجرای ارزیابی ۳۶۰ درجه

ارزیابی ۳۶۰ درجه از قابلیت اجرای نسبتاً خوبی برخوردار است. میزان تکمیل فرم‌ها توسط مشاهده‌گران مناسب گزارش شده است. به عنوان مثال، در مطالعه رایت و همکاران (۲۰۱۲) تنها یک درصد بیماران و پنج درصد همکاران کمتر از ۵۰ درصد آیتم‌های پرسشنامه‌های مربوط را تکمیل کردند. با توجه به این‌که پزشکی حرفه‌ای مبتنی بر کار گروهی است و پزشکان در حین طبابت با افراد زیادی از جمله همکاران پزشک و غیر پزشک سر و کار دارند و همچنین بیماران زیادی را ویزیت می‌کنند، مشکلی در جمع‌آوری تعداد کافی فرم وجود ندارد، اگرچه گاهی پزشکان در جمع‌آوری تعداد کافی فرم با مشکل مواجه می‌شوند مثلاً متخصصان اورژانس، روانپزشکی یا بیهوشی که با بیمارانی که توانایی لازم را ندارند، رو به رو هستند یا پزشکانی که در گروه‌های کوچک یا مناطق جغرافیایی محدود کار می‌کنند. کل زمان صرف شده برای توزیع، تکمیل، تحلیل و ارائه گزارش یک فرم ارزیابی ۳۶۰ درجه کمتر از یک ساعت است که از قابلیت اجرای این روش ارزیابی حمایت می‌کند (آرچر و همکاران ۲۰۰۵). به علاوه استفاده از سیستم‌های آنلاین می‌تواند این فرایند را کوتاه‌تر نیز کند.

1. Brinkman et al.
2. Smither et al.

مقبولیت ارزیابی ۳۶۰ درجه

هر چند ارزیابی توسط همکاران موجب کاهش سوگیری و افزایش پایایی در مقابل ارزیابی کلی پایان دوره توسط استادان می‌شود، در عمل مقاومت‌هایی در پذیرش و اجرای آن وجود دارد. به عنوان مثال دیده شده است که دستیاران تمایلی به ارزیابی همکاران خود ندارند به این دلیل که احساس می‌کنند نقش ارزیابی ارتباطات آن‌ها با همکارانشان را مخدوش می‌کند (ون‌روزندال[1] و جنت ۱۹۹۲). اخیراً مطالعات در مقطع پزشکی عمومی نیز این یافته‌ها را تأیید می‌کند (آرنولد[2] و همکاران ۲۰۰۵، شو[3] و همکاران ۲۰۰۵). بنابراین، در زمینه مشارکت همکاران در ارزیابی ملاحظاتی را باید در نظر داشت. از جمله در ارزیابی حوزه‌هایی پیچیده مانند انسان‌دوستی توسط همکاران باید با احتیاط عمل کرد و از همکارانی که دانش کافی برای قضاوت در مورد این موضوع برخوردار هستند، استفاده نمود.

در مجموع موارد زیر به منظور پژوهش بیشتر در مورد جنبه‌های مبهم و چالش‌برانگیز سودمندی ارزیابی ۳۶۰ درجه پیشنهاد می‌شود:

- بررسی نقش عوامل مؤثر بر ارزیابی مشاهده‌گران مانند میزان آشنا بودن با ارزیابی شوندگان
- فهم و درک عدم توافق با بازخورد ارائه شده
- بررسی روایی معیاری
- بررسی تاثیر آموزشی ارزیابی در طولانی مدت

1. Van Rosendall
2. Arnold
3. Shue

منابع

1. Amin Z, Seng CY, Eng KH. Practical guide to medical student assessment. World Scientific Publishing; 2006.
2. Archer JC, Norcini J, Davies HA. Use of SPRAT for peer review of pediatricians in training. BMJ. 2005;330(7502):1251–3.
3. Archer J, Norcini J, Southgate L, GHeard S, Davies H. mini-PAT (Peer Assessment Tool): a valid component of a national assessment programme in the UK?. Adv Health Sci Educ. 2008;13(2):181–92.
4. Arnold L, Shue CK, Kritt B, Ginsburg S, Stern DT. Medical students' views on peer assessment of professionalism. J Gen Intern Med. 2005;20(9):819–24.
5. Atwater LE, Roush P, Fischtal A. The influence of upward feedback on self and follower rating of leadership. Personnel Psychol. 1995;48(1):35–59.
6. Atwater LE, Yammarino FJ. Does self-other agreement on leadership perceptions moderate the validity of leadership and performance predictions. Personnel Psychol. 1995;45(1):141–64.
7. Atwater LE, Ostroff C. Yammarino FJ, Fleenor JW. Self-other agreement: Does it really matter? Personnel Psycho. 1998;51(3):577–98.
8. Brinkman WB, Geraghty SR, Lanphear BP, Khoury JC, Gonzalez del Rey JA, DeWitt T G, et al. Effect of multisource feedback on resident communication skills and professionalism: a randomized controlled trial. Arch Pediatr Adolesc Med. 2007;161(1):44–9.
9. Cantillon P, Wood D. ABC of Learning and Teaching in Medicine. 2nd ed. West Sussex: John Wiely & Sons; 2010.
10. Dubinsky I, Jennings K, Greengarten M, Brans A. 360-degree physician performance assessment. Healthcare Q. 2010;13(2):71–6.
11. Hall W, Violato C, Lewkonia R, Lockyer J, Fidler H, Toews J et al. Assessment of physician performance in Alberta: the physician achievement review. CMAJ. 1999;161(1):52–7.
12. Holmboe ES, Hawkins RE. Practical guide to the evaluation of clinical competence. Philadelphia: Mosby/Elsevier; 2008.
13. Kluger AN, DeNisi A. The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. Psychol Bull 1996;119:254–84.
14. Lipner RS, Blank LL, Leas BF, Fortna GS. The value of patient and peer ratings in recertification. Acad Med. 2002;77(10):S64–S6.
15. Lockyer J. Multisource feedback in the assessment of physician competencies. J of Contin Educ in Health Prof. 2003;23(1):4–9.
16. Lockyer JM, Viola to C, Fidler H. A multisource feedback program for anesthesiologists. Can J Anesth. 2006;53(1):33–9.

17. Ramsey PG, Wenrich MD, Carline JD, Inui TS, Larson EB, LoGerfo JP. Use of peer ratings to evaluate physician performance. JAMA. 1993;269(13):1655–60.
18. Ramsey PG, Carline JD, Blank LL, Wenrich MD. Feasibility of hospital-based use of peer ratings to evaluate the performances of practicing physicians. Acad Med. 1996;71(4):364–70.
19. Sargeant JM, Mann KV, Ferrier SN. Exploring family physicians' reactions to multisource feedback: perceptions of credibility and usefulness. Med Educ. 2005;39(5):497–504.
20. Sargeant JM, Mann KV, Ferrier SN, Langille D, Muirhead PD, Sinclair DE. Responses of rural family physicians and their colleagues and co-worker raters to a multisource feedback process: a pilot project. Acad Med.10)77;2003):S542–4.
21. Sargeant JM, Mann KV, Sinclair D, van der Vleuten C, Metsemakers J. Understanding the influence of emotions and reflection upon multi-source feedback acceptance and use. Adv Health Sci Educ. 2006;13(3):275-88
22. Shue CK, Arnold L, Stern DT. Maximizing participation in peer assessment of professionalism: the students' speak. Acad Med. 2005;80(10):S1–S5.
23. Smither JW, London M, Reilly RR: Does performance improve following multisource feedback? A theoretical model. Meta analysis and review of empirical findings. Personnel Psychol. 2005;58(1):33–66.
24. Swanwick T. Uderstanding Medical Education: Evidence, Theory and Practice. West Sussex: John Wiely & Sons; 2010
25. Van Rosendall GMA, Jennett PA. Resistance to peer evaluation in an internal medicine residency. Acad Med.1992;67(1):63.
26. Veloski J, Boex JR, Grasberger MJ, Evans A, Wolfson DB. Systematic review of the literature on assessment, feedback and physicians' clinical performance: BEME Guide No. 7. Med Teach. 2006;28(2):117–28.
27. Violato C. Lockyer JM, Fidler H. Multisource feedback: A method of assessing surgical practice. BMJ. 2003;326(7388):546–8.
28. Violato C. Lockyer J, Fidler H. The assessment of pediatricians by a regulatory authority. Pediatrics. 2006;117(3):796–802.
29. Violato C, Marini A, Toews J, Lockyer J, Fidler H. Feasibility and psychometric properties of using peers, consulting physicians, co-workers, and patients to assess physicians. Acad Med. 1997;72(10):S82–S4.
30. Wenrich MD, Carline ID, Giles LM, Ramsey PG: Ratings of the performances of practicing internists by hospital-based registered nurses. Acad Med. 1993;68(9):680–7.
31. Weissman S. Multisource feedback: problems and potential. Acad Med. 2013;88(8):1055.
32. Wright C, Richards SH, Hill JJ, et al. Multisource feedback in evaluating the performance of doctors: The example of the UK General Medical Council patient and colleague questionnaires. Acad Med. 2012;87(12):1668–78.

۷
تعیین حدنصاب قبولی آزمون

فصل ۲۷

مقدمات و کلیات

تعاریف و مفاهیم پایه

در حال حاضر آزمون‌های زیادی در مقاطع و رشته‌های مختلف علوم پزشکی برگزار می‌شوند که هدف آنها اندازه‌گیری میزان توان‌مندی[1] دانشجویان است. اگر هدف از آموزش فراگیران را دستیابی آنها به یک سری پیامد[2] مشخص بدانیم، هدف از برگزاری آزمون، سنجش میزان این دستیابی خواهد بود. از آنجا که نمی‌توان انتظار داشت همه دانشجویان پس از طی یک دوره آموزشی به تمام اهداف تعیین شده رسیده باشند، به طور معمول در پایان دوره از فراگیران آزمون به عمل می‌آید تا سطح آنها از این نظر مشخص گردد. طیف توان‌مندی‌های دانشجویان پس از گذراندن یک دوره گسترده خواهد بود: کسانی که اصلاً به اهداف مورد نظر نرسیده‌اند، کسانی که بخشی از توانایی‌های لازم را کسب کرده‌اند و کسانی که کاملاً توان‌مند[3] محسوب می‌شوند.

همان‌طور که به کرات در این کتاب عنوان شد، برخی از آزمون‌ها حالت تکوینی[4] دارند و هدف از آنها بیشتر این است که با استفاده از نتایجشان به دانشجویان در مورد وضعیت تحصیلی‌شان بازخورد داده شود. به این ترتیب، قبل از اینکه فرصت به پایان برسد، دانشجویان می‌توانند جهت رفع نواقص یادگیری خود اقدام کنند. همچنین مدرس متوجه می‌شود که دانشجویان چه وضعیتی دارند، احیانا چه نکات و مسائلی را خوب نیاموخته‌اند و چه تغییراتی در روند ارائه درس مفید و موثر خواهد بود. با توجه به این توضیحات در این دسته از آزمون‌ها قرار نیست در خصوص رد یا قبول دانشجویان تصمیم‌گیری شود. اما دسته دیگری از آزمون‌ها تجمعی[5] هستند و در نهایت باید وضعیت هر دانشجو را از نظر ردی یا قبولی مشخص کنند. در این موارد، سؤالی که آزمونگران باید به آن جواب دهند این است که در یک امتحان با تعداد مشخصی سؤال، دانشجو چه میزان از سؤالات را باید پاسخ دهد تا واقعاً توان‌مند و شایسته قبولی محسوب شود.

در تمام آزمون‌هایی که به نوعی بحث رد و قبول دانشجویان مطرح است، مهم است که «حداقل نمره قبولی»[6] یا «حدنصاب قبولی» آزمون تعیین گردد تا مرز بین دانشجویان رد و قبول مشخص شود. به این نمره، «نقطه برش»[7] یا «استاندارد» نیز گفته می‌شود. به روند سیستماتیکی که طی آن تصمیم‌گیری و قضاوت انجام می‌شود تا مشخص گردد دانشجویان چه نمره‌ای باید کسب کنند تا در امتحان قبول محسوب شوند، «تعیین استاندارد»[8] می‌گویند. تعیین استاندارد

1. Competency
2. Outcome
3. Competent
4. Formative
5. Summative
6. Minimum Pass Level (MPL)
7. Cut-off
8. Standard setting

در حیطه آزمون، در واقع تعیین نقطه برش یا حداقل نمره قبولی یا حدنصاب قبولی برای تمایز دانشجوی توانمند از غیرتوانمند است. تعیین استاندارد، قضاوتی است که توسط افراد حرفه‌ای انجام می‌شود و محتوای آزمون، هدف آزمون، توانایی دانشجویان و شرایط آموزشی و اجتماعی نیز در آن تاثیرگذار هستند (نورسینی[1] ۲۰۰۳).

آزمونگران تلاش زیادی می‌کنند تا آزمون‌ها در حد امکان، روایی و پایایی بالایی داشته باشند اما باید توجه داشت که بدون تعیین کردن استاندارد آزمون با روش‌های عینی و علمی، پیامد موردنظر به دست نخواهد آمد. معمولاً نمره حدنصاب به صورت قراردادی یک عدد ثابت تعیین می‌شود و این موضوع آن قدر رایج است که تصور وجود روش های دیگر در ابتدا بسیار بعید به نظر می‌رسد اما در حقیقت، روش‌های مختلفی برای تعیین حدنصاب قبولی وجود دارند که که در امتحانات رشته‌های گوناگون از جمله علوم پزشکی توسط دانشگاه‌های متعدد به کار گرفته شده‌اند.

گفتنی است که تقریباً همه روش‌های تعیین استاندارد در علوم پزشکی در ابتدا برای آزمون‌های چندگزینه‌ای استفاده شده‌اند. با معرفی آزمون عینی ساختارمند بالینی[2] از حدود سال ۱۹۷۹ در امتحانات پزشکی، ضرورت تجدید نظر در روش‌های تعیین استاندارد احساس شد (هاردن و گلیسین[3] ۱۹۷۹). البته آزمون OSCE در ابتدا بیشتر به عنوان آزمونی تکوینی و با هدف بازخورد دادن به فراگیران استفاده می‌شد. بنابراین شاید در ابتدا تعیین نمره‌ای به عنوان ملاک و استاندارد قبولی برای آن خیلی حائز اهمیت نبود اما به ویژه از زمانی که OSCE برای امتحانات نهایی و صدور پروانه طبابت به عنوان آزمونی مهم و حساس استفاده شد، مطالعات گسترده‌ای برای تبیین و توسعه روش‌های تعیین استاندارد در آزمون‌های مبتنی بر عملکرد صورت گرفت. البته علی‌رغم این موضوع، روش‌های تعیین استاندارد و ویژگی‌های آنها در آزمون‌های مبتنی بر عملکرد به خوبی و روشنی آزمون‌های کتبی مشخص نشده است (بولت و همکاران[4] ۲۰۰۳).

توضیحات تکمیلی در مورد انواع روش‌ها و جزئیات هر یک از آنها در فصل دوم ارائه می‌شود. آنچه اهمیت دارد این است که بر اساس مطالعات انجام‌شده، گفته می‌شود هیچ روش تعیین استانداردی کامل و بی‌نقص نیست و هیچ روشی نسبت به سایرین برتری ندارد (نیوبل[5] ۲۰۰۴، چیزر و همکاران[6] ۲۰۰۴، وود و همکاران[7] ۲۰۰۶، داویسون و بولاک[8] ۲۰۰۷، بارمان[9] ۲۰۰۸). از طرفی این مسأله نیز در خور توجه است که اگر برای تعیین استاندارد یک آزمون از روش‌های مختلف استفاده شود، هر روش ممکن است نتیجه متفاوتی دهد و هر بار، نمره متفاوتی به عنوان حدنصاب حاصل شود (کافمن و همکاران[10] ۲۰۰۰، کوزیمونو و روتمن[11] ۲۰۰۳). در مورد این موضوع نیز در فصل سوم با جزئیات بیشتر بحث خواهد شد.

ضرورت تعیین استاندارد

یک روش سنتی تعیین استاندارد، مشخص کردن درصد پاسخ‌های صحیحی است که دانشجو باید بدهد تا قبول در نظر گرفته شود. مثلاً وقتی نمره قبولی را ۱۲ از ۲۰ می‌گذاریم، به این معناست که دانشجو باید به ۶۰ درصد سؤالات پاسخ درست بدهد. این روش معایبی دارد از جمله اینکه نمی‌توان برای تمام آزمون‌ها با درجه سختی‌های متفاوت، حداقل نمره قبولی یکسانی در نظر گرفت. از آنجا که مسائلی مانند میزان دشواری سؤالات، شرایط آزمون و نیز سطح و خصوصیات

1. Norcini
2. Objective Structured Clinical Examination (OSCE)
3. Harden and Gleeson
4. Boulet et al
5. Newble
6. Chesser et al
7. Wood et al
8. Davison and Bullock
9. Barman
10. Kaufman et al
11. Cusimono and Rothman

فراگیران در آزمون‌های مختلف متفاوت است، برای اجرای آزمونی عادلانه، معقول و منطقی نیست که حدنصاب قبولی همه آزمون‌ها، نمره‌ای ثابت و از قبل تعیین‌شده (مثلاً ۱۰ یا ۱۲ یا ۱۴) باشد. در واقع مهم است که حدنصاب قبولی هر امتحان برای همان آزمون و با توجه به نکات فوق تعیین شود.

صرف نظر از روش مورد استفاده، مهم است که تعیین استاندارد قبولی آزمون با دقت و حساسیت انجام شود تا قضاوت‌هایی که صورت می‌گیرد، از یک طرف با کمترین خطا و هزینه از نظر لطمه به دانشجویان همراه باشد و از طرف دیگر با آسان‌گیری همراه نشود تا گیرندگان خدمت در جامعه به خاطر آن متضرر گردند.

هنگام تعیین استاندارد هدف آزمون باید در نظر گرفته شود. در آزمون‌های نهایی، حساس، مهم و سطح بالا[1] (از جمله امتحاناتی که به دانشجویان برای طبابت مستقل مجوز و گواهینامه می‌دهند)، موارد مثبت کاذب، یعنی قبولی دانشجویان غیرتوانمند، خطرناک‌تر و تأثیرگذارتر است. بنابراین مطمئن بودن از حدنصاب قبولی در این آزمون‌ها ضرورت بیشتری دارد. در حالی که اگر آزمونی با این هدف برگزار شده است که دانشجویان قبول وارد دوره دیگری شوند که در انتهای آن، ارزیابی مجدد در انتظارشان است، شاید تاکید کمتری بر تعیین استاندارد به شیوه‌های مخصوص شود. به همین ترتیب ممکن است برای آزمون‌های پایان نیمسال یا کلاسی که به صورت معمول برگزار می‌گردند، ضرورت تعیین استاندارد با متدلوژی قوی چندان احساس نشود.

این موضوع بسیار مهم است که قبل از شروع فرایند تعیین استاندارد، ذینفعان و دست‌اندرکاران آزمون، در خصوص ضرورت آن به توافق برسند. از آنجا که در بسیاری از دانشگاه‌ها، تعیین استاندارد به روش‌هایی غیر از نمره ثابت، معمول نیست و همچنین در روند تعیین استاندارد مطمئناً چالش‌های فراوانی پیش خواهد آمد، توصیه می‌شود دستاندرکاران در مرحله اول، مسائلی را از قبیل اینکه چرا باید نمره حدنصاب تعیین شود، این کار چه مزایایی دارد و نتایج آن چه تاثیری بر تصمیم‌گیری‌های بعدی خواهد داشت، برای خود روشن کنند. در همین راستا لازم است اساسا پیامدهایی که از تعیین استاندارد حاصل می‌شوند شناسایی شوند که در فصل سوم به آنها خواهیم پرداخت.

گاهی اوقات دست‌اندکاران مشکلاتی در آزمون احساس می‌کنند و فکر می‌کنند تغییر در روش تعیین استاندارد، می‌تواند راه حلی خوبی برای رفع مشکلات باشد. در حالی که، نمره قبولی به تنهایی، قدرت خیلی کمی برای بهبود آموزش و ارزیابی دارد. همان طور که نمی‌توان صرفاً با اندازه‌گیری قد کودک و تشخیص مناسب بودن آن موجب رشد او شد، برخی از مزایای مورد انتظار از حدنصاب قبولی، بدون ایجاد تغییرات اساسی در سایر اجزای کوریکولوم، دست نیافتنی هستند. همچنین ضروری است که مسؤولان، پیامدهای منفی احتمالی تعیین استاندارد را در نظر بگیرند. مثلاً اینکه برای دانشجویانی که موفق به کسب نمره قبولی نمی‌شوند، چه اتفاقی می‌افتد؟ آیا به آنان کمک و مشاوره ارائه می‌شود؟ چه تعداد ردی قابل قبول است؟ اگر میزان ردی خیلی بالاتر از انتظار باشد، چه باید کرد؟ اگر این موضوعات در همان ابتدا شفاف نشوند و مسؤولان توقعات بیش از حد داشته باشند، فرایند تعیین استاندارد با مشکل مواجه می‌شود و به نتیجه نمی‌رسد (زیکی و همکاران[2] ۲۰۰۶).

تجربه تعیین استاندارد در کشور

در مورد سابقه تعیین استاندارد قبولی در کشور باید گفت که در «راهنـمای برگزاری آزمون بالینی ساختاردار عینی» که توسط دبیرخانه شورای آموزش پزشکی و تخصصی تدوین شده، تعیین استانداردها به روشی درست بسیار با اهمیت عنوان شده است و استفاده از دو روش انگوف تغییریافته و رگرسیون مرزی توصیه شده است (ملکان راد و عین اللهی ۱۳۸۹). همچنین در یکی از مقالات خبرنامه شماره هفده دبیرخانه شورای آموزش پزشکی و تخصصی در مورد تعیین حدنصاب آزمون ذکر شده که علی‌رغم اهمیت تعیین استاندارد هر آزمون به صورت جداگانه و اختصاصی برای همان آزمون، با توجه

1. High stake
2. Zieky et al

به مشکلات اجرایی و سیاست‌گزاری، انجام روش‌های علمی برای تعیین استاندارد عملاً امکان‌پذیر نیست و حتی استاندارد آزمون‌های حساس و مهمی مانند ارتقا و دانشنامه، ثابت و از پیش تعیین‌شده است. در واقع در حد اطلاع، تاکنون برای هیچ یک از امتحانات رسمی در دانشگاه‌های علوم پزشکی کشور، تعیین استاندارد با متدلوژی علمی صورت نگرفته است. در آزمون‌هایی هنجاری عموماً ظرفیت پذیرش، نحوه قبولی را مشخص می‌کند و در اغلب امتحانات معیاری، استاندارد آزمون به صورت قراردادی و ثابت، نمره ۱۰، ۱۲ یا ۱۴ در نظر گرفته می‌شود. هر چند سابقاً در تعداد محدودی از امتحانات معیاری مانند آزمون جامع علوم پایه و آزمون جامع پیش‌کارورزی در دوره پزشکی عمومی، از روش کوهن برای تعیین استاندارد استفاده می‌شد، در حال حاضر استاندارد این آزمون‌ها نیز به نمره ثابت تغییر یافته است.

مراحل تعیین استاندارد

بدیهی است که هر یک از روش‌های تعیین استاندارد، مراحل خاص خودش را دارد که جزئیات آنها در فصل‌های بعدی مورد بررسی قرار خواهد گرفت. اما می‌توان کم و بیش مواردی را به عنوان مراحل مشترک اکثر روش‌ها در نظر گرفت که در اینجا به صورت مختصر به آنها اشاره می‌کنیم:

مشخص کردن مسؤول یا مسؤولان فرایند تعیین استاندارد

پس از اینکه دست اندرکاران و ذی‌نفعان امر آموزش و ارزیابی فراگیران، در مورد ضرورت و دلایل تعیین حدنصاب، به توافق رسیدند، باید فرد (یا افراد) متخصصی به عنوان مسؤول فرایند تعیین استاندارد انتخاب کنند که وظیفه وی، انتخاب روش تعیین استاندارد، انتخاب داوران، آموزش داوران، برنامه‌ریزی برای جلسات، تصمیم‌گیری در مورد امکانات لازم، بودجه بندی و نظارت بر اجرای کل فرایند است.

مشخص کردن روش تعیین استاندارد و مدل نمره‌دهی

مسؤول فرایند، قبل از هر اقدامی باید روشن کند که کدام یک از استانداردهای معیارمحور یا هنجارمحور و کدام مدل نمره‌دهی (جبرانی[1]، غیرجبرانی[2] یا نیمه جبرانی[3]) مدنظر است. سپس روش تعیین استاندارد را تعیین کند. به این منظور، علاوه بر توجه به هدف آزمون، بررسی شواهد درباره روایی و پایایی روش‌های تعیین استاندارد و همچنین قابلیت اجرای آنها مخصوصاً از لحاظ نیروی انسانی و زمان مورد نیاز، باید مورد توجه قرار گیرند.

تعریف سطوح عملکردی و دانشجوی مرزی

تقریباً در تمام روش‌های تعیین استاندارد، برای رسیدن به نمره حدنصاب از مفهوم دانشجوی مرزی[4] استفاده می‌شود. بنابراین، فارغ از اینکه روش تعیین استاندارد چیست، باید مشخص شود که سطوح عملکردی[5] چطور تعریف می‌شوند. به عنوان مثال، فقط دو سطح رد و قبول در نظر گرفته می‌شود یا سه سطح پایه، متوسط و پیشرفته یا شش سطح.

هرچند که بعداً طی فرایند تعیین استاندارد لازم است که داوران در مورد سطوح عملکردی و مفهوم دانشجوی مرزی بحث کنند و ادراکات خود را با یکدیگر در میان بگذارند، اما قبل از شروع فرایند، متخصص تعیین استاندارد باید سطوح عملکردی را تعریف و مشخص کند (زیکی و همکاران ۲۰۰۶).

1. Compensatory
2. Non compensatory
3. Partial compensatory
4. Borderline
5. Performance levels

انتخاب داورها

از آنجا که در فرایند تعیین استاندارد، «قضاوت»[1] نقش کلیدی دارد، کسانی که به امر قضاوت می‌پردازند، باید با دقت انتخاب شوند. مسؤول تعیین استاندارد باید در مورد «خصوصیات» داورها و «تعداد» آنها تصمیم‌گیری کند. همچنین آموزش به داوران، انجام تمرین و دادن بازخورد به آنها در دستیابی به نتایج معتبر بسیار اهمیت دارد (نورسینی ۲۰۰۳).

اجرای روش تعیین استاندارد و محاسبه نمره حدنصاب

چگونگی تعیین استاندارد، بسته به روشی که انتخاب می‌شود، متفاوت است اما عموماً پس از اخذ نظرات خام داوران، بنا به نوع روش، محاسباتی انجام می‌شود و نمره قبولی تعیین می‌گردد. برخی از صاحب‌نظران معتقدند استانداردی که در این مرحله تعیین می‌شود، اولیه و مشروط[2] است و سپس با توجه به مسائلی مانند سیاست‌های آموزشی دانشکده، مسائل اجتماعی و اقتصادی، اهداف ارزیابی، جهت‌گیری بین موارد مثبت کاذب یا منفی کاذب و همچنین احتمال رخداد خطا در فرایند تعیین استاندارد، به استاندارد نهایی و عملی[3] تبدیل می‌شود. به عنوان مثال، ممکن است لازم باشد مسؤول تعیین استاندارد در کنار دست‌اندرکاران اجرایی و سایر ذینفعان در مورد وضعیت دانشجویانی که نمره‌ای نزدیک به نمره حدنصاب دارند، تصمیم بگیرند که آیا با توجه به عوامل مذکور، ترجیح می‌دهند آنها را رد یا قبول کنند. سپس مطابق همین موضع، در استاندارد اولیه تغییر ایجاد کنند (زیکی و همکاران ۲۰۰۶و ریکر[4] ۲۰۰۶).

ارزشیابی و بررسی کیفیت روش تعیین استاندارد

صرف نظر از روشی که برای تعیین نمره حدنصاب انتخاب می‌شود، ضروری است اطمینان حاصل شود که نتایج به دست آمده از کیفیت لازم برخوردار هستند. این موضوع از طرق مختلف می‌تواند کنترل شود: نظرسنجی از ذی‌نفعان، مقایسه با شیوه‌های دیگر، مقایسه نتایج به دست آمده با عملکرد دانشجویان در آینده، محاسبه میزان خطای روش و ... اطلاعات کامل در مورد جزئیات روش‌ها در فصل دوم و بحث تکمیلی در خصوص ارزشیابی روش‌ها در فصل سوم ارائه خواهد شد.

مسائل چالش برانگیز در تعیین استاندارد

علی‌رغم اینکه استفاده از روش‌های تعیین استاندارد در بسیاری از کشورها امری جاافتاده و معمول محسوب می‌شود، بعضی از خصوصیات مرتبط با تعیین استاندارد موجب شده است که کماکان سؤال‌ها و مسائل چالشی در ذهن برخی از افراد باقی بماند. می‌توان تصور کرد که این چالش‌ها در کشور ما که پیشینه کمرنگی در خصوص تعیین استاندارد دارد و اساسا تعیین حدنصاب بحث نوینی محسوب می‌شود، به مراتب بیشتر خواهد بود. در این قسمت، به روشن کردن سه مسأله چالشی عمده در این حوزه می‌پردازیم.

ماهیت ذاتی روش‌های تعیین استاندارد قضاوتی است.

تمام متدهای تعیین استاندارد نیازمند استفاده از قضاوت هستند. این قضاوت به صورت کلی در مورد نحوه عملکرد دانشجویان در برابر یک سؤال است که در برخی از روش‌ها با دیدن عملکرد دانشجویان حین امتحان صورت می‌گیرد و

1. Judgement
2. Provisional
3. Operational
4. Ricker

در برخی دیگر، بدون حضور دانشجو و با تصور کردن عملکرد مورد انتظار آنان.

قضاوتی بودن روش‌های تعیین استاندارد باعث می‌شود که افراد در برابر آنها موضع بگیرند، آنها را غیرعلمی، نسبی و ذهنی[1] بدانند و اساساً آنها را زیر سؤال ببرند. مخصوصا این مسأله که با روش‌های متفاوت، نتایج متفاوتی به دست می‌آید، برای برخی غیرقابل پذیرش است.

نمی‌توان انکار کرد که ماهیت تمام روش‌ها قضاوتی است و مربوط به ذهنیت افراد می‌شود اما در واقع، هیچ کار بهتری نمی‌توان انجام داد. درست است که در برخی از روش‌ها محاسبات آماری و ریاضی انجام می‌شود، اما حقیقت این است که هیچ روش کمی و عینی[2] واقعی برای تعیین حدنصاب وجود ندارد. به هیچ وجه نمی‌توان گفت که حدنصاب در یک آزمون نمره خاصی است که یک گروه از افراد منتخب قرار است با طی مسیر خاصی به آن برسند. نمره حدنصاب، فقط نشان‌دهنده ترکیبی از قضاوت ذهنی افراد مختلف است (زیکی و همکاران ۲۰۰۶).

از آنجا که حتی داوران دارای تجربه و پیش‌زمینه و تخصص مشابه که تحت آموزش یکسان قرار گرفته‌اند، باز هم قضاوت‌های مختلفی خواهند داشت، در نتیجه، بدیهی است که با استفاده از روش‌های گوناگون و یا داوران متفاوت، نمرات مختلفی به عنوان استاندارد تعیین شود (زیکی و همکاران ۲۰۰۶). مانند هر قضاوت دیگری، نمی‌توان انتظار داشت که کسی که به امر داوری می‌پردازد، کل بستر فرهنگی و اجتماعی و سیاسی خود را کنار بگذارد و بدون توجه به آنها حکم دهد. در واقع داور هر چقدر تلاش کند، باز ارزش‌های ناخودآگاه او منبعث از این پیش‌زمینه‌ها خواهد بود. بنابراین، نمره حدنصاب ممکن است تحت تاثیر مسائل شناختی، اجتماعی، اقتصادی، سیاسی و احساسی داوران قرار بگیرد. داوران که از طرفی تمایل دارند نمره قبولی را دست بالا بگیرند و از طرفی نگران رد شدن تعداد زیادی از دانشجویان هستند، خود را در میان یک جنگ درونی احساس می کنند و در حین فرایند تعیین استاندارد فشار روانی بالایی را تحمل می‌کنند (بارمان ۲۰۰۸).

انکار ماهیت قضاوتی روش‌های تعیین استاندارد کاری از پیش نمی‌برد. برای اطمینان‌بخشی به ذی‌نفعان ارزیابی، باید «اعتبار»[3] فرایند را افزایش داد. مهم ترین عاملی که در اعتبار روش نقش دارد، افرادی هستند که به قضاوت می‌پردازند و در مورد ارزش‌ها تصمیم‌گیری می‌کنند. قضاوت درست بستگی به دانش، تجربه و توانمندی داوران هم در زمینه تخصصی آزمون و هم در مورد تعیین استاندارد دارد (بارمان ۲۰۰۸). بنابراین، توجه به عوامل زیر مهم و تاثیرگذار است:

- **تعداد داوران:** اگر امتحان سرنوشت‌ساز و مهمی نیست، مانند امتحان کلاسی، یک نفر داور که معمولا خود مدرس است، ممکن است برای قضاوت کفایت کند اما برای آزمون‌های مهم و سطح بالا تعداد بیشتری از اعضای هیأت علمی بایدی درگیر شوند تا نتیجه قابل قبول باشد.
- **انتخاب داوران:** شاید بیش از اینکه تعداد داوران مهم باشد، خصوصیات ویژگی‌های آنها حائز اهمیت باشد. توصیه می‌شود داورانی از هر دو جنس به صورتی ترکیبی از تخصص‌ها (عمومی، داخلی، جراحی، ...) و نقش‌ها (مدرس، ممتحن، مسؤول اجرایی، ...) برگزیده شوند که تجربه کار و آموزش به دانشجویان همان مقطع مورد نظر را داشته باشند (نورسینی ۲۰۰۳).
- **آموزش داوران:** قبل از شروع فرایند تعیین استاندارد، آموزش به داوران و انجام تمرین باید در نظر گرفته شود. همچنین حین فرایند، دادن بازخوردهای مستمر در دستیابی به نتایج معتبر بسیار اهمیت دارد.

توضیحات بیشتر در خصوص موارد فوق در فصل بعدی و به تفکیک هر روش آمده است.

سطوح عملکردی دانشجویان چگونه باید مشخص شود؟

قبل از انتخاب روش تعیین استاندارد، باید مشخص شود که عملکرد دانشجویان به چند سطح تقسیم می‌شود، نام

1. Subjective
2. Objective
3. Credibility

سطوح چیست و به صورت کلی چگونه تعریف و از هم متمایز می‌شوند. به عبارت دیگر، در هر سطح از دانشجو انتظار می‌رود که چه چیزهایی بداند و چه کارهایی را بتواند انجام دهد. در برخی از متون آموزشی اشاره شده است که در نظر گرفتن سه یا چهار سطح بهتر است و بیش از آن کار داوران را بسیار پیچیده می‌کند (زیکی و همکاران ۲۰۰۶). ممکن است فقط دو سطح رد و قبول در نظر گرفته شود. یعنی عملکرد یک دانشجو یا قابل قبول است یا نیست و غیر از این حالتی ندارد. یا سه سطح پایه، متوسط و پیشرفته در نظر گرفته می‌شود که در آن دانشجو در سطح عملکردی پیشرفته این طور تعریف می‌شود: کسی که در برابر مسائل چالشی دنیای واقعی قادر به نشان دادن توانمندی و به کارگیری دانش و مهارت‌های تحلیلی است.

تصمیم‌گیری در این مورد کار آسانی نیست. به خصوص که تقسیم‌بندی مذکور به صورت کلی برای عموم فراگیران ذکر شده است و باید توجه داشت که شاید برای دانشجویان علوم پزشکی کمی متفاوت باشد. در هر حال، بهترین کار برای تعریف هر سطح این است که مسؤول تعیین استاندارد با کمک طیفی از داوران، رفتارها و توانمندی‌هایی که از دانشجو در هر مقطع و هر موضوع درسی انتظار می‌رود، فهرست کنند. به عنوان مثال، برای تعریف سطوح عملکردی در امتحان پیش‌کارورزی باید از خود بپرسیم از دانشجوی پزشکی که مقطع کارآموزی را تمام کرده است و قرار است وارد مقطع کارورزی شود، انتظار داریم در رشته جراحی چه کارهایی را انجام دهد و چه مطالبی را بداند تا او را در سطح عالی قرار دهیم. بدیهی است که این کار، آسان نیست و باید توسط کسانی انجام شود که با همان رده دانشجویان سر و کار داشته‌اند.

دانشجوی مرزی با توانمندی حداقلی چه مفهومی دارد؟

اغلب روش‌های تعیین استاندارد، با تعریف دانشجوی مرزی با توانمندی حداقلی[1] سر و کار دارند زیرا برای رسیدن به نمره حدنصاب، به نحوی باید دانشجوی مرزی را بیابند و نمره او را به دست آورند.

با وجود اینکه سعی می‌شود تعاریف سطوح عملکردی روشن و واضح باشند، باز هم داوران برای تصور کردن دانشجوی مرزی با مشکل مواجه می‌شوند و علی‌رغم آموزش‌هایی که به آنها ارائه می‌شود، باز هم درک واضح و یکسانی از مفهوم دانشجوی مرزی ندارند. به طوری که برخی از نویسندگان، درک این مفهوم و کار کردن با آن را «مأموریت غیرممکن» خوانده‌اند.

گاهی برخی از داوران، در تلاش برای در نظر گرفتن دانشجوی مرزی، «متوسط» دانشجویان را در نظر می‌گیرند که درست نیست. مثلاً ممکن است عضو هیأت علمی که به عنوان داور حضور پیدا می‌کند، برای درک مفهوم دانشجوی مرزی، ۲۰ دانشجویی را که اخیراً با آنها سر و کار داشته است، تصور کند و برآورد کند که میانگین و متوسط آنها چطور بوده است. باید دقت شود که او با این کار اولاً سطوح عملکردی را که یک مفهوم مبتنی بر معیار است، به حالت هنجاری تبدیل می‌کند که پذیرفته نیست. ثانیاً معمولاً سطح عملکردی دانشجویان متوسط، در حالت مرزی نیست و اتفاقاً بالاتر و در سطح قابل قبول است.

با توجه به این مسائل، لزوم بحث بین داوران قبل از شروع فرایند تعیین استاندارد، انجام چند مورد داوری به عنوان تمرین و ارائه بازخورد حین و بعد فرایند، برای درک بهتر این مفهوم پیچیده بیش از پیش آشکار می شود و توجه به آنها اهمیت پیدا می‌کند.

1. Minimally competent

منابع

1. Barman A. Standard setting in student assessment: is a defensible method yet to come? Ann Acad Med Singapore 2008; 37(11): 957-63.
2. Ben-David MF. AMEE Guide No. 18: Standard setting in student assessment. Medical Teacher 2000; 22(2):120-130
3. Boulet JR, de Champlain A and MCkinley D. Setting defensible performance standards on OSCEs and standardized patient examinations. Medical Teacher 2003;25(3):245–249
4. Chesser A, Laing M, Miedzybrodzka Z, et al. Factor analysis can be a useful standard setting tool in a high stakes OSCE assessment. Medical Education 2004;38:825–831
5. Cusimono MD, Rothman AI. The Effect of Incorporating Normative Data into a Criterion-Referenced Standard Setting in Medical Education. Acad Med. 2003;78(10 suppl):88-90
6. Davison I, Bullock A. Evaluation of the Introduction of the Objective Structured Public Health Examination. The Research Office School of Education University of Birmingham, 2007
7. Harden RM, Gleeson F. Assessment of clinical competence using an objective structured clinical examination (OSCE). Med Edu 1979;13:41-54
8. Hejri SM, Jalili M. Standard setting in medical education: fundamental concepts and emerging challenges. Medical journal of the Islamic Republic of Iran 2014; 28 (34)
9. Kaufman DM, Mann KV, Muijtjens AMM, van der Vleuten CPM. A comparison of standard setting procedures for an OSCE in undergraduate medical education. Acad Med 2000;75:267–71.
10. Newble D. Techniques for measuring clinical competence: objective structured clinical examinations. Med Educ 2004;38:199-203.
11. Norcini JJ. Setting standards on educational tests. Medical education 2003;37(5):464-9.
12. Ricker KL. Setting Cut Scores: Critical Review of Angoff and Modified-Angoff Methods. Alberta Journal of Educational Research 2006;52(1):53-64
13. Wood TJ, Humphrey-Murto SM, Norman GR. Standard setting in a small scale OSCE: a comparison of the Modified Borderline-Group Method and the Borderline Regression Method. Advances in health sciences education: theory and practice2)11;2006):115-22.
14. Zieky M. Perie M, Livingston S. A Primer on Setting Cut Scores on Tests of Educational Achievement, Princeton, N.J: Educational Testing Service, 2006.

۱۵. ملکان راد ا، عین اللهی ب. راهنـمای سـاده برای برگزاری آزمون بالینی ساختاردار عینی. وزارت بهداشت، درمان وآموزش پزشکی معاونت آموزشی و اموردانشجویی، دبیرخانه شورای آموزش پزشکی و تخصصی. دسترسی اسفند ۱۳۸۹ در:
rds.semums.ac.ir/edc/downloads/simple%20help%20for%20OSCE.pdf

۱۶. مقاله علمی طراحی آزمون. خبرنامه شماره هفدهم شورای آموزش پزشکی و تخصصی. دسترسی اسفند ۱۳۸۹ در:
http://dme.hbi.ir/cgme/newsletter/17/pdf17/MAGHALEH-TARAHY.pdf

فصل ۲۸

روش‌های تعیین حدنصاب قبولی آزمون

استانداردهای هنجاری و معیاری

همان‌طور که گفته شد، روش‌های تعیین استاندارد بر اساس هدف آزمون به دو دسته هنجارمحور و معیارمحور تقسیم می‌شوند. در این فصل ابتدا در مورد ویژگی‌های هر کدام از این روشها صحبت می‌کنیم و سپس با جزئیات و مراحل روش‌های مختلف آنها آشنا می‌شویم.

استانداردهای هنجاری

در آزمون‌های هنجارمحور عملکرد هر دانشجو با سایر دانشجویان مقایسه می‌شود. در نتیجه سطح سایر شرکت‌کنندگان بر نتیجه قبول یا رد هر فرد تاثیر می‌گذارد. به عبارت دیگر، لزوماً توانمندی هر شخص یا محتوای آزمون تعیین‌کننده وضعیت قبول یا رد نیستند. به همین دلیل به این دسته، استانداردهای نسبی[1] هم گفته می‌شود.

از آنجا که ممکن است سطح دانشجویان شرکت‌کننده در آزمون متغیر باشد، نمی‌توان مطمئن بود که آیا دانشجوی قبول، واقعاً توانمند است یا صرفاً در مقایسه با سایر دانشجویان عملکرد بهتری داشته است. به همین دلیل علی‌رغم اینکه تعیین استانداردهای هنجاری، آسان و بی دردسر است، توصیه می‌شود برای آزمون‌های گواهینامه و اخذ مجوز مورد استفاده قرار نگیرند. زیرا برای اینکه ببینیم آیا دانشجویی برای فارغ‌التحصیلی و خدمت به بیماران آماده است یا خیر، قیاس عملکرد او با سایر دانشجویان کار معقولی نیست.

مورد استفاده معمول استانداردهای هنجاری در رتبه‌بندی افراد یا آزمونهای پذیرش است که باید تعداد مشخصی از دانشجویان پذیرفته شوند. مثلاً برای ورود به دوره دستیاری از بین تعداد زیادی داوطلب، تعداد محدودی از نفرات برتر انتخاب می‌شوند.

در این روش‌ها همیشه یک تعداد از دانشجویان رد و عده دیگری قبول می‌شوند و میزان قبولی یا ردی هیچگاه صد در صد نیست. بدیهی است که استاندارد آزمون از قبل معلوم نیست و در هر بار تکرار آزمون، با توجه به سطح شرکت‌کنندگان، متغیر خواهد بود. یک نمونه رایج از این نوع، «درصد ثابت» است که ظرفیت پذیرش آزمون، مشخص‌کننده حدنصاب قبولی است. نمونه دیگر از این دسته، «میانگین نمرات» است که به عنوان استاندارد در نظر گرفته می‌شود.

استانداردهای معیاری

در آزمون‌های معیارمحور، عملکرد هر دانشجو با یک نمره استاندارد مقایسه می‌شود و نه با سایر شرکت‌کنندگان. به عبارت دیگر، انتظار می‌رود که دانشجو میزان مشخصی از توان‌مندی‌ها را کسب کرده باشد. این روش به خصوص برای آزمون‌های

1. Relative

حساس، مهم و سطح بالایی که منجر به ارائه مدرک می‌شوند، به کار می‌رود (کوزیمونو ۱۹۹۶ و کریمر و همکاران[1] ۲۰۰۳). به عنوان نمونه، وقتی هدف از آزمون، صدور مجوز ورود به مقطع بالاتر مثلاً کارورزی است، لازم است که هر دانشجو مستقل از نحوه عملکرد سایر فراگیران به حد مطلوب و موردنظر رسیده باشد. در اینجا تعداد کسانی که رد یا قبول می‌شوند از قبل معلوم نیست. ممکن است همه دانشجویان قبول شوند و یا هیچ کس به نمره حدنصاب نرسد. نمره فراگیران روی یکدیگر اثر نمی‌گذارد و ملاک، میزان فعالیت و توان‌مندی هر فرد است. بنابراین به این روشها، روش مطلق[2] هم گفته می‌شود. روش‌های مربوط به استانداردهای معیارمحور به طور کلی به دو دسته مبتنی بر آزمون[3] و مبتنی بر آزمودنی[4] تقسیم می‌شوند.

در روش‌های مبتنی بر آزمون، تمرکز اصلی بر سؤالات امتحان است. معمولاً قبل از آزمون (و گاهی پس از آزمون) از داوران دعوت می‌شود تا در جلسه‌ای شرکت کنند و با بررسی محتویات و ایتم‌های امتحان، استاندارد را تعیین نمایند؛ از جمله این موارد میتوان به روش آنگوف، ابل و ندلسکی اشاره کرد.

در دسته دیگری از روش‌ها که مبتنی بر آزمودنی هستند، نیازی به گروه متخصصان نیست بلکه در حین امتحان از داوران درخواست می‌شود تا در روند تعیین استاندارد شرکت کنند. یکی از مزایای این روش صرفه‌جویی در وقت است؛ مخصوصاً برای هیأت علمی که از نظر زمان مشکل دارند. در اینجا موقعیت مانند روش‌های مبتنی بر آزمون، مجازی و فرضی نیست و در جریان یک آزمون واقعی تصمیم‌گیری صورت می‌گیرد. روش‌های گروه متمایز (متقابل)، گروه مرزی، و رگرسیون مرزی از این دسته هستند

استانداردهای معیاری-هنجاری

روش‌های دیگری وجود دارند که هم خصوصیات استانداردهای معیاری و هم خصوصیات استانداردهای هنجاری را دارند مانند روش هافستی و روش کوهن. تعیین استاندارد به این شیوه آسان‌تر از روش‌های معیاری است ولی در عین حال اگر هدف، افتراق دانشجویان توانمند از غیرتوانمند است، نسبت به روش‌های هنجاری به این هدف نزدیک‌تر است.

در قسمت‌های بعدی با تک‌تک این روش‌ها بیشتر آشنا می‌شویم.

درصد ثابت

یکی از شایع‌ترین روش‌هایی که در آزمون‌های هنجاری استفاده می‌شود، درصد (تعداد) ثابت[5] است. معمولاً ظرفیت پذیرش است که مشخص می‌کند چه درصدی از فراگیران قبول می‌شوند. بنابراین فرقی نمی‌کند که نمره خام هر شخص چقدر بوده است. بلکه جایگاه نمره او در کنار سایر شرکت‌کنندگان و در واقع رتبه او است که در تصمیم‌گیری برای رد و قبل وی اهمیت پیدا می‌کند. به عنوان مثال اگر آزمون پذیرش دستیار تخصصی را به عنوان یک آزمون هنجاری در نظر بگیریم، تعداد دانشجویانی که در یک رشته و دانشگاه خاص پذیرفته می‌شوند، از قبل مشخص است.

گاهی به جای اینکه ظرفیت پذیرش، تعیین‌کننده درصد قبولی باشد، جمع داوران در مورد تعداد قبولی تصمیم می‌گیرند. این روش که اولین بار توسط بیوک[6] در سال ۱۹۸۴ عنوان شد (چین و هرتز[7] ۲۰۰۲)، به این ترتیب است که داوران انتخاب می‌شوند و قبل از برگزاری آزمون، از هر داور سؤال می‌شود که از نظر او چه درصدی از دانشجویان باید در این امتحان قبول شوند. میانگین نظر داوران (با یا بدون احتساب انحراف معیار نظرات) به عنوان ملاک تصمیم‌گیری رد و قبول استفاده می شود.

1. Kramer et al
2. Absolute
3. Exam-centered
4. Examinee-centered
5. Fixed percentage
6. Beuk
7. Chinn and Hertz

ممکن است قبل از محاسبه میانگین، این فرصت در اختیار داوران قرار بگیرد که در مورد این موضوع با هم بحث کنند و دلایل یکدیگر را بشنوند و اگر تمایل داشته باشند، نظراتشان را تغییر دهند. در مورد اینکه داوران چه کسانی باشند، سختگیری چندانی آن گونه که برای برخی از روش‌های دیگر (مانند انگوف) اعمال می‌شود، وجود ندارد.

در روش درصد ثابت، چه بر اساس ظرفیت پذیرش و چه بر اساس نظر داوران باشد، قبولی/ ردی صرفاً ناشی از عملکرد خود شخص نیست. بلکه تحت تاثیر عملکرد سایرین نیز قرار می‌گیرد. اگر به هر دلیلی، سطح داوطلبان در آزمون سال جاری نسبت به سال گذشته خیلی بالا باشد، شانس قبولی فردی که عملکرد متوسط دارد، کم است. در حالی که همین فرد اگر (به عنوان مثال) در امتحان سال گذشته شرکت می‌کرد، بیشتر محتمل بود که قبول شود.

نکته دیگر اینکه در این روش، نمره حدنصاب قبل از برگزاری آزمون مشخص نیست. مثلاً داوطلب کنکور سراسری نمی‌داند که برای قبولی در رشته پزشکی دانشگاه علوم پزشکی تهران دقیقاً لازم است چه نمره خامی کسب کند. اساساً در این روش بیش از اینکه به نمره خام به عنوان حدنصاب توجه شود، رتبه افراد اهمیت پیدا می‌کند.

میانگین نمرات

یکی دیگر از روش‌های تعیین استاندارد هنجاری به این صورت است که پس از اینکه دانشجویان در آزمون شرکت کردند، از نمره آنها میانگین گرفته می‌شود و همان نمره میانگین به عنوان حدنصاب قبولی اعلام می‌شود.

در اینجا نیز، مشابه روش درصد ثابت، قبولی/ردی فراگیر تحت تاثیر عملکرد سایر دانشجویان قرار می‌گیرد. اگر شرکت‌کنندگان قوی باشند، میانگین نمره بالا می‌رود و شانس قبولی فرد متوسط کم می‌شود و برعکس. به علاوه در این روش نیز نمره حدنصاب قبل از برگزاری آزمون مشخص نیست و همیشه تعدادی از دانشجویان در آزمون رد می‌شوند.

گاهی اوقات، برگزارکنندگان آزمون تصمیم می‌گیرند برای تعیین استاندارد، به مقدار یک یا دو خطای معیار اندازه‌گیری[1] به نمره میانگین اضافه یا از آن کم کنند. این کار به دلیل لحاظ کردن احتمال خطا در نمره مشاهده شده انجام می‌شود و نتیجه آزمون را به نمره واقعی فرد نزدیک می‌کند. توضیحات بیشتر در این خصوص در بخش هشتم کتاب ارائه خواهد شد. به عنوان مثال، اگر میانگین آزمون ۱۵ و خطای معیار اندازه‌گیری آن ۰/۵ باشد، حدنصاب قبولی را ۱۶ اعلام می‌کنند (میانگین به اضافه دو خطای معیار). این کار در مواقعی صورت می‌گیرد که امتحان مهم است و باید سخت‌گیری بیشتری اعمال شود تا موارد مثبت کاذب (کسانی که قبول می‌شوند اما توانمند نیستند)، کاهش یابد. برعکس، در برخی از موارد که می‌خواهند منفی کاذب کم شود، به اندازه خطای معیار از میانگین کم می‌کنند.

نمره ثابت

ساده‌ترین و رایج‌ترین روش تعیین استاندارد معیاری، در نظر گرفتن یک نمره ثابت است. برگزارکنندگان آزمون به صورت ساده درنظر می‌گیرند که اگر دانشجویی بتواند نصف سؤالات را پاسخ دهد، یعنی نمره ۱۰ از ۲۰ یا ۵۰ از ۱۰۰ بگیرد، به اندازه کافی خوب هست و قبول اعلام می‌شود. گاهی سخت‌گیری بیشتری می‌شود و نمره ۱۲ (۶۰ درصد) یا ۱۴ (۷۰ درصد) به عنوان نمره قبولی انتخاب می‌شوند.

مشکلی که این روش دارد، این است که برای تمام امتحانات یکسان است. یعنی تفاوتی نمی‌کند که محتوای امتحان چیست یا سطح دشواری سؤالات چگونه است یا شرکت‌کنندگان در چه مقطعی امتحان می‌دهند. در همه حالات، چنانچه درصد مشخصی از سؤالات را جواب دهند، قبول محسوب می‌شوند. به عنوان مثال، اگر آزمون پایان ترم یک درس در

1. Standard error of measurement (SEM)

سال جاری نسبتاً آسان باشد، چون حدنصاب ثابت است، بدیهی است که تعداد بیشتری از دانشجویان، ولو اینکه واقعاً عملکرد خوبی نداشته باشند، قبول می‌شوند. اما چون امتحان مشابه در سال گذشته سخت بوده است، تعدادی از دانشجویان علی‌رغم برخورداری از سطح قابل قبول عملکرد، موفق به گذراندن درس نشده‌اند.

در روش‌های بعدی خواهیم دید که چگونه با بررسی سؤالات آزمون یا عملکرد دانشجویان در آزمون، استاندارد منحصر به همان آزمون تعیین می‌شود.

روش ندلسکی

این روش که اولین بار توسط لئو ندلسکی[1] در سال ۱۹۵۴ معرفی شد، یک روش تعیین استاندارد معیاری از نوع مبتنی بر آزمون است که فقط برای سؤالات چندگزینه‌ای کاربرد دارد. در این روش، برای هر سؤال، گزینه‌های انحرافی که احتمال می‌رود دانشجوی مرزی بتواند حذف کند و تشخیص دهد اشتباه هستند، مشخص می‌شود. مثلاً به نظر داور، دانشجوی مرزی گزینه الف را می‌تواند حذف کند اما تصمیم بین گزینه د (گزینه صحیح) و گزینه ب و ج برای او دشوار است و نمی‌تواند گزینه صحیح را تشخیص دهد. منطق روش در این است که دانشجوی مرزی برای پاسخ به سؤال چندگزینه‌ای از حذف گزینه استفاده می‌کند و سپس از بین گزینه‌های باقی‌مانده به صورت تصادفی یکی را انتخاب می‌کند.

مراحل روش ندلسکی

- داور تک‌تک سؤالات را می‌خواند و دور گزینه‌هایی که فکر می‌کند دانشجوی مرزی می‌تواند به راحتی کنار بگذارد، خط می‌کشد.
- برای هر سؤال گزینه‌هایی که باقی مانده‌اند، یعنی دورشان خط کشیده نشده است، شمارش می‌شوند و عدد یک بر مقدار به دست آمده تقسیم می‌شود تا استاندارد سؤال به دست آید. در مثال بالا، یک گزینه حذف شده بود و سه گزینه باقی مانده بود. پس احتمال پاسخگویی دانشجوی مرزی به این سؤال، یعنی همان استاندارد سؤال، یک‌سوم و معادل ۳۳ درصد است. اگر داور احتمال دهد که دانشجوی مرزی نمی‌تواند گزینه‌ای را حذف کند، از آنجا که دانشجو کماکان می‌تواند با حدس تصادفی یک گزینه را به صورت شانسی انتخاب کند، احتمال پاسخگویی برای وی یک چهارم یعنی ۰/۲۵ در نظر گرفته می‌شود. به صورت کلی مقادیری که می‌تواند به سؤال چهارگزینه‌ای اختصاص یابد، ۰/۲۵، ۰/۳۳، ۰/۵ و ۱ است.
- استاندارد تمام سؤالات با یکدیگر جمع می‌شود تا نمره مورد انتظار برای دانشجوی مرزی به دست آید که همان حدنصاب آزمون است (جدول ۱-۲۸).

جدول ۱-۲۸: تعیین استاندارد آزمون چهارگزینه‌ای با ۵ سؤال به روش ندلسکی توسط یک داور (جمع نمره: ۵)

سؤال	گزینه الف	گزینه ب	گزینه ج	گزینه د	استاندارد
سؤال ۱	*	*	*		۱
سؤال ۲				*	۰/۳۳
سؤال ۳			*	*	۰/۵۰
سؤال ۴		*			۰/۳۳
سؤال ۵					۰/۲۵
کل					۱/۴۱

1. Leo Nedelsky

این روش با قضاوت یک نفر امکان‌پذیر است اما بهتر است برای اعتبار بیشتر، از نظر چندین داور استفاده شود. با توجه به زمانی که برای توافق بین داوران صرف می‌شود، باید مراقب بود که روند تعیین استاندارد بیشتر طول می‌کشد. در این حالت، مراحل زیر انجام می‌شود:

- ابتدا داوران انتخاب می‌شوند و از آنها برای شرکت در جلسه تعیین استاندارد دعوت به عمل می‌آید.
- جلسه تعیین استاندارد تشکیل می‌شود و سؤالات آزمون در اختیار داوران قرار می‌گیرد. هر داور هر سؤال را به صورت انفرادی بررسی می‌کند و گزینه‌هایی را که فکر می‌کند دانشجوی مرزی می‌تواند به راحتی کنار بگذارد، مشخص می‌کند.
- بین داوران بحث مختصری انجام می‌شود. البته این قسمت اختیاری است و می‌تواند بنا به صلاحدید و امکانات و وقت موجود انجام شود یا نشود. برای شروع بحث، مسؤول جلسه از داوران می‌پرسد که چند نفر دور گزینه الف را خط کشیده‌اند و چند نفر خط نکشیده‌اند. سپس یکی از هر گروه به ارائه دلایل خود می‌پردازد. هدف این نیست که افراد قانع شوند و حتما به توافق برسند. بلکه صرفاً به نقطه نظرات یکدیگر گوش می‌دهند و در صورت تمایل نظر خود را عوض می‌کنند.
- پس از اینکه همین روند برای تمام سؤالات انجام شد، استانداردی که توسط هر داور به دست آمده است، محاسبه می‌شود.
- در نهایت استاندارد همه داوران جمع‌آوری می‌شود و برای تعیین استاندارد کل، می‌توان از نظر داوران میانگین گرفت. در برخی از موارد به جای میانگین از شاخص میانه برای تعیین استاندارد کل استفاده می‌شود.

محدودیت‌های روش ندلسکی

از محدودیت‌های اصلی این روش، این است که صرفاً برای آزمون‌های چندگزینه‌ای قابل اجرا است. البته لازم نیست تعداد گزینه‌های تمام سؤالات آزمون یکسان باشد. حتی اگر تعدادی از سؤالات سه‌گزینه‌ای و تعدادی دیگر چهار یا پنج‌گزینه‌ای باشند، می‌توان حدنصاب آزمون را با روش ندلسکی محاسبه کرد.

مسأله دیگر اینکه برای اجرای این روش داوران باید همزمان در جلسه حاضر شوند. همچنین زمان زیادی برای این کار وقت بگذارند که با توجه به محدودیت وقت هیأت علمی، کار دشواری است. باید توجه داشت که ناآشنایی داوران با روش مشکل ایجاد خواهد کرد و روند تعیین استاندارد را کند خواهد نمود. بنابراین مانند بقیه روش‌ها توصیه شده است که قبلاً به آنها آموزش داده شود و مراحل کار را برای چند سؤال به صورت تمرینی اجرا کنند.[1]

نکته دیگری که در خصوص این روش ذکر شده است، محدود بودن احتمالاتی است که برای هر سؤال می‌توان در نظر گرفت. به عنوان مثال همان‌طور که ذکر شد، در یک سؤال چهارگزینه‌ای احتمالات به این صورت است: ۰/۲۵، ۰/۳۳، ۰/۵۰ و ۱. همچنین فواصل این احتمالات با یکدیگر برابر نیست. این مسأله نیز قابل تأمل است که معمولاً داوران از دادن عدد ۱ به سؤال، یعنی حالتی که دانشجوی مرزی بتواند تمام گزینه‌ها را کنار بگذارد، اجتناب می‌کنند و به جای آن حذف دو گزینه و احتمال ۰/۵۰ را در نظر می‌گیرند. این امر که موجب می‌شود استاندارد سؤال و به تبع آن استاندارد آزمون کاهش پیدا کند که باید در جلسات آموزشی داوران توضیح داده شود.

یک مسأله در روش ندلسکی (و همین‌طور روشهای انگوف و ابل) این است که آیا باید جواب صحیح سؤالات را در اختیار داوران قرار داد یا نه. اگر داوران گزینه صحیح را از قبل بدانند، احتمالاً سؤال برایشان آسان‌تر به نظر می‌رسد و نمره حدنصاب بالاتر به دست می‌آید. اگر جواب صحیح در اختیارشان قرار نگیرد، چند احتمال وجود دارد: یکی اینکه داوران احساس می‌کنند مورد ارزیابی قرار گرفته‌اند و فراموش می‌کنند که کار اصلیشان چه بوده است. دوم اینکه روند تعیین استاندارد بیشتر طول می‌کشد. سوم اینکه ممکن است گزینه درست را نادرست فرض کنند و به اشتباه بگویند که دانشجوی مرزی آن را حذف

1. برای بحث بیشتر در خصوص ویژگی و تعداد داوران به روش انگوف مراجعه کنید.

می‌کند. از آنجا که از داوران متخصص برای شرکت در جلسه دعوت می‌شود، احتمال رخداد این اتفاق کم است، اما بالاخره وجود دارد. یک راه حل این است که ابتدا از داوران خواسته شود به سؤالات جواب بدهند. سپس کلید را در اختیارشان گذاشت تا برگه خود را تصحیح کنند و سپس فرایند داوری را شروع کنند.

روش انگوف

این روش که یکی از معروف‌ترین و رایج‌ترین روش‌های تعیین استاندارد معیاری است، در سال ۱۹۷۱ توسط ویلیام انگوف[1] معرفی شد (انگوف ۱۹۷۱). اگرچه این روش مانند روش ندلسکی در ابتدا برای آزمون‌های چندگزینه‌ای طراحی شده بود، اما برای آزمون‌های تشریحی و عملی نیز قابل استفاده است. انگوف یک روش مبتنی بر آزمون است بنابراین از متخصصان درخواست می‌شود تا محتوای آزمون یعنی آیتم‌های مختلف آن را تحلیل و بررسی کنند و بر آن اساس استاندارد را تعیین کنند. چون جلسه داوری معمولاً قبل از برگزاری آزمون تشکیل می‌شود، یکی از خوبی‌های ذکرشده برای آنگوف این است که مانند روش ندلسکی دانشجویان استاندارد را قبل از شرکت در آزمون می‌دانند.

مراحل روش انگوف

آنچه در زیر می‌آید، مراحل روش انگوف برای آزمون چندگزینه‌ای است. اما دقیقاً همین فرایند برای تعیین استاندارد آزمون تشریحی و OSCE نیز به کار می‌رود. فقط کافی است به جای سؤال، ایستگاه (شامل سناریو و چک لیست آن) گذاشته شود.

- ابتدا داوران انتخاب می‌شوند و از آنها برای شرکت در جلسه تعیین استاندارد دعوت به عمل می‌آید.
- جلسه تعیین استاندارد (معمولاً قبل از برگزاری آزمون) تشکیل می‌شود. سؤالات آزمون در اختیار داوران قرار می‌گیرد و هر داور به صورت مستقل به این پرسش پاسخ می‌دهد که چقدر احتمال دارد یک دانشجوی مرزی بتواند به هر سؤال پاسخ صحیح بدهد. این احتمال به صورت یک عدد از صفر تا ۱۰۰ عنوان می‌شود. به این منظور، داور باید سطح دشواری سؤال، مقطع دانشجویان، هدف آزمون، مفهوم دانشجوی مرزی و .. را در نظر داشته باشد.
- احتمالاتی که داورهای مختلف به هر سؤال داده‌اند، جمع‌آوری و میانگین آنها محاسبه می‌شود. به این ترتیب استاندارد آن سؤال به دست می‌آید.
- برای تعیین استاندارد کل آزمون، استانداردهای همه سؤالات جمع‌آوری و میانگین آنها محاسبه می‌گردد (جدول ۲-۲۸).

جدول ۲-۲۸: تعیین استاندارد یک آزمون با ۶ سؤال به روش انگوف با شرکت ۵ داور

سؤال	داور ۱	داور ۲	داور ۳	داور ۴	داور ۵	استاندارد
سؤال ۱	۵۰	۳۰	۶۵	۵۰	۵۵	۵۰
سؤال ۲	۷۰	۵۵	۷۵	۸۰	۶۰	۶۸
سؤال ۳	۳۰	۲۵	۳۵	۳۵	۲۵	۳۰
سؤال ۴	۶۰	۵۵	۵۰	۶۰	۶۵	۵۸
سؤال ۵	۴۵	۴۵	۵۰	۵۵	۶۰	۵۱
سؤال ۶	۵۰	۲۵	۳۰	۴۰	۵۵	۴۰
کل						۴۹/۵

1. William Angoff

انتخاب داوران آنگوف

فرایند قضاوت در تعیین استاندارد بسیار حائز اهمیت است. بنابراین، کسانی که به عنوان داور دعوت می‌شوند، باید شایستگی این کار را داشته باشند.

در مورد ویژگی‌های داوران توصیه شده که بهتر است ترکیبی از نقش‌های متخصص، عمومی، استاد و پزشک را داشته باشند و در کنار آن از نظر جنس، سن و نژاد بین آنها تعادل برقرار باشد و هیچ کدام در این امر ذی‌نفع نباشند. انتخاب داوران آنگوف از دیسیپلین‌ها و موقعیت‌های حرفه‌ای گوناگون با توزیع جنسی و سنی متنوع، روایی مطالعه را بهبود خواهد داد (نورسینی ۲۰۰۳).

داورها باید با ویژگی‌ها و توانایی‌های دانشجویانی که قرار است در آزمون شرکت کنند، به خوبی آشنا باشند؛ با آزمونی که قرار است آن را داوری کنند، آشنایی داشته باشند و در مورد طیف عملکرد و رفتار از عالی تا رضایت‌بخش، ضعیف و غیرقابل قبول دید روشنی داشته باشند. از آنجا که یکی از علل نتایج ضعیف تعیین استاندارد را فقدان آموزش و تمرین داوران می‌دانند (ایمپارا و پلاک[1] ۱۹۹۷)، هرچند پروتکل آموزشی مشخص و استانداردی برای داوران وجود ندارد اما لزوم آموزش و آشناسازی داوران با فرایند تعیین استاندارد، انجام تمرین قبل از آغاز کار اصلی و ارائه بازخوردهای مکرر از جمله مسائلی است که توصیه شده است و باید مورد توجه قرار بگیرد (بولت و همکاران ۲۰۰۳). مخصوصاً برای اینکه همه داوران مفهوم عملکرد دانشجوی مرزی را به وضوح درک نمایند، بحث در این خصوص در آغاز جلسه بسیار کمک‌کننده است. برنامه پیشنهادی یک جلسه تعیین استاندارد به شیوه انگوف در جدول ۳-۲۸ آمده است.

جدول ۳-۲۸: برنامه پیشنهادی جلسه تعیین استاندارد آزمون OSCE با ۱۴ ایستگاه به روش آنگوف با استفاده از بحث

موضوع		روش	زمان
مقدمه	مفهوم و ضرورت تعیین استاندارد آزمون انواع روش های تعیین استاندارد	سخنرانی	۱۰ دقیقه
روش آنگوف	نحوه کار در روش آنگوف	سخنرانی	۱۰ دقیقه
سطوح عملکردی	سطوح عملکردی و توان‌مندی با ذکر مثال مفهوم دانشجوی مرزی	سخنرانی و بحث گروهی	۲۰ دقیقه
تمرین	تعیین استاندارد برای دو ایستگاه آزمون سال گذشته	انفرادی/بحث گروهی	۲۰ دقیقه
تعیین استاندارد ۱	تعیین استاندارد برای ۷ ایستگاه OSCE	کار عملی انفرادی	۴۰ دقیقه
بحث تعیین استاندارد ۱	بحث گروهی در مورد استاندارد هر ایستگاه و بازبینی استاندارد	بحث گروهی/انفرادی	۲۰ دقیقه
استراحت			۱۰ دقیقه
تعیین استاندارد ۲	تعیین استاندارد برای ۷ ایستگاه دیگر	کار عملی انفرادی	۴۰ دقیقه
بحث تعیین استاندارد ۲	بحث گروهی در مورد استاندارد هر ایستگاه و بازبینی استاندارد	بحث گروهی/انفرادی	۲۰ دقیقه

در مورد تعداد داورانی که برای هدایت جلسه آنگوف مناسب است، تحقیقات متعددی صورت گرفته است و نتایج متفاوتی به دست آمده است. از آنجا که افزایش تعداد داور برای رسیدن به پایایی بیشتر لازم است، می‌توان به سادگی گفت

1. Impara and Plake

که هرچه تعداد داوران بیشتر باشد، بهتر است اما عملاً به دلیل محدودیت‌های اجرایی این امر امکان‌پذیر نیست و عموم مطالعات به بررسی این مسأله پرداخته‌اند که چه تعداد داور برای رسیدن به پایایی مطلوب کافی است.

- در یکی از اولین مطالعات، گفته شد که حداقل تعداد داوران برای تعیین استاندارد یک امتحان مهم و حساس، باید شش تا هشت باشد (برنان و لاکوود[1] ۱۹۸۰).
- نورسینی و همکاران در سال ۱۹۸۷ بیان کردند که افزایش تعداد داوران از پنج نفر به ۱۰ نفر، خطا را کم نمی‌کند و پایایی استاندارد را افزایش نمی‌دهد. بنابراین همان پنج نفر را کافی دانستند.
- (نورسینی و همکاران ۱۹۸۸). مارر و همکاران[2] ابتدا خودشان پژوهشی انجام دادند و سپس در مطالعه دیگری به بررسی نتایج پنج مطالعه پرداختند و به این نتیجه رسیدند که بسته به حیطه مورد ارزیابی، شرایط زمینه‌ای و نوع روش انگوف، تعداد مطلوب داوران از ۱۲ تا ۲۰ نفر متغیر است (مارر و همکاران ۱۹۹۱، مارر و الکساندر[3] ۱۹۹۲).
- در سال ۱۹۹۹ مطالعه دیگری روی داده‌های انگوف هشت امتحان گواهینامه در رشته‌های مختلف از جمله پزشکی انجام شد که پایایی انگوف را با استفاده از نظریه تعمیم‌پذیری[4] محاسبه کرد. نتایج نشان داد ۱۰ تا ۱۵ داور بهینه است، هرچند گاهی کمتر از این تعداد هم کافی است. نویسندگان توصیه کردند که به صورت قانون کلی ۱۰ تا ۱۵ داور انتخاب شود ولی هر بار آنالیز به منظور تعیین تعداد مناسب داور برای همان آزمون و همان نوع انگوف صورت گیرد (هرتز و هرتز[5] ۱۹۹۹).
- پس از آن مطالعه برندان[6] نشان داد حداقل ۱۰ داور و به صورت ایده‌ال ۱۵ تا ۲۰ داور برای آزمون‌های چندگزینه‌ای مناسب است (برندان ۲۰۰۴).
- فوئل و همکاران[7] مطالعه‌ای بر روی داده‌های تعیین استاندارد به دو روش آنگوف تغییریافته و ابل انجام دادند که در چهار سال متوالی در دانشگاه لیورپول جمع‌آوری شده بود. بر اساس نتایج این مطالعه، حداقل ۱۰ داور برای تعیین استاندارد به شیوه‌های مبتنی بر آزمون لازم است. نویسندگان اضافه کردند که در صورت گنجاندن جلسه بحث بین داوران، تعداد شش نفر برای دستیابی به پایایی مورد قبول کافی خواهد بود. به عبارت دیگر بحث بین داوران پایایی را بهبود می‌دهد و در صورتی که اجرا شود، به تعداد کمتری داور نیاز است (فوئل و همکاران ۲۰۰۶). البته باید توجه داشت که این مطالعه روی آزمون‌های کتبی انجام شده بود و شاید به راحتی قابل تعمیم به آزمون‌های مبتنی بر عملکرد نظیر OSCE نباشد.

باید توجه داشت که موارد گفته شده، کم و بیش در مورد هر سه روش تعیین استاندارد مبتنی بر آزمون که مستلزم تشکیل گروه داوران هستند (ندلسکی، انگوف و ابل)، قابل استفاده هستند. همچنین، این مطالعات در حیطه‌های مختلف آموزشی انجام شده‌اند و برای تعمیم به حوزه علوم پزشکی و مخصوصا آزمون‌های عملی باید با احتیاط عمل کرد.

انواع انگوف

آنچه ذکر شد پایه و اساس روش انگوف است. در سال‌های بعد، تغییراتی توسط پژوهشگران مختلف در روش انگوف اعمال شده است که تقریباً همه آنها تحت عنوان کلی روش انگوف تغییریافته[8] شناخته می‌شوند و کاربرد دارند. از جمله این تغییرات، بحث بین داوران در فواصل کار و همچنین در انتهای جلسه است که بر اساس آن، داوران می‌توانند در نمراتی که داده‌اند تغییراتی اِعمال کنند. در نوع دیگری از روش انگوف، پیشنهاد شده که قبل از جلسه بحث انتهایی، نمرات واقعی دانشجویان در اختیار داوران قرار گیرد.

- **استفاده از بحث:** برخی پیشنهاد کرده‌اند که در روش انگوف، برای کاهش واریانس بین نظرات داورها و افزایش

1. Brenan and Lockwood
2. Maurer et al
3. Maurer and Alexander
4. Generalizability theory
5. Hurtz and Hurtz
6. Brandon
7. Fowell et al
8. Modified Angoff

توافق بین آنها، در فواصل کار و همچنین در انتهای جلسه، بحث[1] صورت گیرد (هامبلتون و پلاک[2] ۱۹۹۵). یعنی پس از تعیین استاندارد به صورت انفرادی، این فرصت در اختیار داوران قرار می‌گیرد که در مورد دلایل نظر خود صحبت کنند و استدلال داوران دیگر را بشنوند و سپس مجدداً نظر خود را اعلام کنند. داوران مختار هستند که استاندارد خود را عوض کنند یا نکنند. مجدداً اعداد جمع‌آوری می‌شود و استاندارد به همان روش قبلی محاسبه می‌شود. به صورت کلی بحث به منظور کاهش خطا در تعیین نمره قبولی، افزایش پایایی تعیین استاندارد و افزایش توافق بین داوران انجام می‌شود. در واقع فرض بر این است که توافق اگرچه صحت استاندارد را تضمین نمی‌کند، احتمال بهبود آن را افزایش می‌دهد اما باید دید که مطالعات تجربی در این باره چه می‌گویند. در مطالعه استرن و همکاران[3] استاندارد کل بعد از بحث تقریباً مشابه استاندارد اولیه بود اما تفاوت بین داورها بعد از بحث کم شده بود (استرن و همکاران ۲۰۰۵). در مطالعه مرتاض هجری و همکاران، استاندارد پس از بحث افزایش یافت که تغییر آن از نظر آماری معنادار نبود اما ضریب توافق بین داوران پس از بحث افزایش یافت (مرتاض هجری و همکاران ۱۳۹۰).

استرن و همکاران ۲۰۰۵

در این مطالعه تعیین استاندارد یک OSCE با ۱۰ ایستگاه انجام شد. برای روش آنگوف تغییریافته همه افراد ابتدا در یک جلسه شرکت می‌کردند تا نسبت به موضوع آشنا شوند. سپس سناریوی هر ایستگاه را می‌خواندند و تخمین می‌زدند که یک دانشجوی مرزی چند آیتم یک چک‌لیست را می‌تواند به طور صحیح جواب دهد. سپس بحث گروهی اجرا می‌شد و تخمین تکرار می‌شد. در هر ایستگاه داوران در سه حوزه شرح حال، معاینه بالینی و مهارت ارتباطی در مقیاس لیکرت ۵ تایی نمره می‌دادند. تمام داورها استاندارد تقریباً مشابه‌ای ارائه دادند. استاندارد حاصل از آنگوف اولیه و ثانویه هم تقریباً یکسان بود اما تفاوت بین داورها کم شده بود.

مرتاض هجری و همکاران ۱۳۹۰

در این مطالعه، استاندارد قبولی OSCE در مقطع پیش کارورزی دانشگاه علوم پزشکی تهران به روش انگوف تعیین شد. به این ترتیب که ۱۱ داور به صورت مستقل، احتمال قبولی یک دانشجوی مرزی را در هر یک از ایستگاه ها برآورد کردند. میانگین احتمالات تمام داوران در تمام ایستگاه‌ها، استاندارد آزمون محسوب شد. این روند دو بار دیگر، پس از برگزاری جلسه بحث بین داوران و بعد از بررسی نمرات واقعی دانشجویان تکرار شد. استاندارد اولیه ۴۹/۱۵، استاندارد بعد از بحث ۴۹/۹۰ و استاندارد پس از بررسی نمرات واقعی ۵۱/۵۲ به دست آمد. افزایش استاندارد سوم نسبت به استاندارد اول معنادار بود (p=0.02) اما نسبت به استاندارد دوم معنادار نبود. همچنین میزان قبولی به ترتیب ۶۷/۶، ۶۴/۸ و ۵۸/۱ درصد به دست آمد که درصد قبولی سوم نسبت به اول کاهش معنادار داشت (p=2.002). میزان توافق بین داوران به ترتیب ۰/۷۷، ۰/۸۸ و ۰/۹۵ به دست آمد.

بر اساس نتیجه متاآنالیزی که در سال ۲۰۰۳ توسط هرتز و آئروباخ[4] به منظور بررسی تأثیر روش‌های مختلف آنگوف انجام شد، بحث گروهی در مورد خصوصیات دانشجوی مرزی و در مورد استانداردهای نوبت اول، توافق بین داوران را بالا برد و همچنین بعد از بحث، استاندارد به طور معناداری افزایش پیدا کرد. نویسندگان به بررسی علت این تغییرات پرداختند. آنان ذکر کردند که تغییراتی که داوران پس از بحث در مورد برآوردها اعمال می‌کنند، شاید ناشی از دینامیک گروه باشد. اما نامشخص است که این جهت‌گیری لزوماً باعث استاندارد معتبرتر نیز می‌گردد یا نه. در واقع اگر بحث باعث شود جنبه نامشخصی از یک سؤال روشن شود، یعنی یک اثر سیستماتیک و بر پایه انتقال اطلاعات داشته باشد، آن وقت است که می‌توان گفت که اعتبار را بالا برده است اما این تغییرات ممکن است ناشی از دینامیک گروهی باشد و آنچنان که در یک مطالعه ذکر شده بود، برخی داوران تأثیر شدیدی بر جلسه و بحث‌های آن داشته باشند و جهت را به یک سمت عوض کنند. بنابراین، نویسندگان وجود یک تسهیل‌گر برای هدایت بحث گروهی و اجرای یک سری دستورالعمل به منظور اداره جلسه پیشنهاد دادند تا روند جلسه به گونه‌ای پیش رود که تبادل اطلاعات

1. Discussion
2. Hambleton and Plake
3. Stern et al
4. Hurtz and Auerbach

زیاد باشد و در مقابل، اثر عوامل اجتماعی و روانشناختی کمرنگ شود (هرتز و آئروباخ ۲۰۰۳). توضیح دیگری که در این خصوص ارائه شد، این بود که برخی داوران وقتی نظر خود را در جمع مطرح می‌کنند، تمایل دارند استانداردهای حرفه‌ای و شخصی خود را بالاتر نشان دهند. بنابراین، پس از بحث سختگیر می‌شوند و استانداردهای خود را بالا می‌برند (لیری و کولاسکی[1] ۱۹۹۰). در هر صورت چنانچه علت تغییر استانداردها، مسائل مذکور باشد، روایی مطلوب حاصل نشده است و عادلانه هم نیست که ردی یا قبولی دانشجویان تحت تاثیر این موارد قرار بگیرد.

- **استفاده از نمرات واقعی**: در روش آنگوف با بررسی نمرات واقعی[2] پیشنهاد شد که بعد از تعیین استاندارد به صورت انفرادی یا با بحث، نمرات واقعی دانشجویان در اختیار داوران قرار بگیرد. به این منظور پس از برگزاری آزمون، جلسه دیگری تشکیل می‌شود و از همان داوران دعوت به عمل می‌آید. علاوه بر استانداردهایی که در جلسه قبلی تعیین کرده بودند، نتایج عملکرد دانشجویان در هر سؤال در اختیار آنها قرار می‌گیرد. داوران این فرصت را دارند که در صورت تمایل در استانداردهای خود تجدیدنظر کنند و همان مراحل معمول انگوف به اجرا در می‌آید. برخی معتقدند که دیدن نمرات واقعی، باعث اغتشاش روند تعیین استاندارد می‌شود زیرا انگوف روشی مبتنی بر معیار است و در اختیار گذاشتن نتایج عملکرد دانشجویان به منظور اصلاح استاندارد گویی آن را به نوعی به روش هنجاری تبدیل می‌سازد. با این وجود، این نکته را نباید از ذهن دور داشت که تمام استانداردها به شکلی بر پایه هنجارها قرار گرفته‌اند. قضاوت در مورد اینکه یک دانشجو چه کاری «باید» انجام دهد، همیشه بستگی به این دارد که فکر می کنیم چه کاری «می‌تواند» انجام دهد (زیکی و همکاران ۲۰۰۶). به عنوان مثال، هیچ کس انتظار ندارد یک ورزشکار مسافت یک کیلومتر را در زمان یک دقیقه بدود. چون عملاً هیچ ورزشکاری نمی‌تواند این کار را بکند. بنابراین، هدف از ارائه نمرات واقعی آگاه کردن داوران از کاری است که دانشجویان می‌توانند انجام دهند. حتی کسانی که موافق ارائه نمرات واقعی به داوران هستند، درباره نوع اطلاعاتی که باید ارائه شود و روش ارائه آن، همچنان بحث می‌کنند. از اطلاعاتی که می‌توان به داوران داد، ضریب دشواری هر سؤال، میانگین، میانه، حداقل و حداکثر نمره، میزان قبولی با در نظرگرفتن استاندارد قبلی و ... است. اما اینکه لزوماً ارائه همه آنها مفید است یا موجب سردرگمی بیشتر داور می‌شود، مشخص نیست. مسأله‌ای که باید مراقب بود این است که عملکرد میانگین دانشجویان به عنوان عملکرد دانشجوی مرزی تلقی نشود.[3]

در مورد تاثیر این مداخله، مطالعات تجربی مختلف توافق نظر ندارند. گفته می‌شود در این روش چون داوران در فضایی واقعی‌تر کار می‌کنند، راحت‌تر می‌توانند استاندارد را برآورد نمایند (کوزیمونو ۱۹۹۶). نتایج برخی از پژوهش‌ها نشان می‌دهد که با استفاده از نمرات واقعی، استاندارد واقع‌بینانه‌تر و پایین‌تر تعیین می‌شود (کریمر و همکاران[4] ۲۰۰۳). البته برخی از مطالعات هم این قضیه را زیر سؤال برده‌اند و این یافته را تأیید نکرده‌اند و توضیح داده‌اند که از نظر داوران، کاهش استاندارد یک ایستگاه همیشه بهترین راه حل نیست (شونیم–کلین و همکاران[5] ۲۰۰۹). بر اساس یک مطالعه در این روش توافق بین داوران افزایش می‌یابد (بوش و یگر[6] ۱۹۹۰). بر اساس یک مطالعه دیگر، داوران پس از دیدن نمرات واقعی، در ۲۵ درصد موارد برآورد خود را تغییر دادند و تغییرات اعمال شده به ازای هر آیتم بسیار کوچک بودند. علاوه برآن معمولاً آیتم‌هایی تغییر پیدا کردند که در ابتدا تخمین اولیه داوران در مورد آنها بیش از حد بالا یا پایین بوده است (نورسینی و همکاران ۱۹۸۸). بر اساس نتایج متاآنالیزی که توسط هرتز و همکاران انجام شد، استاندارد پس از بررسی نمرات واقعی کاهش پیدا کرد. در توضیح این امر گفته شده که احتمالاً داورها قبل از دیدن عملکرد واقعی دانشجویان، احساس می‌کردند سؤالات آسان بوده‌اند (هرتز و آئروباخ ۲۰۰۳). نتایج مطالعه روی OSCE پیش کارورزی دانشگاه علوم پزشکی تهران نشان

1. Leary and Kowalski
2. Reality check
3. رجوع کنید به فصل اول
4. Kramer
5. Schoonheim-Klein
6. Busch and Jaeger

داد که استاندارد پس از بررسی نمرات واقعی هم در روش انگوف تغییریافته و هم انگوف سه‌سطحی افزایش یافته است و منجر به کاهش میزان قبولی دانشجویان شده است (جلیلی و همکاران ۲۰۱۱، مرتاض هجری و همکاران ۱۳۹۰).

- **روش آنگوف بله/خیر:** در روش آنگوف بلی/خیر[1] پیشنهاد شد که داوران برای این که تعیین کنند چقدر احتمال دارد یک فراگیر مرزی بتواند به یک سؤال/آیتم پاسخ صحیح بدهد، به جای بیان عددی بین صفر تا ۱۰۰، به صورت بله/خیر جواب دهند. در حقیقت به صورت ساده به این پرسش پاسخ دهند که آیا از نظر آنها فراگیر مرزی می‌تواند به این سؤال/آیتم جواب بدهد یا نه. این روش به خاطر سهولتی که در کار ایجاد می‌کرد، برای تعیین استاندارد امتحانات پزشکی هم مورد استقبال قرار گرفت. برای محاسبه عدد استاندارد، برای پاسخ «بله» نمره ۱۰۰ (یا یک) و برای پاسخ «خیر» نمره صفر در نظر گرفته می‌شود (جدول ۴-۲۸).

جدول ٤-۲۸: تعیین استاندارد یک آزمون با ۶ سؤال به روش انگوف بله/خیر با شرکت ۵ داور

سؤال	داور ۱	داور ۲	داور ۳	داور ۴	داور ۵	استاندارد
سؤال ۱	بله	بله	خیر	خیر	خیر	۴۰
سؤال ۲	بله	خیر	خیر	بله	خیر	۴۰
سؤال ۳	خیر	خیر	خیر	بله	بله	۴۰
سؤال ۴	بله	خیر	خیر	خیر	بله	۴۰
سؤال ۵	بله	بله	خیر	بله	بله	۸۰
سؤال ۶	خیر	خیر	خیر	خیر	خیر	۰
کل						۴۰

یک مطالعه که به بررسی اثر این مداخله پرداخت نشان داد که استانداردی که با روش بله/خیر تعیین شد، مشابه استاندارد حاصل از روش آنگوف اصلی بود. محققان ادعا کردند که اگرچه یک داور ممکن است در پیش‌بینی عملکرد یک فرد خیلی دقیق عمل نکند اما وقتی داوران به صورت یک گروه گرد هم می‌آیند قادر به تولید دقیق نمره حدنصاب خواهند بود. از طرفی نویسندگان این احتمال را مطرح کردند که چون از داورها خواسته شده بود در یک جلسه به هر دو روش، استاندارد را تعیین کنند، این امر موجب خطا شده باشد (ایمپارا و پلاک ۱۹۹۷). در مطالعه دیگری پژوهشگران با مقایسه دو روش به طور جداگانه به این نتیجه رسیدند که روش بلی/خیر خیلی آسان‌تر است اما به شدت تحت تأثیر داده‌های واقعی قرار می‌گیرد (چین و هرتز ۲۰۰۲). در سال ۲۰۰۸ مطالعه‌ای انجام شد که در آن برای تعیین استاندارد از آنگوف بله/خیر استفاده کردند و به این نتیجه رسیدند که اگرچه کار با این روش بسیار آسان است اما به علت ماهیت دوتایی پاسخ، داور مجبور می‌شود حتی در مواردی که مطمئن نیست، یکی از دو حالت را انتخاب کند. نویسندگان ذکر کردند که اگرچه در یک آزمون تعداد آیتم‌های حد مرزی - یعنی آن آیتم‌هایی که داور ممکن است در برابر آنها با ابهام مواجه شود و برای هر یک از پاسخ‌های «بله» یا «خیر»، احتمال مساوی یا همان ۵۰ درصدی در ذهن داشته باشد- نادر است، اما باز هم می‌تواند به بروز خطای سیستماتیک و تورش منجر گردد. در واقع در این حالت داوری که ذاتاً سخت‌گیر است، در این گونه موارد صفر می‌دهد و داوری که خیلی سخت‌گیر نیست، ممکن است پاسخ یک بدهد (یودکوفسکی و همکاران[2] ۲۰۰۸).

- **روش انگوف سه‌سطحی:** با توجه به محدودیت‌های روش بله/خیر، یودکوفسکی و همکاران روش آنگوف سه‌سطحی[3]

1. Yes/no Angoff
2. Yudkowsky et al
3. Three level Angoff method

را پیشنهاد کرد که در آن سه پاسخ «بله»، «خیر» و «شاید»[1] در نظر گرفته می‌شود. با این روش، داور این امکان را دارد تا در مواجهه با آیتم‌های حد مرزی، گزینه «شاید» را انتخاب نماید. در این حالت پاسخ «بله» به معنای نمره ۱۰۰ است، برای پاسخ «شاید» نمره ۵۰ در نظر گرفته می‌شود، و پاسخ «خیر» یعنی نمره صفر به این معناست که از نظر داور، یک دانشجوی مرزی قادر نیست به این آیتم جواب صحیح بدهد (جدول ۵-۲۸). تفاوتی که در میزان استفاده داوران مختلف از «شاید» وجود دارد، می‌تواند به خاطر ویژگی‌های فردی آنها باشد و همچنین راحتی آنها در برابر تصمیمات سخت، میزان آشنایی آنها با دانشجویان مرزی و اطمینانی که به توانایی پیش‌بینی خود دارند (یودکوفسکی و همکاران ۲۰۰۸). مطالعاتی که به مقایسه نتایج روش انگوف سه‌سطحی با سایر روش‌های انگوف پرداخته‌اند، محدود هستند. یکی از این مطالعات در شرایط فرضی انگوف سه‌سطحی را با انگوف بله/خیر مقایسه کرده است (یودکوفسکی ۲۰۰۸). مطالعه دیگر، با استفاده از دو جلسه موازی، انگوف سه‌سطحی را با انگوف معمول که احتمالات را به صورت درصدی بیان می‌کند، مقایسه کرده است (جلیلی و همکاران ۲۰۱۱).

جدول ۵-۲۸: تعیین استاندارد یک آزمون با ۶ سؤال به روش انگوف سه سطحی با شرکت ۵ داور

سؤال	داور ۱	داور ۲	داور ۳	داور ۴	داور ۵	استاندارد
سؤال ۱	شاید	شاید	بله	شاید	شاید	۶۰
سؤال ۲	شاید	خیر	خیر	بله	خیر	۳۰
سؤال ۳	شاید	شاید	شاید	بله	بله	۷۰
سؤال ۴	شاید	خیر	خیر	خیر	شاید	۲۰
سؤال ۵	بله	بله	شاید	بله	شاید	۸۰
سؤال ۶	شاید	خیر	خیر	خیر	خیر	۱۰
کل						۴۵

یودکوفسکی ۲۰۰۸

در این پژوهش، انگوف سه سطحی همراه با استفاده از نمرات واقعی با انگوف بله/خیر مقایسه شد. پنج داور شامل یک رزیدنت ارشد و چهار عضو هیأت علمی شرکت کردند که همگی سابقه آموزش و کار با دانشجویان سال سه و چهار را داشتند. پس از آزمون، خلاصه‌ای از آزمون و case ها، چک‌لیست‌ها و داده‌های مربوط به عملکرد ۱۶۳ دانشجو در آزمون در اختیار داوران قرار گرفت و از آنها خواسته شد تا به صورت بلی/خیر/شاید برای هر آیتم تعیین کنند که «آیا یک دانشجوی مرزی می‌تواند به آن پاسخ صحیح بدهد». هر داور به طور مستقل این کار را انجام داد و به خاطر کمبود وقت، جلسه بحث برگزار نشد. مطابق مدل جبرانی، تمام آیتم‌ها روی هم ریخته شد و میانگین آنها محاسبه گشت. در ادامه، نویسندگان دو حالت دیگر انگوف را شبیه‌سازی کردند: انگوف بله/خیر سختگیرانه[1] و انگوف بله/خیر سهلگیرانه[2]. یعنی پژوهشگران امتحان کردند که اگر پاسخ‌های «شاید» با پاسخ‌های «خیر» و یا «بله» جایگزین می‌شد، نتایج چگونه تغییر می‌کرد.

یافته‌های مطالعه نشان داد که میانگین نمرات دانشجویان در کل ۶۳ (از ۱۰۰) بود. در مجموع پاسخ‌های تمام داوران، یعنی از ۶۰۵ آیتم، ۵۸ بار از گزینه «شاید» استفاده شده بود (۱۰ درصد) که در ایستگاه‌های مختلف از ۳ درصد تا ۱۳ درصد متفاوت بود. حدود نصف آیتم‌ها، حداقل از یک داور گزینه «شاید» دریافت کرده بودند. سه نفر از داورها اصلاً شاید به کار نبردند و یک داور در هر هفت ایستگاه از آن استفاده کرده بود؛ یعنی به طور متوسط در هر ایستگاه هفت بار. حدنصاب قبولی که از نظرات تمام داوران در تمام ایستگاه‌ها به دست آمد، ۴۳ (از ۱۰۰) بود. با در نظر گرفتن رویکرد سخت‌گیرانه و سهلگیرانه، نمره قبولی به ترتیب به ۴۷ و ۳۹ رسید. در حالی که با روش سه‌سطحی هیچ کس از آزمون رد نشد، در روش سخت‌گیرانه، میزان ردی یک درصد شد. البته نویسندگان اذعان داشتند که در کنار محدودیت‌هایی مانند کم بودن تعداد ایستگاه‌ها، کم بودن تعداد داوران و نبود جلسات بحث، مسأله دیگری هم وجود داشت و آن اینکه برخلاف آنچه که نویسندگان فرض کردند، در آنگوف بلی/خیر واقعی این‌طور نیست که همه موارد «شاید» به «بله» ختم شوند. بنابراین تفاوتِ دیده شده در این مطالعه ممکن است بیش از مقدار واقعی، تخمین زده شده باشد.

1. severe yes/no Angoff
2. lenient yes/no Angoff

1. Maybe

جلیلی و همکاران ۲۰۱۱

در این مطالعه نتایج دو روش تعیین استاندارد OSCE پیش کارورزی با یکدیگر مقایسه شدند. دو گروه جداگانه تشکیل شد که یکی با استفاده از انگوف معمولی و دیگری با استفاده از انگوف سه‌سطحی به تعیین استاندارد پرداختند. بحث و بررسی نمرات واقعی در هر دو گروه وجود داشت.

در هر دو روش استاندارد پس از بررسی نمرات واقعی نسبت به استاندارد پس از بحث افزایش یافته بود. استاندارد حاصل از روش سه‌سطحی بالاتر از استاندارد انگوف معمولی بود. همچنین، در روش انگوف سه‌سطحی، توافق بین داوران پایین‌تر (۰/۸۱ در برابر ۰/۹۴) و بازه اطمینان بازتر (۱۱/۲۹ در برابر ۳/۲۲ نمره) بود. نویسندگان نتیجه‌گیری کردند که روش انگوف معمول مخصوصاً با بحث و بررسی نمرات واقعی نتایج معتبرتری به دست می‌دهد.

محدودیت‌های روش انگوف

بر اساس نتایج مطالعات، هرچه تعداد داورها بیشتر باشد، نتیجه پایاتری در انگوف به دست می‌آید اما از نظر عملی افزایش تعداد داوران محدودیت ایجاد می‌کند. علاوه بر تعداد داورهای مورد نیاز، میزان وقت و در نتیجه هزینه‌ای که برای داوران صرف می‌شود، همیشه به عنوان دغدغه مطرح شده است (شونیم–کلین و همکاران ۲۰۰۹). از طرفی، کاری که از داوران خواسته می‌شود، پیچیده و دشوار است. در اجرای روش آنگوف، داوران معمولاً دو مشکل عمده دارند (بورسیکوت و رابرتز[1] ۲۰۰۶):

یکی در نظر گرفتن و تصور کردن خصوصیات دانشجوی مرزی که از بزرگ‌ترین چالش‌های پیش روی آنگوف است. داوران باید تصور کنند که دانشجو با حداقل توان‌مندی چه خصوصیاتی دارد و چگونه است (ایمپارا و پلاک ۱۹۹۷). این روند بیشترین نقد را متوجه روش آنگوف کرده است؛ به طوری که برخی این عمل را یک توانایی شناختی تقریباً غیرممکن خواندند و برخی دیگر آن را فراتر از قدرت ذهنی افراد برآورد کردند (ریکر ۲۰۰۶). پژوهش‌ها در این زمینه به نتایج متناقضی رسیده‌اند. نتایج یک مطالعه نشان داده است که داورها به خوبی توانستند فعالیت دانشجوی مرزی را پیش‌بینی کنند (پلاک و ایمپارا ۲۰۰۱). در حالیکه مطالعات دیگر این یافته را تأیید نکردند (نورسینی ۱۹۹۴، فهرمان و همکاران[2] ۱۹۹۱).

مسأله پیچیده دیگر علاوه بر تصور کردن مفهوم دانشجوی مرزی، تبدیل این مفهوم مبهم به عدد است. تخصیص دادن یک مقدار عددی به این مفهوم، پیچیده است (نورسینی ۱۹۹۳). همان‌طور که قبلاً ذکر شد، روش‌های انگوف بله/خیر و سه سطحی برای مقابله با این مشکل ایجاد شدند.

گفته شده که فضای مجازی و فرضی حاکم بر جلسه می‌تواند کار را پیچیده‌تر کند. مثلاً درک داور را از عملکرد دانشجوی مرزی تحت تاثیر قرار دهد و پیامدهایی در پی داشته باشد. به طوری که استاندارد بالای تعیین شده در روش آنگوف در برخی از مطالعات به این امر منتسب شده است. علت دیگری که در توضیح استاندارد بالای روش آنگوف عنوان می‌گردد، این است که اکثر داوران در رشته تخصصی خود سخت‌گیری می‌کنند و حد بالایی از توان‌مندی را برای دانشجویان لازم می‌دانند. به همین دلیل است که برخی استفاده از نمرات واقعی را برای تعدیل فضای فرضی جلسه انگوف مفید می‌دانند.

روش ابل

روش ابل[3]، یکی دیگر از روش‌های تعیین استاندارد معیاری و مبتنی بر آزمون است که دو مرحله‌ای است. از آنجا که داوران باید سطح دشواری و میزان اهمیت هر سؤال را تعیین کنند، توصیه می‌شود قبل از شروع فرایند، این دو مفهوم توضیح داده شوند و به بحث گذاشته شوند.

1. Boursicot and Roberts
2. Fehrmann et al
3. Ebel

مراحل روش ابل

- داور در مرحله اول، هر یک از سؤالات را از نظر میزان دشواری و میزان اهمیت آن بررسی می‌کند. برای دشواری، سه سطح (آسان، متوسط و سخت) و برای میزان اهمیت، چهار سطح (ضروری، مهم، قابل قبول و مشکوک[1]) در نظر گرفته می‌شود (جدول ۶-۲۸). در این مرحله، اگر ضریب دشواری سؤالات بر اساس عملکرد دانشجویان محاسبه شده باشد، می‌توان آن را برای اطلاع داور در اختیار وی قرار داد.

جدول ۶-۲۸: مرحله اول تعیین استاندارد ابل: مشخص کردن سطح دشواری و میزان اهمیت ۵ سؤال از ۱۰۰ سؤال یک آزمون فرضی

سؤال	دشواری			اهمیت			
	آسان	متوسط	سخت	ضروری	مهم	قابل قبول	مشکوک
سؤال ۱	❊					❊	
سؤال ۲	❊			❊			
سؤال ۳		❊			❊		
سؤال ۴			❊	❊			
سؤال ۵		❊					❊

- از داده‌های فوق، ۱۲ حالت ترکیبی (۴ × ۳) استخراج می‌شود. تعداد سؤالاتی که در هر حالت قرار گرفته‌اند، شمارش می‌شوند. به عنوان مثال، یکی از حالت‌ها متوسط/مهم خواهد بود که تعداد سؤالات آن در یک آزمون ۱۰۰ سؤالی، مثلاً ۲۲ تا است (جدول ۷-۲۸).

جدول ۷-۲۸: خلاصه مرحله اول تعیین استاندارد به شیوه ابل: مشخص کردن تعداد سؤالات در هر یک از ۱۲ حالت

	آسان	متوسط	سخت
ضروری	۴ سؤال	۳۰ سؤال	۵ سؤال
مهم	۶ سؤال	۲۲ سؤال	۱۰ سؤال
قابل قبول	۲ سؤال	۱۸ سؤال	۰
مشکوک	۰	۳ سؤال	۰

- در مرحله بعدی، داور احتمال پاسخگویی دانشجوی مرزی را برای هر یک از ۱۲ حالت فوق تعیین می‌کند. در واقع به این پرسش جواب می‌دهد که اگر دانشجوی مرزی با سؤالات در هر یک از این حالت‌ها مواجه شود، چقدر احتمال دارد که پاسخ درست بدهد (جدول ۸-۲۸). به عنوان مثال، یک داور این عدد را برای حالت متوسط/مهم، ۴۵ درصد در نظر گرفته است. یعنی به نظر او دانشجوی مرزی ۴۵ درصد احتمال دارد که بتواند یک سؤال متوسط و مهم را به درستی پاسخ دهد.
- برای تعیین استاندارد هر حالت، تعداد سؤالات هر حالت (محصول مرحله ۲) در احتمال اختصاص یافته به همان حالت

1. Questionable

(محصول مرحله ۳) ضرب می‌شود. به عنوان مثال برای حالت متوسط/مهم:

۲۲×۴۵%=۹/۹

جدول ۸-۲۸ : مرحله دوم تعیین استاندارد به شیوه ابل: مشخص کردن احتمال پاسخگویی یک دانشجوی مرزی به هر یک از حالات

	آسان	متوسط	سخت
ضروری	۸۰ درصد	۵۵ درصد	۳۰ درصد
مهم	۷۵ درصد	۴۵ درصد	۲۵ درصد
قابل قبول	۵۵ درصد	۴۵ درصد	۱۰ درصد
مشکوک	۲۰ درصد	۱۰ درصد	۵ درصد ۰

- مرحله ۴ برای تمام ۱۲ حالت تکرار می‌شود و تمام اعداد به دست آمده با هم جمع می‌شوند تا نمره حدنصاب به دست آید.

اگر قرار است بیش از یک داور در فرایند تعیین استاندارد به شیوه ابل شرکت کنند، عملاً همان روشی که برای ندلسکی و انگوف به کار رفت، در همین جا پیاده می‌شود اما کار کمی پیچیده‌تر است. چون فرایند داوری در دو مرحله صورت می‌گیرد، توصیه می‌شود که بحث بین داوران دو بار در پایان هر مرحله انجام شود. به صورت زیر:

- داوران انتخاب می‌شوند و از آنها برای شرکت در جلسه دعوت می‌شود.
- هر داور هر سؤال را به صورت انفرادی بررسی می‌کند و سطح دشواری و میزان اهمیت آن را تعیین می‌کند.
- بین داوران بحث مختصری انجام می‌شود. مسؤول جلسه در مورد هر سؤال از داوران می‌پرسد که به نظر چند نفر، سؤال آسان یا متوسط یا دشوار بوده است. سپس افراد به ارائه دلایل خود می‌پردازند.
- در ادامه، از آنها سؤال می‌شود که میزان اهمیت سؤال را چطور ارزیابی کرده‌اند و دوباره بحث صورت می‌گیرد.
- به داوران فرصت داده می‌شود تا در صورت تمایل نظر خود را عوض کنند. در اینجا هم مانند روش‌های انگوف و ندلسکی، هدف بحث، الزاماً قانع شدن داوران و رسیدن به توافق نیست.
- پس از اینکه همین روند برای تمام سؤالات انجام شد، مرحله بعدی آغاز می‌شود یعنی هر داور احتمال پاسخگویی دانشجوی مرزی را برای هر حالت تعیین می‌کند.
- داورها در مورد اعدادی که به هر حالت اختصاص داده‌اند، بحث می‌کنند و از دیدگاه خود دفاع می‌کنند. یک راه خوب برای درگرفتن بحث این است که سؤال شود چه کسی بیشترین یا کمترین احتمال را عنوان کرده است و از او خواسته شود تا دلایلش را برای جمع توضیح دهد.
- به داوران فرصت داده می‌شود تا در صورت تمایل نظر خود را عوض کنند و احتمالات را تغییر دهند. در حین فرایند، باید به صورت مدام به آنها یادآوری شود که در مورد احتمال پاسخگویی یک دانشجوی «مرزی» فکر کنند.
- اطلاعات جمع می‌شود و استاندارد حاصل از داوری هر نفر محاسبه می‌شود.
- برای ترکیب نظر داوران، به همان روش انگوف یا ندلسکی، میانگین یا میانه استانداردهای انفرادی محاسبه می‌شود.

محدودیت‌های روش ابل

این روش مانند سایر روش‌های مبتنی بر آزمون، نیازمند صرف وقت و انرژی توسط هیأت علمی است واز آنجا که پیچیده‌تر و دو مرحله‌ای است، زمان بیشتری می‌برد و کار دشوارتری است.

ملاحظاتی که برای انتخاب داوران و همچنین آموزش آنها در روش انگوف و ندلسکی گفته شد، در اینجا نیز باید رعایت شود.

روش گروه متمایز (متقابل)

گروه متمایز (متقابل)[1] یک روش تعیین استاندارد معیاری از نوع مبتنی بر آزمودنی است. یعنی حدنصاب قبولی پس از برگزاری امتحان و با بررسی عملکرد دانشجویان تعیین می‌شود. نکته حائز اهمیت این است که روش گروه متمایز هم برای سؤالات چندگزینه‌ای و هم تشریحی قابل استفاده است.

مراحل روش گروه متمایز (متقابل)

- پس از امتحان برگه تعدادی از دانشجویان به صورت تصادفی انتخاب می‌شود.
- هر داور در مورد هر دانشجو قضاوت می‌کند که او را در گروه قبول قرار دهد یا در گروه افتاده‌ها. برای این ارزیابی کلی مهم است که مقطع دانشجویان در نظر گرفته شود و با سطح خودشان قیاس شود، نه سطح دانشجویان سال بالاتر. این قضاوت کلی و ذهنی ممکن است به دو شکل انجام شود: گاهی از داوران خواسته می‌شود با توجه به تجربه قبلی خود در مورد عملکرد دانشجویان تصمیم‌گیری کنند. بدیهی است که در این حالت، داور حتماً باید برخورد قبلی با دانشجویان داشته باشد مثلاً مدرس دوره باشد. شکل دیگر بدین نحو است که تصمیم‌گیری بر اساس پاسخی که دانشجو به تمام سؤالات امتحان داده است انجام می‌شود. در این حالت مهم است که قضاوت بر اساس عملکرد کلی باشد نه بر اساس جمع نمرات دانشجو در آزمون.
- سپس تمام داوران در مورد آن دانشجو با هم بحث می‌کنند و به توافق می‌رسند.
- همین کار برای تمام دانشجویان منتخب تکرار می‌شود.
- در ادامه، نمرات واقعی این دانشجویان بر حسب پاسخ‌های درست آنها به سؤالات محاسبه می‌شود.
- عملکرد این دانشجویان بر حسب نمره‌ای که از امتحان کسب کرده‌اند و گروه رد/قبول خلاصه می‌شود (جدول ۹-۲۸).

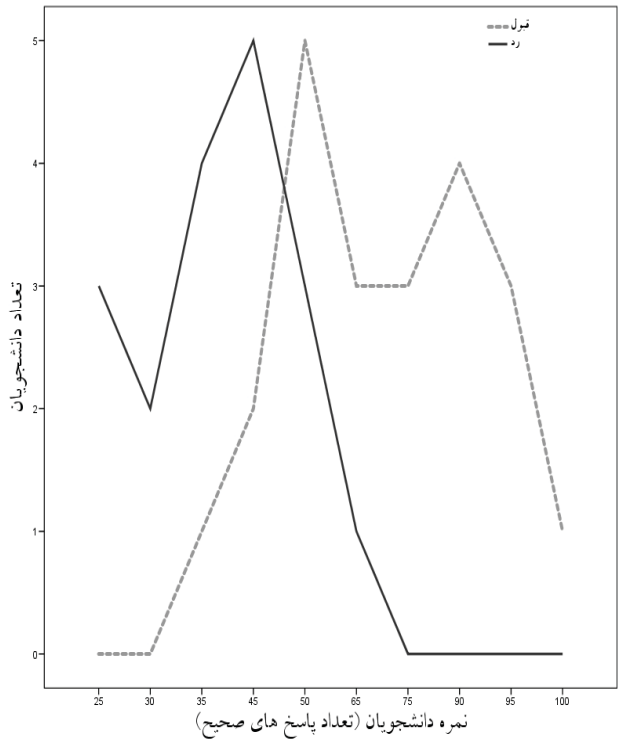
جدول ۹-۲۸ : تعیین استاندارد به روش گروه متمایز: انتخاب ۴۰ دانشجو به صورت تصادفی، قضاوت کلی در مورد عملکرد آنها به صورت رد یا قبول در گروه داوران و محاسبه نمره برگه هریک از آنها از نمره کل

نمره دانشجو	۲۵	۳۰	۳۵	۴۰	۵۰	۶۵	۷۵	۹۰	۹۵	۱۰۰
تعداد دانشجو در گروه رد	۳	۲	۴	۵	۳	۱	۰	۰	۰	۰
تعداد دانشجو در گروه قبول	۰	۰	۱	۲	۵	۳	۳	۴	۳	۱

- روی یک نمودار که محور x آن، پاسخ‌های صحیح و محور y آن، تعداد دانشجویان است، دو منحنی جداگانه برای هر یک از گروه‌های رد و قبول رسم می‌شود (شکل ۱-۲۸).
- در نهایت محل تلاقی دو منحنی به عنوان استاندارد در نظر گرفته می‌شود. به عنوان نمونه در مثال فوق، نمره حدنصاب حدود ۴۸ خواهد بود (شکل ۱-۲۸). با این وجود، معمول است که در تعیین نقطه برش، صرفاً به نقطه تلاقی اتکا نشود و هدف امتحان نیز در نظر گرفته شود: آیا می‌خواهیم آزمونی با موارد مثبت کاذب اندک داشته باشیم؟ آیا

1. Contrasting group method

افزایش تمایز بین گروه ضعیف و قوی مد نظر است؟ آیا کاهش موارد منفی کاذب مورد تاکید است؟ به عنوان مثال، در امتحانات اعطای مجوز پزشکی توصیه می‌شود که جهت تضمین کیفیت مراقبت‌های ارائه شده به بیماران، مقادیر مثبت کاذب بالا نباشند. در مثال فوق، اگر داوران بخواهند سختگیرانه عمل کنند تا آزمون مثبت کاذب نداشته باشد، باید نمره ۷۵ را به عنوان حدنصاب قبولی اعلام کنند. البته از آنجا که دو مقدار مثبت کاذب و منفی کاذب خلاف جهت هم حرکت می‌کنند، اگرچه منصفانه به نظر نمی‌رسد، ناگزیر مقادیر منفی کاذب، یعنی دانشجویان توانمندی که در آزمون رد اعلام می‌شوند، بالا می‌رود.



شکل ۱-۲۸: تعیین استاندارد به روش گروه متمایز: توزیع نمرات ۴۰ دانشجو به تفکیک گروه رد یا قبول

محدودیت‌های روش گروه متمایز (متقابل)

در روش گروه متمایز، همانند سایر روش‌هایی که نیازمند فرایند قضاوت سیستماتیک هستند، اینکه بتوان بعد از امتحان داوران را جمع کرد و از ایشان خواست تا برگه‌های دانشجویان را با این هدف بررسی کنند، با توجه به محدویت زمانی اعضای هیأت علمی، کار دشواری است.

در مورد اینکه برگه چه تعداد دانشجو باید بررسی شود، توافق وجود ندارد. برخی از منابع این تعداد را به تعداد سطوح عملکردی مرتبط دانسته‌اند و ذکر کرده‌اند در هر سطح عملکردی، ۱۰۰ دانشجو باید بررسی شوند (زیکی و همکاران ۲۰۰۶). ولی با این وجود، مطالعات در این خصوص کم هستند و نمی‌توان صرفاً به این مورد اتکا کرد.

روش گروه مرزی

روش گروه مرزی[1] یک روش معیاری و مبتنی بر آزمودنی است. یعنی حدنصاب قبولی، مشابه روش گروه متمایز، بر اساس عملکرد دانشجویان تعیین می‌شود با این تفاوت که فرایند داوری، حین برگزاری آزمون و با مشاهده مستقیم عملکرد دانشجویان صورت می‌پذیرد. این مسأله باعث شده است که روش گروه مرزی مخصوصاً برای امتحانات بالینی و عملی روش مناسبی باشد.

برای تعیین استاندارد به روش گروه مرزی، لازم است علاوه بر نمره‌ای که دانشجو از پاسخ به سؤالات گرفته است، عملکرد وی به طور کلی[2] توسط داور ارزیابی شود. اگر یک آزمون OSCE را در نظر بگیریم، به صورت معمول برای هر دانشجو در هر ایستگاه، چک لیستی توسط آزمونگر مستقر در ایستگاه تکمیل می‌شود. در اینجا لازم است آزمونگر علاوه بر پر کردن چک‌لیست و دادن ریزنمرات به یک مهارت، ارزیابی گلوبال خود را نسبت به عملکرد دانشجو در یک مقیاس لیکرت برآورد کند.

مقیاس لیکرت به شکل‌های گوناگون ممکن است مورد استفاده قرار بگیرد. مثلاً: ضعیف، مرزی، قابل قبول، خوب و عالی؛ که یک لیکرت پنج‌تایی است. به همین ترتیب، مقیاس خیلی ضعیف، ضعیف، رد مرزی، قبول مرزی، عالی، فراتر از انتظار؛ یک لیکرت شش‌تایی است.

دو مسأله باید برای تکمیل مقیاس لیکرت مورد توجه قرار بگیرد:

- ارزیابی گلوبال باید بدون توجه به نمره‌ای که دانشجو از چک‌لیست کسب کرده است، انجام شود. نمره چک‌لیست نباید به نمره ارزیابی کلی تبدیل شود. مثلاً برخی از آزمونگران برای راحتی با خود قرار می‌گذارند که اگر دانشجویی از ایستگاه ده نمره‌ای، چهار گرفت، در ازریابی کلی او را مرزی اعلام کنند. از این امر باید کاملاً اجتناب شود زیرا ارزیابی کلی باید مستقل از نمره چک‌لیست باشد.
- برای مشخص کردن جایگاه دانشجو در طیف عملکردی ضعیف تا عالی، مهم است که مقطعی که در آن امتحان برگزار می‌شود، مدنظر باشد و قضاوت طبق انتظاری که از دانشجویان همان مقطع می‌رود، صورت گیرد. به عنوان مثال، در امتحانی که برای دانشجویان کارآموز برگزار می‌گردد، اگر استادی که معمولاً با دستیاران تخصصی سر و کار دارد، این موضوع را در نظر نگیرد، سطح توقع بالایی دارد و عملکرد دانشجویان را کمتر از میزان واقعی برآورد می‌کند.

مراحل روش گروه مرزی در آزمون OSCE

- در زمان برگزاری OSCE، آزمونگر مستقر در هر ایستگاه، نتیجه مشاهده خود و ارزیابی دانشجو را به هر دو صورت تکمیل چک‌لیست و ارزیابی گلوبال (با لیکرت از قبل مشخص‌شده) ثبت می‌کند.
- پس از اتمام آزمون، دانشجویانی که در ارزیابی گلوبال هر ایستگاه، مرزی تشخیص داده شده‌اند، مشخص می‌شوند.
- نمره چک‌لیست مربوط به این دسته از دانشجویان (دانشجویان مرزی ایستگاه) محاسبه و استخراج می‌شود.
- میانگین نمرات چک‌لیست دانشجویان مرزی در یک ایستگاه محاسبه می‌شود که همان استاندارد ایستگاه است (جدول ۱۰). در برخی موارد ممکن است میانه نمرات چک‌لیست دانشجویان مرزی به عنوان استاندارد ایستگاه در نظر گرفته شود.
- میانگین استاندارد همه ایستگاه‌ها به عنوان استاندارد کل آزمون لحاظ می‌شود.

همان‌طور که در جدول ۱۰–۲۸ مشخص است در ایستگاه ریه، نمره گلوبال دانشجوی شماره ۴ و دانشجوی شماره ۱۶ برابر دو است. به عبارت دیگر، داور مستقر در این ایستگاه این دو دانشجو را از نظر عملکرد کلی به عنوان دانشجوی مرزی تشخیص داده است. نمره چک‌لیست این دو دانشجو به ترتیب ۳۸ و ۴۲ است که میانگین این دو نمره ۴۰ می‌شود. بنابراین حدنصاب ایستگاه معادل ۴۰ اعلام می‌شود.

1. Borderline group method
2. Global rating

جدول ۱۰-۲۸ : تعیین استاندارد سه ایستگاه OSCE از طریق روش گروه مرزی برای ۲۰ دانشجو (نمره مرزی در مقیاس کلی:۲)

	ایستگاه سردرد		ایستگاه ریه		ایستگاه قلب			
شما ر ه دانشجو	چک لیست	نمره کلی	چک لیست	نمره کلی	چک لیست	نمره کلی	نمره دانشجو	وضعیت قبولی
۱	۵۵	۳	۴۵	۳	۳۰	۱	۴۳/۳۳	۱
۲	۶۰	۴	۵۰	۳	۵۵	۱	۵۵	۱
۳	۷۰	۳	۶۶	۳	۱۴	۳	۵۰	۱
۴	۴۵	۴	۳۸	۲	۳۵	۱	۳۹/۳۳	۰
۵	۷۰	۴	۷۰	۴	۷	۳	۴۹	۱
۶	۸۵	۵	۶۵	۴	۵۷	۳	۶۹	۱
۷	۸۰	۵	۶۵	۳	۱۵	۱	۵۳/۳۳	۱
۸	۴۰	۱	۶۵	۳	۵۸	۴	۵۴/۳۳	۱
۹	۵۵	۳	۵۸	۳	۳۰	۲	۴۷/۶۷	۱
۱۰	۳۰	۱	۵۰	۳	۱۵	۱	۳۱/۶۷	۰
۱۱	۵۰	۴	۵۵	۳	۲۲	۳	۴۲/۳۳	۱
۱۲	۶۵	۳	۶۰	۴	۰	۱	۴۱/۶۷	۱
۱۳	۸۰	۵	۷۲	۴	۳۰	۱	۶۰/۶۷	۱
۱۴	۷۵	۴	۴۵	۳	۳۵	۱	۵۱/۶۷	۱
۱۵	۷۵	۵	۶۲	۳	۷۰	۴	۶۹	۱
۱۶	۵۵	۲	۴۲	۲	۱۵	۱	۳۷/۳۳	۰
۱۷	۸۰	۴	۷۰	۴	۴۲	۲	۶۴	۱
۱۸	۸۵	۴	۴۰	۳	۱۵	۲	۴۶/۶۷	۱
۱۹	۹۵	۵	۷۵	۴	۸	۱	۵۹/۳۳	۱
۲۰	۷۰	۳	۷۵	۴	۱۵	۳	۵۳/۳۳	۱
	حدنصاب ایستگاه ۱ ۵۵		حدنصاب ایستگاه ۲ ۴۰		حدنصاب ایستگاه ۳ ۲۹		حدنصاب کل ۴۱/۳۳	میزان ردی کل ۱۵ درصد

محدودیت‌های روش گروه مرزی

برای انجام روش گروه مرزی، توجه به چند نکته در مورد آزمونگران حائز اهمیت است. همان طور که قبلاً گفته شد، این روش، مبتنی بر آزمودنی است و در آن قضاوت تخصصی که برای تعیین استاندارد ضروری است، در حین آزمون صورت می‌گیرد. به عبارت دیگر، آزمونگران OSCE همان داوران متخصص هستند. بنابراین ضروری است که تمام ملاحظاتی که برای تعیین داوران در روش‌های دیگر مثلاً انگوف برشمرده شد، در اینجا هم رعایت شود. به عبارت دیگر، شاید آزمونگری که کم و بیش

آشنا به مطلب مورد آزمون است، برای تکمیل چک‌لیست کافی باشد اما برای اینکه ارزیابی گلوبال به درستی صورت گیرد، لازم است از آزمونگر متخصص و باتجربه که سابقه کار با دانشجویان همان مقطع را دارد، استفاده شود تا در تعیین طیف عملکردی و شناسایی دانشجویان مرزی به درستی عمل کند. همین موضوع، ممکن است محدودیت اجرایی ایجاد کند زیرا در بسیاری از آزمون‌های OSCE از رزیدنت‌ها، دانشجویان سال‌های بالاتر و بیماران استانداردشده به عنوان ارزیاب استفاده می‌شود که نمی‌توان از آنها انتظار داشت ارزیابی گلوبال را به صورت قابل اعتمادی انجام دهند. در یکی از مطالعاتی که در همین رابطه انجام شده، یکی از اشکالاتی که به روش مرزی وارد شده است، وقت‌گیر بودن آن برای اعضای هیأت علمی است (ویلکینسون و همکاران[1] ۲۰۰۱). در OSCEهایی که به طور معمول از اعضای هیأت علمی جهت ارزیابی استفاده می‌شود و حضور اعضای هیات علمی صرفاً به قصد تعیین استاندارد نیست، مشکلی از این نظر وجود ندارد.

در هر حال، حتی اگر آزمونگران، افراد متخصص و با تجربه باشند، قبل از آزمون، برگزاری جلسه توجیهی برای آشنایی آزمونگران با مقیاس لیکرت ضروری است. آنها معمولاً با چک‌لیست و نحوه تکمیل آن آشنایی دارند اما اگر قرار است برای اولین بار ارزیابی گلوبال انجام دهند، باید با مفهوم و کاربرد آن آشنا شوند و نکاتی که قبلاً در مورد ارزیابی کلی ذکر شد، برایشان روشن شود. همان‌طور که گفته شد، تاکید بر دو نکته مستقل بودن نمره گلوبال از چک لیست و لحاظ کردن مقطع شرکت‌کنندگان حین ارزیابی گلوبال حائز اهمیت است.

یکی از مشکلات روش گروه مرزی این است که اگر آزمونگر حین آزمون متوجه شود که به عده زیادی از دانشجویان نمره ضعیف داده یا آنها را رد کرده است، احتمال دارد به دانشجویان بعدی نمره بالاتری دهد. این مسأله روایی آزمون را مخدوش می‌کند و در واقع آزمون مبتنی بر معیار را به آزمون مبتنی بر هنجار تبدیل می‌کند (داویسون و بولاک ۲۰۰۷).

محدودیت دیگری روش گروه مرزی این است که احتمال دارد تعداد دانشجویانی که بر اساس لیکرت در گروه مرزی قرار گرفته‌اند، کم باشد و نتوان آنالیز را به درستی انجام داد (اسمی و بلک مور[2] ۲۰۰۱). مثلاً اگر در یک آزمون اصلاً هیچ دانشجوی مرزی وجود نداشته باشد، نمی‌توان استاندارد را تعیین کرد.

محدودیت دیگر، تعداد کل دانشجویان شرکت‌کننده در امتحان است. هرچند که در متون دقیقاً نیامده است که چه تعداد دانشجو برای محاسبه استاندارد به این روش لازم است اما ذکر شده است که روش گروه مرزی برای آزمونی با تعداد دانشجوی کم مناسب نیست. در یک مطالعه از ۶۱ نفر دانشجو در ایستگاه‌های مختلف، بین ۱۲ تا ۳۹ نفر به عنوان مرزی رتبه‌دهی شدند و پژوهشگران به این نتیجه رسیدند که همین تعداد دانشجو برای آنالیز مناسب بوده است (هامفری–مورتو و مک فادین[3] ۲۰۰۲). شونیم–کلین و همکاران با مقایسه نتایج مطالعه خود که روی ۱۱۹ دانشجو انجام شده بود، با دو مطالعه دیگر که با ۵۸ و ۸۹ نفر دانشجو نتایج مشابه‌ای به دست آورده بودند، نشان دادند که در روش مرزی، بازه اطمینان استانداردها تا حدی به تعداد دانشجویان وابسته است. این پژوهشگران نتیجه گرفتند که این روش با حداقل ۵۰ دانشجو قابل انجام است (شونیم–کلین و همکاران ۲۰۰۹، وود و همکاران ۲۰۰۶).

ویلکینسون و همکاران ۲۰۰۱

۱. در این مطالعه، برای سه OSCE در دانشگاه Otago در فواصل سال‌های ۱۹۹۷ تا ۱۹۹۹ که هر یک شامل ۱۸ ایستگاه ۵ دقیقه‌ای بودند، با روش گروه مرزی استاندارد تعیین شد. عملکرد دانشجویان در هر ایستگاه توسط یک چک‌لیست ۲۰ آیتمی و همچنین ارزیابی کلی (رد، مرزی، قابل قبول، بالاتر از حد انتظار) مورد ارزیابی قرار گرفت. در هر ایستگاه از مجموع نمره چک‌لیست دانشجویانی که مرزی تشخیص داده شده بودند، میانگین گرفته شد. سنجش روایی از طریق مقایسه با نمرات همین دانشجویان در یک آزمون کتبی چندگزینه‌ای (MCQ) و همچنین جمع نمرات آنها در هر دوره صورت پذیرفت. آزمون سال اول برای کسب تجربه بود و کسی رد نشد. در سال دو و سوم هر یک چهار نفر رد شدند. میزان عدم توافق آزمون‌گران در سال‌های اول، دوم و سوم به ترتیب ۱/۳ درصد، ۱/۴ درصد و ۲/۹ درصد بود.

1. Wilkinson et al
2. Smee and Blackmore
3. Humphery-Murto and Macfadyen

روش رگرسیون مرزی

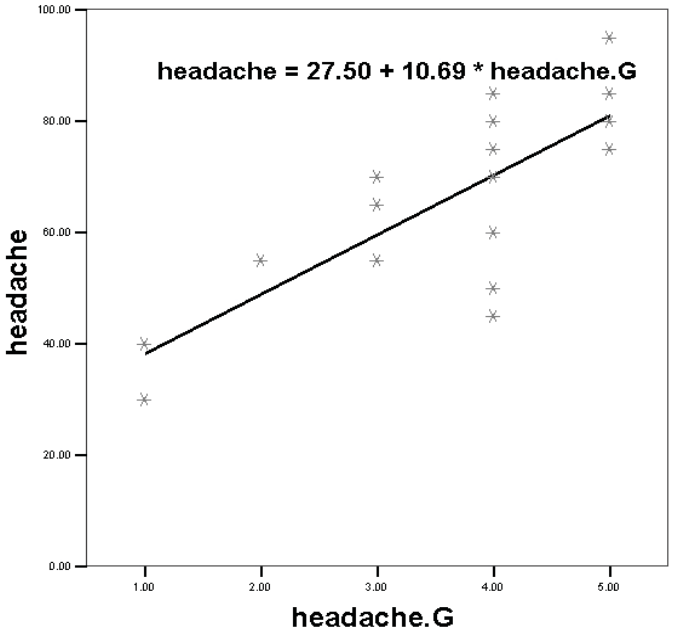
روش رگرسیون مرزی[1] یک روش معیاری و مبتنی بر آزمودنی است. از لحاظ مراحل اجرایی مشابه روش گروه مرزی است اما با این تفاوت که در هنگام محاسبه استاندارد، از نمره همه دانشجویان (نه فقط مرزی) استفاده می‌شود.

مراحل روش رگرسیون مرزی

- هر فراگیر توسط آزمونگر مستقر در ایستگاه به هر دو صورت چک‌لیست و ارزیابی گلوبال مورد سنجش قرار می‌گیرد.
- برای هر ایستگاه با در نظر گرفتن نمره چک‌لیست تمام دانشجویان به عنوان یک متغیر وابسته و نمره حاصل از ارزیابی کلی (در مقیاس لیکرت) آنها به عنوان یک متغیر مستقل، یک معادله رگرسیون به دست می‌آید.
- نمره چک‌لیست دانشجوی مرزی با قرار دادن عدد معادل مرزی در معادله پیش‌بینی می‌شود که به عنوان استاندارد ایستگاه محسوب می‌شود.
- میانگین استاندارد تمام ایستگاه‌ها به عنوان استاندارد کل آزمون اعلام می‌شود.

برای درک بهتر برای ایستگاه سردرد جدول ۱۰ از طریق روش رگرسیون مرزی، حدنصاب تعیین می‌گردد. در شکل ۲-۲۸ نمرات چک‌لیست دانشجویان در ایستگاه سردرد به عنوان متغیر وابسته وارد شده است و با در نظر گرفتن نمره گلوبال عملکرد دانشجویان در همین ایستگاه، نمودار رگرسیون رسم شده است و معادله رگرسیون مربوط به آن به دست آمده است. از آنجا که در این مثال، عدد ۲ در مقیاس لیکرت به عنوان دانشجوی مرزی در نظر گرفته شده بود، با جایگزین کردن عدد ۲ در معادله، پیش‌بینی می‌شود که نمره یک دانشجوی مرزی در این ایستگاه چیست. به این ترتیب، استاندارد ایستگاه به دست می‌آید.

$$۲۷/۵ + (۱۰/۶۹ \times ۲) = ۴۸/۸۸$$



شکل ۲-۲۸: توزیع نمرات چک لیست و ارزیابی کلی دانشجویان در ایستگاه سردرد

1. Borderline regression method

گفته می‌شود که روش رگرسیون مرزی نسبت به روش گروه مرزی ارجح است چرا که اولاً در آن از نمرات همه دانشجویان استفاده می‌شود و صرفاً به عملکرد تعداد محدودی از دانشجویان بسنده نمی‌شود. در ثانی، اگر دانشجویی به عنوان مرزی اعلام نشده باشد، مشکلی برای تعیین استاندارد پیش نمی‌آید و جای نگرانی نیست. چون با استفاده از معادله رگرسیون در واقع حتی اگر دانشجوی مرزی نداشته باشیم، نمره دانشجوی مرزی از روی نمرات سایر دانشجویان در رده‌های دیگر قابل پیش‌بینی است.

محدویت‌های روش رگرسیون مرزی

در این روش نیز مانند آنچه برای روش گروه مرزی گفته شد، لزوم توجه به انتخاب آزمونگران متخصص و آموزش آنها وجود دارد.

محدودیت دیگر که البته برای گروه مرزی هم صدق می‌کند، این است که گفته می‌شود چون تا حدی وابسته به عملکرد طیف دانشجویان حاضر در یک امتحان خاص است و بنابراین کاملاً مبتنی بر معیار نیست. تا جایی که مک‌ایلنمی و اور [1] معتقد بودند که نباید از این روش در پزشکی استفاده شود (مک ایلنمی و اور ۲۰۰۲).

وود و همکاران ۲۰۰۶

۱. در این مطالعه، در دانشگاه اوتاوا برای OSCE با ۱۰ ایستگاه روی ۵۹ دانشجو به دو روش رگرسیون مرزی و روش گروه مرزی تغییریافته استاندارد تعیین شد. در این مطالعه فقط ۸ ایستگاه که برخورد با بیمار داشت، بررسی شدند. ۲ ایستگاه دیگر سؤال کتبی داشتند و از مطالعه کنار گذاشته شدند. در روش گروه مرزی تغییریافته، بر اساس روتین قبلی دانشگاه از چک‌لیست زیر استفاده شد: Inferior، poor، excellent و good، borderline satisfactory، borderline unsatisfactory. استاندارد حاصل از روش رگرسیون به طور متوسط کمتر از گروه مرزی تغییریافته بود. در مورد تک تک ایستگاه‌ها هم، استاندارد ۶ ایستگاه در روش رگرسیون کمتر بود. بازه اطمینان در روش رگرسیون کوچک‌تر بود. درصد قبولی در روش رگرسیون ۴ درصد بیشتر به دست آمد.

روش بوک‌مارک

روش بوک‌مارک [2] روش نسبتاً جدیدی برای تعیین استاندارد است که در اوایل دهه ۱۹۹۰ معرفی شد. این روش، برای امتحاناتی به کار می‌رود که اساس تحلیل آنها، نظریه سؤال پاسخ [3] است. نامگذاری این روش از آنجا نشأت می‌گیرد که تمام سؤالات به ترتیب آسان تا سخت در یک دفترچه مرتب می‌شوند؛ به طوری که در هر صفحه فقط یک سؤال نوشته شده است. از داوران خواسته می‌شود به ترتیب سؤالات را بخوانند و هر جا به سؤالی رسیدند که فکر می‌کنند احتمالاً دانشجوی مرزی نمی‌تواند به درستی به آن جواب دهد، در دفترچه نشانه‌گذاری کنند.

مراحل روش بوک‌مارک

- در ابتدا تمام سؤالات بر اساس پارامتر دشواری، که با استفاده از نظریه سؤال پاسخ به دست آمده است، از آسان به سخت در یک دفترچه مرتب می‌شوند.
- داور از سؤال اول شروع می‌کند و به این پرسش جواب می‌دهد که آیا احتمالاً دانشجوی مرزی می‌تواند به این سؤال به درستی پاسخ دهد؟ اگر جواب منفی است، همان جا را نشانه‌گذاری می‌کند و اگر جواب مثبت است، باید به سراغ سؤال بعدی بروند تا به جواب خیر برسد.
- برای تعیین استاندارد آزمون توسط یک داور، سؤالی که توسط وی در دفترچه نشانه‌گذاری شده است، استخراج می‌شود و به داده‌های تحلیل آزمون بر مبنای نظریه سؤال–پاسخ رجوع می‌شود تا مشخص شود میزان توانایی مربوط

1. Mcillnemy & Orr
2. Bookmark
3. Item-Response Theory

به آن سؤال (θ) چقدر بوده است. سپس معادل این مقدار تتا در نظریه کلاسیک محاسبه و به عنوان حدنصاب آزمون اعلام می‌شود.

- در صورتی که تعیین استاندارد توسط چند داور انجام شد، داوران در مورد نظرات خود و جایگاهی که نشانه‌گذاری کرده‌اند، با یکدیگر بحث می‌کنند و در صورت تمایل نظر خود را عوض می‌کنند.
- برای ترکیب نظر داوران، میانگین (یا میانه) تتای به دست آمده از داوران مختلف محاسبه می شود و سپس معادل نمره آن در نظریه کلاسیک به عنوان حدنصاب اعلام می شود.

مزایای روش بوک‌مارک

کاری که داوران باید برای قضاوت و اجرای این روش انجام دهند، بسیار کمتر از روش‌هایی مانند انگوف و ندلسکی است. به همین دلیل روش بوک‌مارک نسبت به سایر روش‌ها کمتر وقت می‌گیرد. همچنین داده‌های کمتری در این روش تولید می‌شود که باعث می‌شود فرایند وارد کردن داده‌ها و محاسبات آنها آسان‌تر باشد.

مزیت دیگر روش بوک‌مارک این است که برای سؤالات تشریحی نیز با تغییراتی قابل اجرا است. در واقع اگر یک آزمون از هر دو نوع سؤال تستی و تشریحی تشکیل شده باشد، می‌توان با روش بوک‌مارک حدنصاب آن را تعیین کرد.

مزیت دیگر این روش این است که با استفاده از آن می‌توان بیش از یک استاندارد برای آزمون تعیین کرد. به عنوان مثال، چنانچه لازم باشد در یک آزمون دانشجویان با سطح ضعیف، قابل قبول و عالی جداگانه تعیین شوند، مراحل ذکر شده دو بار انجام می‌شود. یک بار برای تعیین مرز بین ضعیف و قابل قبول و بار دیگر برای تعیین مرز بین سطح قابل قبول و عالی.

روش Body of work

روش Body of work از لحاظ اصول کلی مشابه روش گروه متقابل است اما از اوایل دهه ۱۹۹۰ مورد استفاده قرار گرفته است و روش نوینی محسوب می‌شود. در واقع، روشی است که در آن داوران با دیدی جامع[1] همه سؤالات همه دانشجویان را مرور و ارزیابی می‌کنند. کاربرد اصلی این روش برای آزمون‌های تشریحی است و استفاده از آن برای سؤال چندگزینه‌ای دشوار است. همچنین این روش مانند روش بوک‌مارک می‌تواند در صورتی که لازم باشد تعیین استاندارد برای سطوح عملکردی مختلف (به عنوان مثل ضعیف، قابل قبول و عالی) انجام شود، این کار را انجام دهد.

مراحل روش Body of work

- تمام پاسخ‌های یک دانشجو در یک دفترچه جمع‌آوری می شود.
- پنج تا هشت دفترچه انتخاب می‌شوند. هر داور به تنهایی و به نوبت تک‌تک دفترچه‌ها را مطالعه می‌کند و بر اساس پاسخ‌هایی که دانشجو ارائه داده است، مشخص می‌کند که دانش و مهارت او مطابق کدام سطح عملکردی است. نکته قابل توجه اینجاست که نمره دانشجو به داوران گفته نمی‌شود.
- سپس داوران نظرات خود را با یکدیگر به اشتراک می‌گذارند و بحث می‌کنند. تا اینجا هدف، آموزش و تمرین داوران است.
- سپس حدود ۳۰ دفترچه در اختیار داوران قرار می‌گیرد. این بار هم داوران بدون آگاهی از نمره دانشجو، سطح عملکرد او به صورت کلی را از روی پاسخ‌هایی که داده است، می‌سنجند. مجدداً بحث بین داورها انجام می‌شود. این کار برای تمام دفترچه‌ها تکرار می‌شود تا تمام شوند.

1. Holistic

- توزیع نمرات دانشجویان بر اساس ارزیابی گلوبال داوران در یک جدول خلاصه می‌شود. هدف این است که نمره یا نمراتی که به بهترین شکل سطوح مختلف عملکردی را از هم جدا می‌کنند، مشخص شوند. برای درک بهتر، جدول ۱۱-۲۸ را مشاهده کنید. توزیع نمرات در این مثال نشان می‌دهد که نمرات ۴۶ تا ۶۰ برای سطوح عملکردی پایه و متوسط همپوشانی دارند. همچنین نمرات ۷۲ تا ۸۵ توسط داوران هم برای سطح متوسط و هم برای سطح پیشرفته در نظر گرفته شده است.

جدول ۱۱-۲۸ : توزیع نمرات دانشجویان و ارزیابی گلوبال دفترچه‌ها توسط داوران برای تعیین استاندارد به روش Body of work

نمره	ارزیابی کلی
۰ تا ۴۶	سطح پایه توسط همه داوران
۴۶ تا ۶۰	سطح پایه توسط برخی داوران سطح متوسط توسط برخی داوران
۶۱ تا ۷۱	سطح متوسط توسط همه داوران
۷۲ تا ۸۵	سطح متوسط توسط برخی داوران سطح پیشرفته توسط برخی داوران
۸۵ تا ۱۰۰	سطح پیشرفته توسط همه داوران

- برای رسیدن به نمره دقیقی که سطوح دارای همپوشانی را از یکدیگر افتراق دهد، در ادامه ۲۰ تا ۳۰ دفترچه که نمرات آنها در دامنه‌های مذکور قرار دارد انتخاب می‌شوند. به عنوان نمونه در مثال فوق، دفترچه‌هایی که نمره آنها در بازه ۴۶ تا ۶۰ است، جدا می‌شوند. از داوران خواسته می‌شود باز هم بدون آگاهی از نمره دانشجو، هر دفترچه را مجدداً ارزیابی کنند و در یکی از دو سطح مورد نظر (پایه یا متوسط) قرار دهند. مجدداً بحث بین داورها صورت می‌گیرد.
- اکنون هر یک از ۲۰ تا ۳۰ دفترچه، یک ارزیابی گلوبال (به صورت پایه یا متوسط) و یک نمره دارد. از اینجا به بعد مشابه روش گروه متقابل برای تعیین استاندارد اقدام می‌شود.

روش هافستی

روش هافستی[1] که در سال ۱۹۸۳ توسط ویلیام هافستی معرفی شد، مطلقاً در دسته استانداردهای معیاری یا مطلقاً در دسته استانداردهای هنجاری قرار نمی‌گیرد. در واقع، این روش ترکیبی از دو نوع استاندارد فوق است. این روش نیز در ابتدا برای آزمون چندگزینه‌ای استفاده شد اما بدون تغییر برای سایر آزمون‌ها مانند تشریحی یا عملی هم قابل استفاده است. در روش هافستی هر داور باید چهار عدد را تعیین کند که در ظاهر و نسبت به سایر روش‌ها شاید کار سخت‌تری به نظر برسد اما از آنجا که این برآورد برای کل آزمون است و نه تک تک سؤالات، فرایند داوری آسان‌تر است و وقت چندانی نمی‌گیرد.

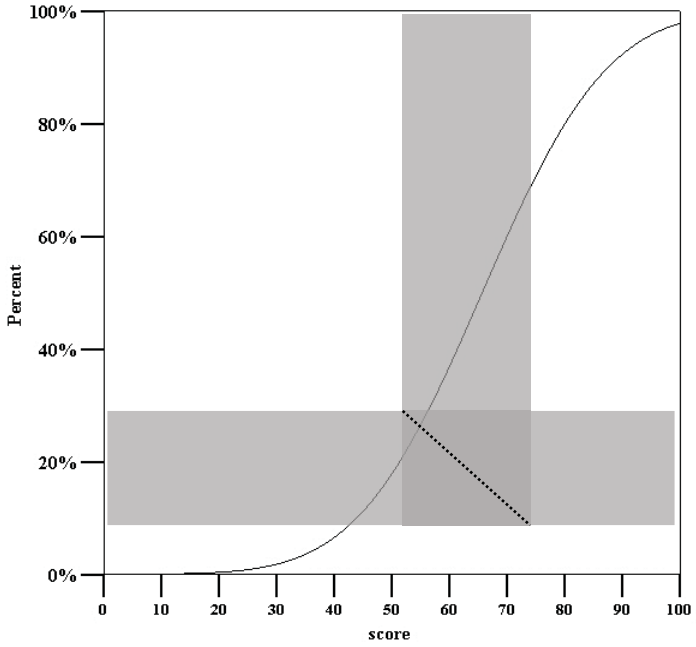
مراحل روش هافستی

- از هر داور خواسته می‌شود که موارد زیر را مشخص کند:
  - حداقل نمره حدنصاب قبولی که قابل قبول است (جنبه معیاری).
  - حداکثر نمره حدنصاب قبولی که قابل قبول است (جنبه معیاری).

1. Hofstee

- پایین‌ترین درصد ردی که قابل قبول است (جنبه هنجاری).
- بالاترین درصد ردی که قابل قبول است (جنبه هنجاری).

- داوران اعداد خود را به بحث می‌گذارند و در صورت تمایل نظر خود را تغییر می‌دهند.
- اعداد جمع‌آوری می‌گردد و میانگین نظر داوران مختلف حساب می‌شود (جدول ۱۲-۲۸).
- بر اساس مقادیر به دست آمده، نموداری رسم می‌شود که محور x آن نمره و محور y آن، میزان ردی است. از تقاطع این خطوط یک مستطیل ایجاد می‌شود که قطر آن رسم می‌گردد.
- نمودار تجمعی نمرات دانشجویان نیز رسم می‌شود.
- محل تقاطع قطر مستطیل با نمودار نمرات دانشجویان، حدنصاب آزمون است. در مثال زیر، نمره ۵۴ به عنوان حدنصاب در نظر گرفته می‌شود (شکل ۳-۲۸).



شکل ۳-۲۸: توزیع تجمعی نمرات دانشجویان و میانگین نظر داوران برای تعیین استاندارد به روش هافستی

جدول ۱۲-۲۸: نظرات داوران درمورد هر یک از مولفه های زیر برای تعیین استاندارد به روش هافستی

	داور ۱	داور ۲	داور ۳	داور ۴	داور ۵	میانگین نظر داوران
حداقل حدنصاب قبولی که قابل قبول است.	۵۵	۵۰	۴۵	۵۵	۶۰	۵۳
حداکثر حدنصاب قبولی که قابل قبول است.	۷۰	۸۰	۶۵	۷۵	۸۵	۷۵
پایین‌ترین میزان ردی که قابل قبول است.	۵	۱۰	۱۵	۱۰	۵	۹
بالاترین میزان ردی که قابل قبول است	۲۵	۲۵	۴۰	۳۰	۱۵	۲۷

محدودیت‌های روش هافستی

مطالعات بسیار اندکی در مورد روش هافستی و مقایسه آن با سایر روش‌ها انجام شده است اما به صورت کلی استفاده از این روش برای تعیین حدنصاب امتحانات سطح بالا چندان معمول نیست. یکی از محدویت‌های تاثیرگذار این روش، نبود برنامه کامپیوتری برای انجام محاسبات آن است. مراحلی که شرح داده شد، باید به صورت دستی انجام شود که موجب می‌شود احتمال خطای انسانی بالا باشد. یکی از کاربردهای این روش، رویکرد کمکی در تعیین استاندارد است که آن هم چندان توصیه نمی‌شود (چیزک و بانچ ۲۰۰۷). در رویکرد کمکی، استاندارد آزمون ابتدا با یکی از روش‌های مبتنی بر معیار مانند انگوف تعیین می‌شود و در انتهای جلسه، که داوران به مسائلی مانند هدف آزمون و سطح سؤالات اشراف پیدا کردند، روش هافستی نیز اجرا می‌گردد.

روش کوهن

نام روش کوهن شاید خیلی مشهور نباشد اما متد آن کم و بیش آشنا است. در واقع، در داخل کشور نیز تا سال‌ها مشابه این روش برای تعیین حدنصاب قبولی آزمون‌های جامع علوم پایه و پیش‌کارورزی مورد استفاده قرار می‌گرفت. اولین بار در سال ۲۰۱۰ کوهن–شوتانوس پژوهشگر هلندی به بیان تجربه خود در استفاده از این روش پرداخت و بعد از آن به همین نام شناخته می‌شود. محاسبه استاندارد در روش کوهن بسیار ساده است. از آنجا که نمره دانشجویان برتر را مبنا قرار می‌دهد، استاندارد به طیف شرکت‌کنندگان بستگی دارد و هنجاری است اما چون در عین حال ملاک ثابتی را نیز در نظر می‌گیرد، یک روش هنجاری و معیاری است.

مراحل روش کوهن

- نمرات دانشجویان محاسبه و از زیاد به کم مرتب می‌شود.
- نمره صدک ۹۵ یعنی ۵ درصد اول دانشجویان مشخص می‌شوند (جنبه هنجاری). به عنوان مثال در یک آزمون ۱۰۰ نمره‌ای، میانگین نمره صدک ۹۵، ۹۰ شده است.
- سپس ۶۰ درصد عدد فوق به عنوان استاندارد آزمون در نظر گرفته می‌شود (جنبه معیاری) که در مثال فوق، معادل نمره ۵۴ است. البته در روش کوهن اصلی برای تعیین استاندارد آزمون‌های چندگزینه‌ای، یک ضریب اعمال می‌شود که به منظور کاهش اثر حدس زدن است.

کوهن–شوتانوس و ون‌درولوتن [1] ۲۰۱۰

در این مطالعه جمعاً ۱۰۶ امتحان دو دانشگاه هلند طی ۹ سال بررسی شد. دانشگاه گرونینگن[2] از نمره ثابت ۰۶ برای تصمیم‌گیری در مورد رد و قبول دانشجویان استفاده کرده بود. در حالی که دانشگاه ماستریخ[3] با استفاده از روش کوهن استاندارد هر آزمون را محاسبه کرده بود. نتایج نشان داد در دانشگاه گرونینگن با اینکه نمره قبولی ثابت بود، میزان ردی در امتحانات مختلف بسیار متفاوت به دست آمد (از ۱۷ درصد تا ۹۷ درصد). در دانشگاه ماستریخت، نمره قبولی امتحانات مختلف متفاوت بود (از ۱۵ تا ۴۶) اما میزان ردی طی این سال‌ها تقریبا ثابت و حدود ۷۱ درصد بود. همچنین میانگین میزان ردی در امتحانات مختلف در دانشگاه گرونینگن ۵۳ درصد به دست آمد در حالی که این مقدار در دانشگاه ماستریخت ۱۷ درصد بود. به عبارت دیگر تعداد دانشجویان افتاده در سیستم نمره ثابت بسیار بیشتر بود.
در ادامه، چهار روش تعیین استاندارد برای امتحانات دانشگاه گرونینگن به کار رفت و نتایج مقایسه شد: نمره ثابت ۶۰، میانگین منهای یک انحراف معیار، ۶۰ درصد از بالاترین نمره و ۶۰ درصد صدک ۵۹ .
نتایج نشان داد که پایین‌ترین میزان قبولی از روش هنجاری (میانگین منهای یک انحراف معیار) و بالاترین میزان قبولی از روش معیاری (نمره ثابت ۶۰) به دست آمد. همچنین پایدارترین نتایج در میان آزمون‌های سال‌های مختلف از روش کوهن (۰۶ درصد صدک ۹۵) به دست آمد.

1. Cohen-Schotanus & van der Vleuten2. lenient yes/no Angoff
2. Groningen
3. Maastricht

محدودیت‌های روش کوهن

اساسی‌ترین نکته‌ای که در مورد روش کوهن گفته می‌شود این است که بیشتر در دسته استانداردهای هنجاری قرار می‌گیرد تا معیاری. این در حالی است که توصیه می‌شود برای آزمون‌های سنجش توانمندی مانند OSCE استاندارد هنجاری مناسب نیست و بهتر است از نوع معیاری استفاده شود. به عبارت دیگر، ضریب ۶۰ درصد اگرچه با نگاه به عدد ثابت حدنصاب آزمون‌ها در نظر گرفته شده است اما بر خلاف تصور، به عنوان یک ملاک عمل نمی‌کند. بلکه ضریبی برای یک هنجار است.

از طرف دیگر، ذکر شده است که خود عدد ۶۰ یک نمره قراردادی و ذهنی است و به نظر می‌رسد از آنجا که به صورت روتین حدنصاب آزمون های هلند ۵۵ تا ۶۰ درصد بوده است، به صورت دم دستی انتخاب شده است و برای یافتن مناسب‌ترین ضریب، نیاز به مطالعه است.

از آنجا که امتحانات مهم و سطح بالا باید مبتنی بر معیار باشند و همچنین بهتر است فرایند قضاوت در آنها به جای اینکه در مورد کل آزمون صورت گیرد، برای تک‌تک سؤالات انجا شود، استفاده از این روش در این امتحانات توصیه نمی‌شود. کاربرد روش کوهن می‌تواند به عنوان جایگزین نمره ثابت برای حدنصاب امتحانات نه چندان سطح بالا باشد.

یکی دیگر از انتقاداتی که به روش کوهن شده است، ضریبی است که در فرمول تعیین استاندارد آزمون چندگزینه‌ای برای تصحیح اثر حدس زدن و پاسخ تصادفی به کار رفته است. در واقع این مسأله همانند در نظر گرفتن نمره منفی برای سؤالات چندگزینه‌ای برای افزایش پایایی به کار رفته است اما همان‌طور که در بخش دوم کتاب مور بحث قرار گرفت، احتساب نمره منفی مخالفان جدی دارد. در اینجا نیز تلاش برای از بین بردن احتمال پاسخ حدسی کار چندان مقبولی نیست.

با توجه به این مسائل، روش کوهن تغییریافته توسط تایلر[1] در سال ۲۰۱۱ منتشر شد که به جای صدک ۹۵، صدک ۹۰ را در نظر گرفته است، تصحیح فرمول برای پاسخ حدسی را حذف کرده است و به جای ضریب ۶۰ درصد، آن را به روش دیگری محاسبه می‌کند.

مقایسه روش‌ها: توصیه‌های کلی

در این فصل انواع روش‌های تعیین استاندارد معرفی شدند. البته روش‌هایی که برای مشخص کردن نمره حدنصاب قبولی به کار می‌روند، تنها شامل این موارد نیست و شیوه‌های دیگری نیز وجود دارند. در هر حال، با مشاهده تنوع روش‌ها این سؤال مطرح می‌شود که در نهایت از کدام روش استفاده کنیم. مهم است که توجه داشته باشیم که این موضوع مهم جواب روشن و قطعی ندارد. البته این مطلب به این معنا نیست که اصلاً توصیه‌ای حتی به صورت کلی نیز نمی‌توان مطرح کرد اما بهتر است که دست‌اندرکاران، در هر شرایطی با توجه به امکانات و ویژگی‌های آزمون که خودشان شناخت بیشتری از آن دارند، تصمیم‌گیری کنند. آنچه در اینجا ذکر می‌شود، عمدتاً برگرفته از دستورالعمل مرکز سنجش آموزش[2] آمریکا برای تعیین استاندارد آزمون‌ها است. توجه به این امر ضروری است که این متن عمومی است و به صورت اختصاصی برای آزمون‌های گروه علوم پزشکی نوشته نشده است. بنابراین، اگر چه روش‌ها در کلیت و ماهیت مستقل از رشته آموزشی هستند، شاید تعمیم توصیه‌ها به حوزه آموزش پزشکی و کاربرد آنها برای تعیین حدنصاب آزمون‌های علوم پزشکی، نیاز به دقت و احتیاط بیشتر داشته باشد.

در مرحله اول، انتخاب روش بستگی به این دارد که شما به چه نوع قضاوتی باور و اطمینان دارید و همچنین چه نوع قضاوتی برایتان امکان‌پذیر و در دسترس است.

1. Tayler
2. Educational Testing Service (ETS)

اگر این امکان برای شما وجود دارد که داوران حین آزمون، عملکرد دانشجویان را به صورت مستقیم مشاهده کنند یا نمونه کار آنها را ببینند، استفاده از روش گروه متقابل در درجه اول توصیه می‌شود. این روش، مخصوصاً برای آزمون‌های تشریحی و آزمون‌های عملی مناسب است. برای آزمون چندگزینه‌ای، استفاده از روش گروه متقابل در صورتی توصیه می‌شود که مطمئن باشید داوران اولاً در قضاوت خود همان ویژگی‌هایی را مدنظر قرار خواهند داد که اساساً آزمون آنها را مورد سنجش قرار می‌دهد و ثانیاً به شدت مراقب خواهند بود سطح و مقطع دانشجویان را هنگام داوری در نظر بگیرند. از نظر نویسندگان دستورالعمل ETS، پایه‌های تئوری این روش نسبت به سایر روش‌ها مستحکم‌تر و منطقی‌تر است. در واقع، تنها روشی است که این امکان را ایجاد می‌کند تا میزان هر دو نوع خطای تصمیم‌گیری (یعنی مثبت کاذب و منفی کاذب) تخمین زده شود. البته چالش‌ها و سؤالت بی‌پاسخی در خصوص این روش وجود دارد: به عنوان نمونه سخت است که گفته شود حداقل چه تعداد برگه باید بررسی شوند زیرا عوامل متعددی در این موضوع موثر هستند؛ مشخص نیست اگر دانشجویان به اشتباه در دو گروه رد و قبول قرار گیرند، چه عواقبی دارد و موضوع تا چه حد جدی است؛ اینکه میزان توافق بین داوران در تقسیم‌بندی دانشجویان چقدر باید باشد و تعداد سطوح عملکردی بهتر است چندتا باشد نیز جواب روشنی ندارند.

اگر این امکان برای شما وجود ندارد که نظر چندین داور را در مورد عملکرد تعدادی از دانشجویان بگیرید، ولی می‌توانید یک داور را به این کار اختصاص دهید که عملکرد دانشجویان را به صورت مستقیم مشاهده کند و سطح آنها را ارزیابی کند، روش مرزی مناسب است. البته با توجه به نقائص روش گروه مرزی، روش رگرسیونِ مرزی پیشنهاد می‌شود.

این دستورالعمل همچنین توصیه میکند که اگر امکان مشاهده عملکرد دانشجویان اصلاً وجود ندارد، از یکی از روش‌های مبتنی بر آزمون استفاده کنید: ندلسکی، ابل، انگوف و بوک‌مارک. از هر یک از این روش‌ها که استفاده می‌کنید، پیشنهاد می‌شود که بررسی نمرات واقعی دانشجویان حتماً به آنها اضافه شود. استفاده از این روش‌ها مخصوصاً در مواردی که آزمون برای تعداد زیادی از افراد به منظور فارغ‌التحصیلی برگزار می‌شود، مناسب است.

توجه کنید که در روش‌های مبتنی بر آزمون، لزومی ندارد که همیشه جلسه تعیین استاندارد را قبل از آزمون برگزار کرد و می‌توان آن را به بعد آزمون موکول کرد. این کار دو مزیت دارد: یکی اینکه نمرات دانشجویان بعد از برگزاری آزمون آماده است و می توان به راحتی آنها را برای بررسی نمرات واقعی در اختیار داوران گذاشت و دیگر اینکه در این حالت، نگرانی درباره مسائل مربوط به امنیت آزمون و لو رفتن سؤالات نیز کمتر است.

اینکه از بین روش‌های مبتنی بر آزمون، کدام روش انتخاب شود، نیز بستگی دارد به نوع قضاوتی که می‌خواهید و می‌توانید داشته باشید: داوران در انگوف باید دانشجوی مرزی را تصور کنند و سپس احتمالی را برای تک‌تک سؤالات برآورد کنند. این کار از لحاظ ذهنی سخت است ولی این مزیت را دارد که ارتباطی با درجه دشواری سؤال ندارد. در حالی که در روش ابل داوران هم اهمیت و هم دشواری تک‌تک سؤالات را باید در نظر بگیرند و از این لحاظ زمان بیشتری باید به این روش اختصاص داد. در روش بوک‌مارک، تنها درجه دشواری سؤال باید بررسی شود. اگر داده‌های لازم در دسترس باشند، احتمالاً بوک‌مارک آسان‌ترین روش تعیین حدنصاب است زیرا لازم نیست هر داور تک‌تک سؤالات را بررسی کند.

توجه به این موضوع نیز حائز اهمیت است که نوع سؤال هم در انتخاب روش اثرگذار است: روش ندلسکی فقط می‌تواند برای آزمون چندگزینه‌ای استفاده شود که انتظار می‌رود در هر سؤال یک گزینه، بهترین پاسخ صحیح باشد اما مکرراً دیده می‌شود که سؤالی در ظاهر چندگزینه‌ای است اما از کلمات منفی استفاده کرده است یا در واقع، سؤال درست–نادرست یا سؤال ترکیبی را در خود دارد. در این صورت استفاده از روش ندلسکی با مشکل مواجه می‌گردد.

روش‌های انگوف و بوک‌مارک را هم برای سؤال چندگزینه‌ای و هم برای سؤال تشریحی می‌توان استفاده کرد. روش body of work نیز مخصوصاً برای سؤالات تشریحی مناسب است.

منابع

1. Angoff WH. Scales, norms and equivalent scores. In: Thorndike RL, ed. Educational Measurement. 2nd ed. Washington, DC: American Council on Education; 1971:508-600
2. Bandaranayake RC. Setting and maintaining standards in multiple choice examinations: AMEE Guide No.37. Medical Teacher 2008;30(9):836-845
3. Ben-David MF. AMEE Guide No. 18: Standard setting in student assessment. Medical Teacher 2000; 22(2):120-130
4. Boulet JR, de Champlain A and MCkinley D. Setting defensible performance standards on OSCEs and standardized patient examinations. Medical Teacher 2003;25(3):245–249
5. Boursicot K, Roberts T. Setting Standards in a Professional Higher Education Course: Defining the Concept of the Minimally Competent Student in Performance-Based Assessment at the Level of Graduation from Medical School. Higher Education Quarterly 2006;60(1):74–90.
6. Brandon PR. Conclusions about frequently studied modified Angoff standard-setting topics. Applied Measurement in Education 2004;17:59–88.
7. Brennan RL, Lockwood RE. A comparison of the Nedelsky and Angoff cutting score procedures using generalisability theory. Appl Psychol Measurement 1980;4:219–40
8. Busch JC, Jaeger RM. Influence of type of judge, normative information, and discussion on standards recommended for the National Teacher Examinations. Journal of Educational Measurement 1990;27:145-163.
9. Chinn RN, Hertz NR. Alternative approaches to standard setting for licensing and certification examinations. Applied Measurement in Education 2002;15:1-14.
10. Cizek GJ, Bunch MB. Standard setting: A guide to establishing and evaluating performance standards for tests. Thousand Oaks, CA: Sage Publications, Inc. 2007:20–22.
11. Cohen-Schotanus J, van der Vleuten C. A standard setting method with the best performing students as point of reference: Practical and affordable. Medical Teacher 2010;32:154–160
12. Cusinamo MD. Standard setting in medical education. Acad Med120–1996:71:112 .
13. Davison I, Bullock A. Evaluation of the Introduction of the Objective Structured Public Health Examination. The Research Office School of Education University of Birmingham, 2007
14. Fehrmann ML, Woehr DJ, Arthur W. The Angoff cutoff score method: The impact of frame-of-reference rater training. Educational and Psychological Measurement 1991;51(4):857-872.
15. Fowell SL, Fewtrell R, McLaughlin PJ. Estimating the minimum number of judges required for test-centred standard setting on written assessments. do discussion and iteration have an influence? Adv Health Sci Educ Theory Pract 2008;13(1):11-24.
16. Hambleton RK, Plake BS. Using an extended Angoff procedure to set standards on complex perfor-

mance assessments. Applied Measurement in Education 1995;8:41-55.

17. Hejri SM, Jalili M, Muijtjens AMM, van der Vleuten CPM. Assessing the reliability of the borderline regression method as a standard setting procedure for OSCE. Journal of Research in Medical Sciences (in press).
18. Humphrey-Murto S, Macfadyen JC. Standard setting: a comparison of case-author and modified borderline-group methods in a small-scale OSCE, Academic Medicine 2002;77(7):729–732.
19. Hurtz GM, Auerbach MA. A Meta-Analysis of the Effects of Modifications to the Angoff Method on Cutoff Scores and Judgment Consensus. Educational and Psychological Measurement 2003;63:584-601
20. Hurtz GM, Hertz NR. How many raters should be used for establishing cutoff scores with the angoff method? A generalizability theory study. Educational and Psychological Measurement 1999;59(6):885-897
21. Impara JC, Plake BS. Standard setting: An alternative approach. J Educ Meas 1997;34:353-66
22. Jalili M, M Hejri S, Norcini JJ. Comparison of two methods of standard setting: the performance of the three-level Angoff method. Medical Education 2011;45(12):1199–1208
23. Kramer A, Muijtjens A, Jansen K, Dusman H, Tan L, van der Vleuten C. Comparison of a rational and an empirical standard setting procedure for an OSCE. Objective structured clinical examinations. Medical Education 2003;2:132–139.
24. Leary MR, Kowalski RM. Impression management: A literature review and two component model. Psychological Bulletin 1990;107(1):34-47.
25. Maurer TJ, Alexander RA, Callahan CM, Bailey JJ, Dambrot FH. Methodological and psychometric issues in setting cutoff scores using the Angoff method. Personnel Psychology 1991;44:235-262.
26. Maurer TJ, Alexander RA. Methods of improving employment test critical scores derived by judging test content: A reviewand critiquee. Personnel Psychology1992;45:727-762
27. McKinley DW, Norcini JJ. How to set standards on performance-based examinations: AMEE Guide No. 85. Medical teacher 2013, 1–14
28. McIlhenny C, Orr G. Standard setting in an objective structured clinical examination: use of global ratings of borderline performance to determine the passing score. Medical Educcation 2002;36:388–95.
29. Norcini JJ. Research on Standards for Professional Licensure and Certification Examinations. Evaluation & the Health Professions 1994;17(2):160-177.
30. Norcini JJ. Setting standards on educational tests. Medical education 2003;37(5):464-9.
31. Norcini JJ, Lipner RS, Langdon LO, Strecker CA. A comparison of three variations on a standard-setting method. Journal of Educational Measurement 1987;24:56-64.
32. Norcini JJ, Shea JA, Kanya DT. The Effect of Various Factors on Standard Setting. Journal of Educational Measurement 1988;25(1):57-65

33. Norcini JJ, Stillman PL, Sutnick AI, et al. Scoring and Standard Setting with Standardized Patients. Evaluation & the Health Professions 1993;16(3):322-332.
34. Pell G, Fuller R, Homer M, Roberts T. How to measure the quality of the OSCE: A review of metrics. AMEE guide no. 49. Medical teacher 2010;32:802–811
35. Plake B.S and Impara J.C. Ability of panelists to estimate item performance for a target group of candidates: An issue in judgmental standard setting. Educational Assessment 2001; 7: 87-97.
36. Ricker KL. Setting Cut Scores: Critical Review of Angoff and Modified-Angoff Methods. Alberta Journal of Educational Research 2006;52(1):53-64
37. Rothman AI, Blackmore D, Dauphinee WD, Reznick R. The use of global ratings in OSCE station scores. Advances in Health Sciences Education 1996;1(3):215-219.
38. Schoonheim-Klein M, Muijtjens A, Habets L and et al. Who will pass the dental OSCE? Comparison of the Angoff and the borderline regression standard setting method. Eur J Dent Educ 2009;13:162–171
39. Smee SM, Blackmore DE. Setting standards for an objective structured clinical examination: the borderline group method gains ground on Angoff. Medical Education 2001;35(11):1009- 1010.
40. Stern, DT, Ben-David, MF, De Champlain, A, Hodges, B. Ensuring global standards for medical graduates: a pilot study of international standard-setting. Medical Teacher 2005;27(3):207–213

41. Taylor CA. Development of a modified Cohen method of standard setting. Med Teach. 2011;33(12):e678-82. doi: 10.3109/0142159X.2011.611192.
42. Van Der Vleuten C. Setting and maintaining standards in multiple choice examinations: Guide supplement 37.1. Medical Teacher 2010;32:174–176
43. Wilkinson TJ, Newble DI, Frampton CM. Standard setting in an objective structured clinical examination: use of global ratings of borderline performance to determine the passing score. Medical education 2001;35(11):1043-9.
44. Wood TJ, Humphrey-Murto SM, Norman GR. Standard setting in a small scale OSCE: a comparison of the Modified Borderline-Group Method and the Borderline Regression Method. Advances in health sciences education: theory and practice2)11;2006):115-22.
45. Yudkowsky R, Downing SM, Wirth S. Simpler standards for local performance examinations: The yes/no Angoff and whole test Ebel. Teach Learn Med 2008;20:212–217.
46. Yudkowsky R, Downing SM, Popescu Mihaela. Setting standards for performance tests: A pilot study of a three-level Angoff method. Academic Medicine 2008;83(10 suppl):13-16
47. Zieky M. Perie M, Livingston S. A Primer on Setting Cut Scores on Tests of Educational Achievement, Princeton, N.J: Educational Testing Service, 2006.

۴۸. مرتاض هجری سارا، جلیلی محمد، لباف علی. تعیین نمره حدنصاب قبولی آزمون عینی ساختارمند بالینی به روش انگوف و ارزیابی تأثیر بحث و بررسی نمرات واقعی. مجله ایرانی آموزش در علوم پزشکی ۱۳۹۰؛ ۱۱ (۷): ۸۸۵ تا ۸۹۴

۴۹. جلیلی محمد، مرتاض هجری سارا. مقایسه نمره حدنصاب و میزان قبولی شرکت کنندگان در آزمون عینی ساختارمند بالینی با استفاده از چهار روش مختلف تعیین استاندارد: نمره ثابت، آنگوف، رگرسیون مرزی و کوهن. مجله گام‌های توسعه در آموزش پزشکی۱۳۹۱ ؛ ۹ (۱): ۷۷ تا ۸۴

فصل ۲۹

ارزشیابی تعیین حدنصاب قبولی

ضرورت ارزشیابی

ارزشیابی تعیین استاندارد در دو مرحله مورد توجه قرار می‌گیرد: یکی در ابتدای فرایند به منظور انتخاب روش مناسب و دیگری، در انتهای فرایند به دلیل اطمینان یافتن از درستی نتایج حاصله. در مطالعات مربوط، معمول است که برای ارزشیابی تعیین استاندارد، به این موارد پرداخته شود:

- نمره حدنصاب محاسبه‌شده
- میزان قبولی به دست‌آمده
- اعتبار یا روایی روش
- ثبات یا پایایی روش
- قابلیت اجرای روش

اگر قرار باشد به عنوان مسؤول تعیین استاندارد در ابتدای فرایند و برای انتخاب روش مناسب، به نتایج ارزشیابی مطالعات دیگر اتکا کنیم، باید به دنبال اطلاعاتی حول شاخص‌های فوق باشیم و همچنین اگر قرار است به عنوان مسؤول تعیین استاندارد، در انتهای روشی که اجرا کرده‌ایم، به ارزشیابی آن بپردازیم، در نظر گرفتن چارچوب فوق برای ارزشیابی ضروری است.

ضرورت توجه به ارزشیابی در ابتدای تعیین استاندارد

در فصل گذشته، انواع روش های تعیین استاندارد معرفی شدند و مراحل کار برای تک‌تک آنها شرح داده شد. اکنون باید به این سؤال جواب داد که نهایتاً بهترین روش برای تصمیم‌گیری در مورد رد و قبول دانشجویان کدام است. قطعاً برای پاسخ به این سؤال، باید به منابع و شواهد مراجعه نمود. حقیقت این است که پژوهش‌های فراوانی برای بررسی نقاط قوت و ضعف روش‌های تعیین استاندارد صورت گرفته است و مطالعات مختلف به ارزشیابی متدهای گوناگون تعیین استاندارد، چه برای آزمون‌های کتبی و چه برای آزمون‌های عملکردی، پرداخته‌اند اما همان‌طور که قبلاً ذکر شد، در مورد ارجحیت روش‌ها نسبت به یکدیگر و همچنین مقایسه پایایی و اعتبارشان به نتیجه قطعی نرسیده‌اند و کماکان گفته می‌شود که هیچ روش استاندارد طلایی که برای استفاده همگان توصیه شود، وجود ندارد (زیکی و همکاران ۲۰۰۶، وود و همکاران ۲۰۰۶، بارمان ۲۰۰۸).

بنابراین، مسؤولان امر ارزیابی دانشجویان با توجه به عوامل مختلفی که در روند تعیین استاندارد تاثیرگذار هستند و همچنین پیامدهای گوناگونی که ممکن است از فرایند تعیین استاندارد حاصل شوند، باید تلاش کنند تا بهترین روش را برای آزمون خود با شرایط ویژه‌ای که دارد، انتخاب کنند. به همین دلیل ضروری است تا قبل از شروع کار و در مراحل نخستین، به نتایج ارزشیابی روش‌ها که از تجربه سایر دانشگاه‌ها حاصل شده ست، توجه شود.

ضرورت توجه به ارزشیابی در انتهای تعیین استاندارد

همان‌طور که در فصل اول گفته شد، آخرین مرحله در تعیین استاندارد، ارزشیابی است زیرا تعیین استاندارد به هر روشی که انجام شود، در انتها سؤالاتی از این دست برای دست‌اندکاران و دانشجویان مطرح می‌شود: آیا نمره حدنصاب به درستی تعیین شده است، آیا نتایج منطقی و قابل پذیرش هستند، آیا واقعاً دانشجویان توانمند، قبول شده‌اند و آیا دانشجویانی که لایق نمره قبولی نبوده‌اند، به درستی شناسایی شده‌اند؟ بنابراین، حتماً لازم است که حداقل بعد از اجرای بار اول فرایند، تلاش نظام‌مندی صورت گیرد تا با پاسخ به این سؤالات با ادله محکم و قوی به دفاع از روش مورد استفاده پرداخت و ثانیاً روند تعیین استاندارد را برای دفعات بعدی بهبود بخشید. برخی از کارکردهای ارزشیابی تعیین استاندارد، شامل موارد زیر هستند:

- ارزشیابی اگر درست انجام شود، با نشان دادن روند قابل دفاع تعیین استاندارد، شواهدی از اینکه قضاوت و عملکرد داوران «معتبر» بوده است، ارائه می‌کند.
- اگر پس از جمع آوری و تعیین استاندارد، آنالیزهای تکمیلی انجام شود، می‌توان با نشان دادن میزان واریانس بین استانداردها شواهدی در مورد «ثبات» روش ارائه کرد.
- ارزشیابی با جمع‌آوری داده‌های مرتبط می‌تواند «تاثیر» مداخلات گوناگون مانند برنامه‌های توجیهی و آموزشی، بحث بین داوران، بازخورد به آنها و همچنین ارائه نمرات واقعی را نشان دهد.
- با نظرسنجی و پرسش از داوران می‌توان میزان اطمینان داوران نسبت به خودشان طی فرایند تعیین استاندارد را استخراج کرد که به عنوان شاخصی از «اعتبار» روش برای دفاع از روش به کار می‌رود.
- با نظرخواهی از داوران و گرفتن بازخورد از آنها می‌توان در مورد نحوه «بهبود» فرایند، اطلاعات و نظرات مفیدی کسب کرد.

برای دستیابی به کارکردهای فوق ضروری است از آغاز، برای ارزشیابی برنامه‌ریزی شود. باید داده‌های مربوط به تعیین استاندارد از جمله نمره‌ها و درصدهایی که داوران مشخص می‌کنند، در هر مرحله به صورت مرتب و منظم و مستمر جمع‌آوری و ثبت گردد و همچنین در انتها، جلسات نظرخواهی از داوران برگزار شود یا از طریق پرسشنامه نظرات آنها گردآوری شود.

در این فصل سعی می‌کنیم به صورت خلاصه به ابعاد مختلف ارزشیابی تعیین استاندارد بپردازیم.

اعتبار روش تعیین استاندارد

یکی از شاخص‌هایی که در هر اندازه‌گیری مهم است، این مسأله است که اندازه‌گیری تا چه میزان روا[1] بوده است. البته در این موضوع که آیا تعیین استاندارد را می‌توان یک نوع اندازه‌گیری دانست یا خیر، جای بحث وجود دارد که در قسمت بعدی یعنی پایایی روش تعیین استاندارد به آن می‌پردازیم. در اینجا در مورد این موضوع صحبت خواهیم کرد که صاحب‌نظران در مورد روایی[2] تعیین استاندارد چه می‌گویند. از آنجا که مفهوم روایی در بحث ارزیابی فراگیر و امتحانات به ذهن نزدیک‌تر است، قبل از ورود به بحث روایی در تعیین استاندارد، مختصری به روایی آزمون می‌پردازیم.

روایی آزمون به صورت کلی به این مسأله می‌پردازد که واقعاً چقدر همان چیزی که مدنظر بوده است، سنجیده شده است. همان‌طور که در بخش‌های پیشین کتاب ذکر شد، برای روایی، انواع مختلفی چون روایی محتوایی[3]، روایی صوری[4]، روایی سازه[5] و ... در نظر می‌گیرند. ما در اینجا بحث خود را به روایی محتوایی محدود می‌کنیم. به عنوان مثال، آزمون پایان ترم درس جراحی برای دانشجویان پزشکی به صورت ۵۰ سؤال چندگزینه‌ای طراحی شده است. سؤال این است که روایی آزمون چگونه بوده است یعنی چقدر همان چیزی را سنجیده است که قرار بوده بسنجد. برای پاسخ به این سؤال، باید بدانیم

1. Valid
2. Validity
3. Content validity
4. Face validity
5. Construct validity

که هدف از ارائه درس جراحی برای دانشجویان پزشکی چه بوده است و انتظار می‌رفته که دانشجویان با گذراندن آن به چه توانمندی‌هایی رسیده باشند. اگر سؤالات را با اهداف درس مطابقت دهیم، قاعدتاً می‌توانیم به این سؤال جواب دهیم که هر سؤال تا چه اندازه مطابق با اهداف درس طراحی شده است. همان‌طور که می‌دانیم، آزمون‌های کتبی تنها قادر به سنجش اطلاعات و معلومات در حیطه شناختی هستند و قادر به ارزیابی مهارت‌های عملی نیستند. ممکن است ۵۰ سؤال فوق به خوبی اطلاعات و محفوظات و حتی نحوه استدلال دانشجویان را در برخورد با بیماران مبتلا به مشکلات جراحی ارزیابی کرده باشند اما اگر مواردی مثل انجام پروسیجر، ارتباط با بیمار، گرفتن شرح حال و معاینه فیزیکی هم در اهداف دوره بوده‌اند، روایی آزمون زیر سؤال می‌رود. زیرا توانایی دانشجو در انجام معاینه باید ارزیابی می‌شده است که نشده است. پس به صورت کلی، سؤال اصلی در روایی این است که چقدر همان چیزی که قرار بوده است، اندازه‌گیری شده است.

هر چند به کار بردن کلمه روایی برای تعیین استاندارد در متون معمول است، برخی معتقد هستند کار چندان صحیحی نیست و به جای آن، از لفظ «اعتبار»[1] استفاده می‌کنند.مسأله مهم برای اطمینان از روایی، «روند قابل دفاع تعیین استاندارد» است. یعنی مراحلی که طی شده است و فرایندی که برای تعیین نمره حدنصاب صورت پذیرفته است، تا چه حد منطقی و معقول بوده‌اند. مسؤول تعیین استاندارد باید تمام مراحل کار را مستند کند، اطلاعات کافی در مورد نحوه کار در اختیار مخاطبان بگذارد و در مورد روندی که منجر به تعیین نمره شده با ارائه مستندات و مدارک به صورت کامل و بدون دخالت نظر شخصی خود توضیح دهد. سپس خود مخاطبان ارزشیابی تصمیم می‌گیرند که متقاعد شده‌اند یا خیر. دلیل این امر هم به این بر می‌گردد که همه بر سر این موضوع توافق دارند که ماهیت تعیین استاندارد، قضاوتی است. یعنی در نهایت قضاوت غلط و درست نداریم. بلکه این که قضاوت چطور انجام شده است مهم است. مثلاً آیا فقط نظر یک شخص ملاک قرار گرفته است یا جمعی از داوران در تصمیم دخیل بوده است؟ آیا قضاوت با بررسی شواهد و مدارک و اطلاعات مرتبط صورت گرفته است یا به شکل یک حکم کلی و قراردادی، از پیش تعیین شده است؟ همان‌طور که مشخص است استانداردی که با استفاده از نظر تخصصی چندین نفر و با در نظرگرفتن فرصت بحث و بررسی نمرات واقعی تعیین شده است، روند منطقی‌تری دارد تا در نظر گرفتن نمره ثابت ۱۲ برای تمام امتحانات.

به صورت خلاصه، اولاً، قبل از شروع تعیین استاندارد، باید مسیری را در پیش گرفت که اعتبار تا حدی زیاد تامین شود که شامل انتخاب روش مناسب، انتخاب داوران معتبر، مستندسازی مراحل و ... است. ثانیاً، پس از اتمام فرایند تعیین استاندارد لازم است اعتبار روش سنجیده شود تا مطمئن شویم نتایج منطقی حاصل شده است. برای سنجش اعتبار روش‌ها، راه‌هایی پیشنهاد شده است که مجموعاً می‌توان به موارد زیر اشاره نمود (بولت و همکاران ۲۰۰۳، نورسینی ۲۰۰۳):

- بررسی و مقایسه نحوه عملکرد دانشجویان در آینده، مثلاً بررسی مهارت و توانمندی کسانی که در این آزمون با این روش قبول یا رد شده‌اند.
- مقایسه میزان قبولی با سایر آزمون‌های همان دانشجویان
- مقایسه میزان قبولی با میزان قبولی حاصل از سایر روش‌های تعیین استاندارد؛ هرچند انتظار می‌رود روش‌های مختلف نمرات مختلفی را به عنوان حدنصاب تعیین کنند، این مقایسه اطلاعات خوبی در اختیار دست‌اندرکاران می‌گذارد.
- مقایسه نتیجه کار داوران مختلف، مثلاً از طریق برگزاری جلسات موازی با افراد متفاوت
- نظرسنجی از خود داوران در خصوص میزان اطمینانی که به فرایند داوری و قضاوت خود داشتند یا در خصوص اثربخشی مواردی مانند بحث و بررسی نمرات واقعی
- نظرسنجی از دست اندرکاران اجرایی در مورد مقبول بودن میزان ردی

هیچ یک از راهای فوق به تنهایی اعتبار روش تعیین استاندارد را ثابت نمی‌کنند اما در کنار هم می‌توانند تصویر قابل قبولی از میزان تلاش دست‌اندرکاران برای اقزایش اعتبار روش به دست دهند و به ذی‌نفعان اطمینان بخشند که نتایج به دست‌آمده، قابل‌اتکا هستند.

1. Validity/Credibility

پایایی روش تعیین استاندارد

یکی دیگر از شاخص‌های ارزشیابی تعیین استاندارد، پایایی[1] نتایج است که گاهی به جای آن از کلمه ثبات[2] استفاده می‌شود. علی‌رغم تفاوت‌های مفهومی، این دو گاهی به صورت جایگزین هم به کار می‌روند. در اینجا نیز ما ثبات و پایایی را معادل یکدیگر در نظر می‌گیریم. پایایی به این معنا است که در صورت تکرار اندازه‌گیری، تا چه میزان همان نتایج به دست می‌آیند. این سؤال در هر نوع اندازه‌گیری حائز اهمیت است که آیا نتیجه‌ای که حاصل شده، همان اندازه واقعی است و چرا در صورت تکرار اندازه‌گیری، ممکن است نمره متفاوتی به دست آید.

از آنجا که پایایی در بستر آزمون و ارزیابی فراگیر آشناتر است، قبل از پرداختن به مفهوم پایایی در تعیین استاندارد، در مورد پایایی آزمون صحبت می‌کنیم. اگر نمره دانشجو امروز در یک امتحان صد نمره‌ای، ۸۰ شده است، شاید اگر آزمون را مجدداً تکرار کنیم، نمره او ۸۲ یا در نوبت بعدی ۷۵ شود. معلوم نیست کدام یک از این نمرات واقعی هستند. تفاوت بین نمرات ناشی از تاثیر خطای اندازه‌گیری است که خطایی تصادفی است.

باید گفت در هر اندازه‌گیری مواردی وجود دارند که باعث می‌شوند به صورت تصادفی در اندازه‌گیری خطا رخ دهد و با نام منابع خطا شناخته می‌شوند. در حوزه ارزیابی دانشجو، خطاها می‌توانند ناشی از اختلاف نظر مصححان حین تصحیح برگه‌های تشریحی باشد یا به دلیل زمان و شرایطی که آزمون در آن برگزار می‌گردد.

حال که مفهوم پایایی در حوزه آزمون تا حدی روشن شد، به مفهوم پایایی در تعیین استاندارد می‌پردازیم. صاحب‌نظران معتقد هستند مفهوم اندازه‌گیری در آزمون با مفهوم اندازه‌گیری در تعیین استاندارد متفاوت است. آنها ارزیابی فراگیر را یک اندازه‌گیری با رویکرد سوژه–محور[3] می‌دانند که در آن دانشجویان تمرکز اصلی اندازه‌گیری هستند. در حالی که تعیین استاندارد، یک اندازه‌گیری با رویکرد محرک–محور[4] است که در آن اندازه‌گیری در رابطه با سؤالات است. اعتقاد بر این است که برای سنجش پایایی هر دو، می‌توان از نظریه کلاسیک استفاده کرد، با این تفاوت که در تعیین استاندارد، به جای فرم‌های موازی که حاوی سؤالات متفاوت هستند، جلسات موازی داریم که شامل داوران متفاوتی است (نیکولز و همکاران[5] ۲۰۱۰). تفاوت‌ها و شباهت‌های موجود بین مفهوم پایایی و راه‌های اندازه‌گیری آن در آزمون و تعیین استاندارد در نظریه کلاسیک در جدول ۱–۲۹ آمده است.

جدول ۱–۲۹ : مقایسه روش‌های ارزیابی پایایی در آزمون و تعیین استاندارد

پایایی	آزمون	تعیین استاندارد
ثبات درونی	ارائه سؤالات در یک نوبت	جمع آوری نظرات داوران در یک نوبت
اندازه‌گیری مجدد	استفاده از همان سؤالات در نوبت‌های مختلف ارائه سؤالات در مدت زمانی یکسان	استفاده از همان داوران در نوبت‌های مختلف جمع آوری نظرات داوران در مدت زمانی یکسان
فرم‌های موازی	یکسان بودن توزیع نمرات و واریانس در دو فرم ارائه سؤالات در مدت زمانی یکسان	یکسان بودن توزیع نمرات و واریانس در دو گروه جمع‌آوری نظرات داوران در مدت زمانی یکسان
هر سه مورد	تعداد مناسب سؤال انتخاب سؤالات با توجه به شاخص‌های دشواری و تمیز یکسان بودن نوع سؤالات طراحی سؤالات بر اساس دستورالعمل ها	تعداد مناسب داور انتخاب داوران با توجه به سیاست‌ها و دیدگاه‌های بیرونی یکسان بودن روش انتخاب داوران بر اساس ویژگی‌های مدنظر

1. Reliability
2. Consistency
3. Subject-centered
4. Stimulus-centered
5. Nichols et al

در نظریه کلاسیک، در حالت سوژه–محور وقتی واریانس بین دانشجویان زیاد و واریانس بین سؤالات کم است، پایایی افزایش می‌یابد. در مقابل، در رویکرد محرک –محور، پایایی بالا در صورت افزایش واریانس بین سؤالات وکاهش واریانس بین افراد به دست می‌آید

تفاوت این دو در نظریه تعمیم‌پذیری به این صورت خواهد بود که در حوزه آزمون، دانشجویان، محور هستند و آزمون باید بتواند بین دانشجویان با سطوح مختلف توانمندی افتراق دهد و تعیین کند که نتیجه کسب شده توسط دانشجویان تا چه اندازه به سایر شرایط قابل تعمیم است. یعنی در صورت تغییر شرایط آزمون، مثلاً تغییر مصححان یا تغییر زمان برگزاری آزمون یا تغییر سؤالات، آیا باز هم همان نتیجه را کسب می‌کنند یا خیر. در حالی که در حوزه تعیین استاندارد که محرک–محور است، قصد این است که ببینیم نمره حدنصاب تعیین شده تا چه حد به سایر شرایط قابل تعمیم است. منظور از سایر شرایط در اینجا، تغییر داوران، پیش گرفتن روش‌های مختلف و اضافه کردن مواردی مانند بحث و بررسی نمرات واقعی است.

تجلی مفهوم ثبات و پایایی در تعیین استاندارد به این صورت خواهد بود که در صورت تکرار فرایند، نتایج تا چه اندازه یکسان خواهند بود. منظور از نتیجه، نمره حدنصابی است که تعیین شده و میزان قبولی که از آن به دست آمده است. از آنجا که کل فرایند تعیین استاندارد، قضاوتی است، برخی معتقد هستند که به کار بردن کلمه «پایایی» برای آن چندان صحیح نیست زیرا در تعیین استاندارد نمی‌توان از میزان نزدیک بودن نتایج به واقعیت صحبت کرد. اساساً نمره حقیقی به عنوان حدنصاب وجود ندارد. استاندارد با هر روشی که محاسبه شود، باز هم، قراردادی و ذهنی است و با هر روشی، نمره متفاوتی به عنوان حدنصاب یک آزمون استخراج می‌شود که این موضوع به معنای اشتباه بودن آن نیست.

در تفسیر ضرایب پایایی روش‌های تعیین استاندارد باید با احتیاط عمل کرد. سنجش پایایی یا ثبات روش تعیین استاندارد هرچند در مورد کیفیت روش اطلاعات خوبی در اختیار ذی‌نفعان می‌گذارد اما در عین حال درست بودن نمره حدنصاب را به دلیلی که ذکر شد، تضمین نمی‌کند. مسأله دیگر اینکه، بالا بودن ضریب پایایی لزوماً و همیشه نشانه خوبی نیست. به عنوان مثال، در روش انگوف، پایایی حاصل از تعیین استاندارد انفرادی را با پایایی تعیین استاندارد بعد از بحث مقایسه می‌کنیم و می‌بینیم که بحث باعث افزایش توافق بین داوران، کاهش واریانس بین نظرات و در نتیجه افزایش پایایی شده است. با بررسی جریان جلسه متوجه می‌شویم که داوران تحت تاثیر یک داور خاص که بحث را به دست گرفته است، قرار گرفته‌اند. یعنی در واقع، با برجسته شدن و تاثیر یک نفر به خصوص در جریان بحث، داوران احساس فشار می‌کنند و سعی می‌کنند که حتماً نظر خود را تغییر دهند تا مشابه بقیه شود. در موارد این چنینی، افزایش ضریب پایایی لزوماً نشانه فرایند مطلوب نیست.

تاکنون ما در مورد پایایی در آزمون و پایایی در تعیین استاندارد صحبت کردیم. نکته آخر در مورد پایایی که توسط برخی از صاحب‌نظران مورد توجه قرار گرفته است، جمع شدن خطای این دو است. دیدیم که خطا احتمال دارد در خود آزمون رخ دهد که تجلی آن در «نمرات» دانشجویان مشهود است. این مسأله‌ای است که به صورت معمول در آزمون‌ها مورد بررسی قرار می‌گیرد. وقتی آزمونی برگزار می‌شود و پایایی آن سنجیده می‌شود، در حقیقت، میزان احتمال خطا در نمرات دانشجویان برآورد می‌شود. از طرفی خطا در روند تعیین استاندارد هم ممکن است به عنوان یک فرایند مستقل رخ دهد که تجلی آن، در «وضعیت رد یا قبول» دانشجویان است. بنابراین، این احتمال وجود دارد که دو خطای فوق مجموعاً بر نمره قبولی اثر بگذارند. تاثیر این دو خطا می‌تواند موجب مخدوش شدن تقسیم‌بندی دانشجویان شود. برخی معتقد هستند این دو خطا به هم وابسته نیستند (برنان و لاکوود ۱۹۸۰) اما برخی معتقد هستند که این دو خطا با یکدیگر ارتباط دارند (کین و ویلسون[1] ۱۹۸۴) و این کواریانس حاصله است که مشخص می‌کند چقدر حدنصاب در بستر اهداف همان آزمون خوب تعیین شده است. کواریانس منفی، باعث خطای زیاد در دسته‌بندی دانشجویان می‌شود.

1. Kane & Wilson

برای روشن شدن موضوع، به مثال زیر توجه کنید:

نمره دانشجو در آزمونی صد نمره ای، ۴۵ شده است. نمره حدنصاب آزمون با استفاده از روش انگوف، ۵۱ برآورد شده است. یعنی می‌توانیم اعلام کنیم دانشجو موفق به کسب استاندارد نشده و در آزمون رد است.
حال با بررسی پایایی آزمون و پایایی روش تعیین استاندارد متوجه می‌شویم که خطای معیار آزمون ۲ برآورد شده است. به عبارت دیگر، نمره حقیقی دانشجو به احتمال ۶۸ درصد بین ۴۳ تا ۴۷ است. یعنی ممکن است نمره دانشجو ۴۷ باشد و به خاطر خطای آزمون کمتر گزارش شده باشد. در هر حال، با توجه به نمره حدنصاب ۵۱، این دانشجو حتی با احتساب خطا، باز هم در آزمون رد است.
حال در می‌یابیم که که خطای معیار تعیین استاندارد، ۵ برآورد شده است. این بدان معنا است که استاندارد می‌تواند بین ۴۶ تا ۵۶ باشد و دقیقاً ۱۵ نیست. به عنوان مثال اگر نمره واقعی دانشجو ۴۷ و استاندارد ۴۶ باشد، در این حالت آیا وضعیت دانشجو قبول است یا رد؟ از آنجا که بین دو بازه فوق، همپوشانی وجود دارد، نمی‌توان با اطمینان گفت که دانشجو قبول نشده است. می‌بینیم که اعلام وضعیت افتاده برای دانشجو چندان هم قطعی نیست و ممکن است ناشی از خطای تصادفی اندازه‌گیری باشد.

نمره حدنصاب و میزان قبولی

همان‌طور که در قسمت پایایی عنوان شد، روش‌های مختلف موجب تعیین استانداردهای مختلف می‌شوند؛ مخصوصاً اگر داورها هم متفاوت باشند. این مسأله را باید در نظر داشت که استاندارد با هر روشی که محاسبه شود، نهایتاً قراردادی و ذهنی است. هیچ فرمول عینی و ریاضی برای تعیین حدنصاب وجود ندارد. این موضوع که با هر روشی، نمره متفاوتی به عنوان حدنصاب یک آزمون استخراج می‌شود، به معنای اشتباه بودن نتایج نیست؛ بلکه ذات تعیین استاندارد همین است.

بدیهی است که اگر نمره حدنصاب حاصل از دو روش متفاوت باشد، درصد قبولی منتج شده نیز متفاوت خواهد بود. از طرفی، گاهی تفاوت بین دو نمره حدنصاب چندان محسوس نیست اما توزیع نمرات دانشجویان شرکت‌کننده در امتحان به گونه‌ای است که درصدهای قبولی به صورت چشم‌گیر متفاوت خواهند شد. به عنوان مثال، اگر میانه نمرات خیلی نزدیک به استانداردِ تعیین شده باشد، به این معنا است که با تغییر کم حدنصاب، وضعیت رد و قبول تعداد زیادی از دانشجویان تغییر می‌کند.

در مورد اینکه کدام روش نهایتاً سختگیرانه‌تر است و استانداردهای بالاتر و میزان قبولی پایین‌تر دارد، نمی‌توان نظر قطعی داد. مطالعات مختلفی که به مقایسه روش‌ها برای تعیین استاندارد یک آزمون پرداخته‌اند، به نتایج متفاوتی رسیده‌اند: برخی پژوهش‌ها نشان داده‌اند که روش‌های مبتنی بر آزمون تعداد قبولی بالاتری نسبت به روش‌های مبتنی بر آزمودنی داشته‌اند (نورسینی ۱۹۹۳)؛ در حالی‌که نتایج مطالعات دیگر حاکی از آن است که روش‌های مبتنی بر آزمون سختگیرانه بوده‌اند و انواع مبتنی بر آزمودنی، تعداد قبولی بیشتری داشته‌اند (روتمن و همکاران ۱۹۹۶، کین و همکاران ۱۹۹۹، کریمر و همکاران ۲۰۰۳، شونیم–کلین و همکاران ۲۰۰۹). بر اساس برخی مطالعات نیز نمرات حدنصاب دو گروه مشابه بودند (کافمن و همکاران ۲۰۰۰).

قابلیت اجرا

یکی دیگر از شاخص‌های مهم که در ارزشیابی روش‌های تعیین استاندارد باید مورد توجه قرار گیرد، قابلیت اجرای روش است. به عبارت دیگر، از لحاظ عملی چقدر با امکانات و شرایط موجود، انجام یک روش خاص برای تعیین حدنصاب قبولی آزمون امکان‌پذیر است. البته هر روش، دشواری‌ها و چالش‌های اجرایی خاص خود را دارد و هیچ شیوه آسانی که در عین حال معتبر هم باشد، نمی‌توان یافت. بنابراین، مسأله‌ای که مورد تاکید است، ضرورت «توجه» به قابلیت اجرا است. هر دانشکده‌ای باید امکانات و شرایط خود را ارزیابی کند و به صورت آگاهانه در جهت رفع مشکلات و جلب مشارکت مسؤولان و تامین هزینه‌ها گام بردارد.

هرچند که هر روش اقتضائات خاص خودش را دارد اما به صوت کلی مواردی را می‌توان نام برد که کم و بیش همه جا روی قابلیت اجرا اثر دارند. در اینجا به برخی از این مشکلات و راه‌حل‌های آنها می‌پردازیم:

اول اینکه روش‌های مبتنی بر آزمون که نیاز به تشکیل جلسه دارند، ممکن است از لحاظ اجرا با مشکل مواجه شوند.

زیرا اعضای هیأت علمی احساس می‌کنند وقت خود را می‌توانند به کارهای مهم‌تر اختصاص دهند. بنابراین، محتمل است که عده‌ای از آنها در جلسه شرکت نکنند. غیبت داوران از چند جهت مهم است و باید از آن جلوگیری شود.

- اولاً تعداد داوران برای رسیدن به روایی و پایایی قابل‌قبول باید مناسب باشد که قبلاً در مورد آن صحبت شد. در نظر بگیرید حداقل تعداد توصیه‌شده را در نظر گرفته‌ایم و مثلاً ۶ نفر را برای شرکت در جلسه دعوت کرده‌ایم. اگر دو نفر غیبت کنند، عملاً روایی و پایایی روش به شدت زیر سؤال خواهد رفت.
- تعداد داوران مخصوصاً اگر قرار باشد بحث گروهی صورت گیرد، بسیار اهمیت دارد. معمولاً در تعداد کم تبادل نظرات به خوبی شکل نمی‌گیرد و در واقع بحث موثر نخواهد بود.
- اگر قرار باشد جلسه دومی پس از برگزاری آزمون تشکیل شود تا بررسی نمرات واقعی صورت گیرد، عملاً داوری که در جلسه اول حضور داشته است اما در جلسه دوم غیبت کرده است، از محاسبات کنار گذاشته می‌شود. پس اهمیت دارد که در تمام جلسات تمام داوران حاضر باشند.

برای مقابله با چالش فوق مهم است که تدابیری اندیشیده شود از جمله:

- از قبل با اعضای هیأت علمی هماهنگی صورت گیرد، اهمیت جلسه به صورت رسمی و غیررسمی یادآوری شود و زمان و مکان جلسه به اطلاع ایشان برسد.
- زحمتی که داوران متقبل می‌شوند و زمانی که صرف جلسه می‌کنند، به شکل مناسب جبران شود.
- حتی‌المقدور از حواشی جلسات کاسته شود تا زمان جلسه طولانی نشود. مثلاً اگر قرار است جلسه توجیهی و آشنایی با روش تعیین استاندارد در همان روز صورت گیرد، از سخنرانی‌های کسل‌کننده و طولانی پرهیز شود و بیشتر کار عملی و بحث انجام شود. اگر قرار است محاسباتی در جلسه صورت گیرد و نتایج همانجا به رویت داوران برسد، از قبل ترتیبی اتخاذ گردد که داده‌ها سریعاً جمع‌آوری گردند و وارد فایل شوند.
- اگر قرار است بررسی نمرات واقعی انجام شود، همیشه تشکیل دو جلسه الزامی نیست. می‌توان جلسه قبل از آزمون را حذف کرد. بنابراین داوران فقط یک بار به جلسه دعوت می‌شوند و ابتدا بدون مشاهده نتایج آزمون استاندارد را تعیین می‌کنند. سپس در دور بعدی، نمرات دانشجویان در اختیار آنها گذاشته می‌شود تا در صورت تمایل تجدید نظر کنند.

دومین مسأله اینکه احتمال دارد روش‌هایی که مستلزم فعالیت ذهنی هستند، از لحاظ اجرایی با مشکل مواجه شوند. البته تقریباً در تمام روش‌های تعیین استاندارد از آنجا که داور باید به صورت مسمتر قضاوت کند، احتمالات مختلف را بالا و پایین کند، مفهوم دانشجوی مرزی را تجسم کند و ...، فعالیت ذهنی مدام مورد نیاز است اما برخی از روش‌ها روند پیچیده‌تری دارند. شیوه‌های مبتنی بر آزمون که چند مرحله‌ای هستند یا نیاز دارند که داور در آن واحد چند مقوله را در ذهن نگه دارد و بسنجد، از جمله مواردی هستند که روند پیچیدهای دارند. خستگی ذهنی داوران بر روند قضاوت و تصمیم‌گیری آنها اثر می‌گذارد. بنابراین، در نظر گرفتن فرصت استراحت در فواصل کار و حتی مسائل ساده‌ای مانند پذیرایی حائز اهمیت هستند.

سومین چالش اجرایی، نیاز به داوران متخصص و با تجربه است که البته باز هم برای تمام روش‌های تعیین استاندارد صدق می‌کند اما برای برخی بیشتر محسوس است. دوباره دو محدودیتی را که در بالا به آنها اشاره شد، در نظر بگیرید. شاید این طور به نظر برسد که روش‌های مبتنی بر آزمودنی از لحاظ اجرایی ساده‌تر هستند. اما حقیقت این است که در برخی از روش‌های مبتنی بر آزمودنی، مشکل دیگری وجود دارد و آن اینکه دسترسی به داوران متخصص و با تجربه دشوارتر است. به عنوان مثال روش گروه مرزی را برای تعیین استاندارد یک OSCE در نظر بگیرید. هرچند تصور می‌شود در این حالت، فکر کردن به مفاهیم فرضی و تخیلی آنچنان که در روش انگوف لازم است وجود ندارد و کار داوران راحت‌تر اس، اما از طرفی معمول است که در بسیاری از آزمون‌های OSCE به جای اعضای هیأت علمی خبره، از بیماران استانداردشده، دانشجویان سال بالاتر یا دستیاران برای نمره‌دهی در ایستگاه‌ها استفاده شود. هرچند شاید این افراد به خوبی از عهده تکمیل چک‌لیست برآیند، برای ارزیابی کلی در مقیاس لیکرت که برای روش گروه مرزی احتیاج است، نیاز به قضاوت معتبرتری است.

چهارمین مشکلی که در مورد روش‌های مبتنی بر آزمودنی و کلاً هر روش دیگری که در آن باید نتایج عملکرد دانشجویان بررسی شود وجود دارد، تصحیح برگه‌های دانشجویان است که هر چند کار پیچیده‌ای نیست، ولی تصحیح برگه‌ها، جمع‌آوری نمرات و انجام تحلیل‌های اولیه در مورد آنها از لحاظ اجرایی کار می‌برد و نیازمند صرف وقت است.

پنجم این که روش‌هایی که نیاز به محاسبات پیچیده دارند، شاید در همه جا قابل استفاده نباشند. البته عموم روش‌های تعیین استاندارد، نیاز به محاسبات پیچیده ندارند و صرفاً محاسباتی در حد میانگین و میانه و توزیع نمرات کفایت می‌کند. این مسأله شاید برای روشی مانند رگرسیون مرزی پیش آید که برای احتساب معادله رگرسیونی، نیروی متخصص مورد نیاز باشد. هرچند که این محاسبه نیز واقعاً چندان پیچیده نیست.

به صورت خلاصه، هیچ روش تعیین استاندارد معتبری بدون صرف وقت و هزینه و انرژی، انجام‌پذیر نیست. همان‌طور که قبلاً ذکر شد، منظور از موارد مطرح‌شده این نیست که این روش‌ها اساساً مورد استفاده قرار نگیرند بلکه دست‌اندرکاران باید تمهیداتی بیاندیشند تا ضمانت اجرایی آن را زیاد کنند. امکانات و شرایط موجود، منابع مالی، دسترسی به اعضای هیأت علمی و به صورت کلی، قابلیت اجرا در انتخاب روش تعیین استاندارد آزمون بسیار تاثیرگذار است.

مدل نمره‌دهی

تاکنون در مورد انواع روش‌های تعیین استاندارد صحبت شد. موردی که باقی می‌ماند، مدل نمره‌دهی است که مستقل از روش تعیین استاندرد است. به عبارت دیگر، فارغ از اینکه حدنصاب قبولی با چه روشی تعیین شده است، در خصوص مدل نمره‌دهی باید به صورت جداگانه تصمیم‌گیری شود. چون بر اساس تجربه دیده شده است که معمولاً در خلال بحث تعیین استاندارد، بحث‌هایی در مورد مدل نمره‌دهی صورت می‌گیرد، در اینجا به مدل نمره‌دهی نیز می‌پردازیم.

در مورد هر آزمونی مخصوصاً آنهایی که از چند قسمت تشکیل شده‌اند، باید تصمیم گرفت که نمره نهایی چگونه حاصل می‌شود. مدل نمره‌دهی دلالت دارد به اینکه نمرات قسمت‌های مختلف آزمون را با چه روشی ترکیب کنیم تا وضعیت رد و قبول دانشجو را در کل معلوم کنیم. اگر نمرات حاصل از تمام قسمت‌ها را با هم جمع کنیم و در مجموع بسنجیم که آیا دانشجو قبول شده یا نه، به صورت کاملاً جبرانی[1] عمل کرده‌ایم. دو حالت دیگر، غیرجبرانی[2] (یا کونژنکتیو[3]) و نیمه جبرانی[4] هستند.

مثلاً حالتی را در نظر بگیرید که امتحان پایان ترم با نمره قبولی ۱۰ از ۲۰، شامل آزمون کتبی (۱۲ نمره)، آزمون شفاهی (۴ نمره) و ارائه پروژه (۴ نمره) باشد. در مدل نمره‌دهی جبرانی، کمبود نمره در یک بخش، توسط نمره بالای بخش دیگر قابل جبران است. یعنی، اگر دانشجویی پروژه ارائه ندهد ولی مجموع نمره کتبی و شفاهی وی ۱۴ شود، قبول محسوب می‌شود اما در مدل غیرجبرانی حتی با وجود اینکه دانشجو حدنصاب کلی را کسب کرده است، از آنجا که موفق به کسب حدنصاب در تمام قسمت‌ها نشده است، مردود در نظر گرفته می‌شود (جدول ۲-۲۹).

جدول ۲-۲۹ : مقایسه مدل نمره دهی جبرانی و غیرجبرانی در وضعیت یک دانشجو

	حداکثر نمره	نمره دانشجو	حدنصاب با مدل جبرانی	حدنصاب با مدل غیرجبرانی
آزمون کتبی			-	قبول
آزمون شفاهی	۴	۳	-	قبول
ارائه پروژه	۴	۰	-	رد
نتیجه کل	۲۰	۱۴	قبول	رد

1. Total compensatory (TC)
2. Non compensatory (NC)
3. Conjunctive
4. Partial compensatory (PC)

این مسأله درون یک آزمون نیز صدق می‌کند که با ذکر چند مثال توضیح داده می‌شود:

یک آزمون OSCE با ۱۰ ایستگاه را برای دستیاران رشته طب اورژانس در نظر بگیرید. گاهی ما فقط به نمره کل دستیار نگاه می‌کنیم و عملکرد او در هر یک از ایستگاه‌ها چندان اهمیتی ندارد. در این صورت اگر جمع نمره او بیشتر از استاندارد به دست آمده باشد، قبول محسوب می‌شود. به این حالت، مدل نمره‌دهی جبرانی گفته می‌شود. یعنی شاید در یک ایستگاه نمره خوبی نگرفته باشد اما نمره ایستگاه دیگر آن را جبران کرده است. گاهی برای ما مهم است که دستیاران در تک‌تک ایستگاه‌ها به حدنصاب برسد، در این صورت مدل غیرجبرانی نامیده می‌شود. یعنی اگرچه در کل نمره دستیار به حدنصاب رسیده است اما از آنجا که در یک ایستگاه کمتر از حد مورد انتظار ظاهر شده است، در نهایت در آزمون قبول نمی‌شود. نوع سومی هم وجود دارد که در واقع مخلوط این دو حالت است به این معنا که علاوه بر اینکه نمره کل باید به حدنصاب برسد، حدنصاب برای تعدادی از سؤالات/ ایستگاه‌ها و نه تمام آنها گذاشته می‌شود. مثلاً کمیته آزمون تصمیم می‌گیرد که از ده ایستگاه آزمون فوق، ایستگاه انتوباسیون برای رزیدنت طب اورژانس بسیار مهم است و حتما باید در آن نمره قبول بیاورد. به این مدل نیمه‌جبرانی گفته می‌شود.

در مثال دیگر، یک آزمون OSCE برای پایان دوره کارآموزی دانشجویان پزشکی در نظر بگیرید که شامل ایستگاه‌هایی از جمله اخذ شرح حال، مهارت ارتباطی، معاینه فیزیک و .. است. از نظر کمیته آزمون، حوزه مهارت‌های ارتباطی بسیار مهم است و دانشجو حتماً باید در آن نمره قبولی کسب کرده باشد و بدون آن در کل امتحان قبول محسوب نمی‌شود. در اینجا باید به رابطه‌ای که ممکن است بین چند ایستگاه با یکدیگر وجود داشته باشد، توجه کرد. مثلاً اگر در یک ایستگاه مهارت شرح‌حال مورد ارزیابی قرار می‌گیرد و در ایستگاه دیگر مهارت برقراری ارتباط، باید توجه داشت که چگونگی برقراری ارتباط دانشجو با بیمار، روی مهارت شرح‌حال‌گیری او هم اثر خواهد گذاشت. یعنی نمرات ایستگاه مهارت ارتباطی و ایستگاه شرح حال با هم ارتباط دارند. استفاده از روش غیرجبرانی در صورتی که بین محتوای مورد ارزیابی در ایستگاه‌های مختلف رابطه وجود نداشته باشد و یا این ارتباط قابل چشم‌پوشی باشد، نتایج بهتری می‌دهد (سیرل[1] ۲۰۰۰).

شونیم-کلین و همکاران ۲۰۰۹

در این مطالعه، سه روش گروه مرزی، آنگوف بلی/خیر و آنگوف تغییریافته برای تعیین استاندارد یک OSCE با ۱۴ ایستگاه برای ارزیابی عملکرد ۱۱۹ دانشجوی سال سوم دندانپزشکی با هم مقایسه شدند تا روش بهینه‌ای که قادر باشد دانشجویان توان‌مند و غیرتوان‌مند را از هم جدا کند، معرفی گردد. حوزه‌های توان‌مندی مورد نظر شامل: تعهد حرفه‌ای، مهارت‌های ارتباطی، گردآوری اطلاعات بالینی، تشخیص، ارائه طرح درمان و بهداشت دهان-دندان بود که در چهار کلاستر که هر یک شامل سه یا چهار ایستگاه بودند، مورد ارزیابی قرار گرفتند. برای هر ایستگاه چک‌لیستی شامل ۱۰ آیتم طراحی شده بود؛ به جز یک ایستگاه که پنج آیتم داشت. همچنین آیتم‌ها بر اساس اهمیتشان وزن‌دهی شده بودند. علاوه بر آن هر آزمون‌گر یک چک‌لیست دیگر داشت تا ارزیابی کلی خود را از عملکرد دانشجو در آن ایستگاه به صورت لیکرتی از یک تا پنج مشخص کند. امتحان در چهار روز و در شش جلسه برگزار شد. فرمت کلی ایستگاه‌ها در روزها یکسان بود ولی تغییرات جزئی مثلاً در سناریو یا کلیشه‌های رادیوگرافی اِعمال شده بود. در روزهای متفاوت در اکثر ایستگاه‌های مشابه، آزمون‌گر عوض شده بود که پایایی این روش در مقاله قبلی همین نویسندگان ثابت شده بود. ۴۲ استاد دندانپزشکی به عنوان آزمون‌گر در این چهار روز شرکت کردند. تعیین استاندارد به شیوه آنگوف قبل از برگزاری OSCE و با دعوت از ۱۳ داور انجام شد. در ابتدای جلسه در مورد تعریف و ویژگی‌های دانشجوی مرزی بحث شد و سپس هر کس با پاسخ بلی/خیر به قضاوت در مورد تک‌تک آیتم‌ها پرداخت. گفته شده است که در روش آنگوف، داوران استاندارد را بالا تعیین می‌کنند. بنابراین بعد از OSCE همان داورها دوباره جمع شدند تا بررسی تمرات واقعی انجام گیرد: در این روش اطلاعات مربوط به میانگین نمرهٔ دانشجویان در هر ایستگاه، میزان قبولی و استانداردهایی که قبلاً تعیین شده بود، در اختیار داوران قرار داده شد. پروسه قضاوت مجدداً تکرار شد و داورها مختار بودند که در تصمیم خود برای هر آیتم تجدیدنظر کنند. به این ترتیب آنگوف تغییریافته هم اجرا شد. برای هر ایستگاه و هر داور نسبت قبولی دو روش محاسبه شد و سپس میانگین گرفته شد. به نظر نویسندگان نتایج حاصل از روش مرزی قابل قبول‌تر بود. از نظر میزان قبولی در روش مرزی بیشتر از روش آنگوف و حتی آنگوف تغییر یافته بود. بنابراین می‌شود گفت که در این مطالعه روش مرزی سخت‌گیری کمتری نشان داده است. در مدل نسبتاً جبرانی (PC) که از نظر نویسندگان در مقایسه با مدل‌های جبرانی کامل و غیر جبرانی، بهترین و قابل قبول‌ترین استاندارد را به دست داده بود، میزان قبولی در روش آنگوف I و II و مرزی به ترتیب ۳۰/۳ درصد، ۳۴/۵ درصد و ۶۱/۳ درصد بود. همچنین شاخص پایایی (RMSE) نشان‌دهنده پایایی بیشتر روش مرزی بود. برای مدل جبرانی نسبی (PC) این میزان به ترتیب ۳/۷-۲/۰ درصد، ۲/۲-۱/۸ درصد و ۰/۷-۰/۶ درصد به دست آمد

1. Searle

منابع

1. Barman A. Standard setting in student assessment: is a defensible method yet to come? Ann Acad Med Singapore 2008; 37(11): 957-63.
2. Boulet JR, de Champlain A and MCkinley D. Setting defensible performance standards on OSCEs and standardized patient examinations. Medical Teacher 2003;25(3):245–249
3. Brennan RL, Lockwood RE. A comparison of the Nedelsky and Angoff cutting score procedures using generalisability theory. Appl Psychol Measurement 1980;4:219–40
4. Kane M, Wilson J. Errors of measurement and standard- setting in mastery testing. Applied Psychological Measurement 1984;8:107-1 5.
5. Kane MT, Crooks TJ, Cohen AS. Designing and evaluating standard-setting procedures for licensure and certification tests. Adv Health Sci Educ Theory Pract 1999:4:195–207.
6. Kaufman DM, Mann KV, Muijtjens AMM, van der Vleuten CPM. A comparison of standard setting procedures for an OSCE in undergraduate medical education. Acad Med 2000;75:267–71.
7. Kramer A, Muijtjens A, Jansen K, Dusman H, Tan L, van der Vleuten C. Comparison of a rational and an empirical standard setting procedure for an OSCE. Objective structured clinical examinations. Medical Education 2003;2:132–139.
8. Nichols P, Twing j, Mueller CD, O'Malley K. Standard-Setting Methods as Measurement Processes. Educational Measurement: Issues and Practice 2010; 29(1):14-24
9. Norcini JJ, Stillman PL, Sutnick AI, et al. Scoring and Standard Setting with Standardized Patients. Evaluation & the Health Professions 1993;16(3):322-332.
10. Norcini JJ. Setting standards on educational tests. Medical education 2003;37(5):464-9.
11. Rothman AI, Blackmore D, Dauphinee WD, Reznick R. The use of global ratings in OSCE station scores. Advances in Health Sciences Education 1996;1(3):215-219.
12. Schoonheim-Klein M, Muijtjens A, Habets L and et al. Who will pass the dental OSCE? Comparison of the Angoff and the borderline regression standard setting method. Eur J Dent Educ 2009;13:162–171
13. Searle J. Defining competency – the role of standard setting. Med Educ 2000;34:363–366.
14. Wood TJ, Humphrey-Murto SM, Norman GR. Standard setting in a small scale OSCE: a comparison of the Modified Borderline-Group Method and the Borderline Regression Method. Advances in health sciences education: theory and practice2)11;2006):115-22.
15. Zieky M. Perie M, Livingston S. A Primer on Setting Cut Scores on Tests of Educational Achievement, Princeton, N.J: Educational Testing Service, 2006.

۸

تحلیل آزمون

فصل ۳۰

کلیات تحلیل آزمون

مقدمه

پس از اینکه آزمون برگزار شد و نمرات و نتایج آن اعلام شدند، هنوز کار مسؤولان امر ارزیابی فراگیران تمام نشده است. برای اینکه مطمئن باشیم آزمونی که برگزار کرده‌ایم، از کیفیت لازم برخوردار بوده است و به اهداف خود رسیده است، لازم است که روند برگزاری آزمون را ارزشیابی و نتایج حاصله را تحلیل کنیم. ضرورت این مسأله مخصوصاً در آزمون‌های سطح بالا و مهم علوم پزشکی که هم برای دانشجویان سرنوشت‌ساز هستند و هم باید پاسخگوی انتظارات جامعه باشند، محسوس‌تر است.

در حالی که همگان بر ضرورت و اهمیت این امر واقف هستند، معمولاً این مرحله یا اصلاً انجام نمی‌شود و یا به صورت بسیار مختصر، نامنظم و غیرمستمر انجام می‌شود. در حالی که به نظر می‌رسد تنها با پایش مستمر فرایند و پیامدهای ارزیابی، اخذ بازخورد از جنبه‌های مختلف و اِعمال نکات اصلاحی در دفعات بعدی و بررسی مجدد تاثیر آنها می‌توان امیدوار بود که در نحوه ارزیابی بهبود دیده شود. در غیر این صورت، تکرار اشتباهات بدون آن که اساساً متوجه وقوع آنها باشیم و معمول شدن یک رویه ثابت بدون اصلاح نارسایی‌ها و بازنگری فرایندها، بسیار محتمل است.

نکته اینجاست که حتی در موارد معدودی که تحلیل آزمون صورت می‌گیرد، معمولاً نتایج آن به دست صاحبان اصلی یعنی طراحان آزمون نمی‌رسد. گاهی نیز اعداد خامی به آنها ارائه می‌شود که چون تفسیر و کاربرد آن برای ایشان روشن نیست، در نهایت سود چندانی حاصل نمی‌شود. در حالی که تحلیل آزمون تنها در همکاری نزدیک دو گروه طراح و تحلیل‌گر معنا پیدا می‌کند. هر تلاشی که در این راستا نباشد، مانند حلقه‌های ناپیوسته زنجیری است که نهایتاً نمی‌توان انتظار خاصی از آن داشت. در واقع تفسیر یافته‌های آماری که از تحلیل آزمون به دست آمده است، به تنهایی از عهده تحلیل‌گر برنمی‌آید و طراح آزمون و مدرس باید در این خصوص همفکری نمایند تا بتوانند با توجه به نتایج به دست آمده و شرایط خاص درس و آزمون پیشنهادهایی برای بهبود ارائه دهند. جالب توجه است که بهترین نتیجه زمانی حاصل می‌شود که این رابطه دوسویه باشد. همان‌طور که انتظار داریم نتایج تحلیل به دست طراح برسد و بر اساس آن تغییراتی اعمال شود، منطقی است که انتظارات و سفارش‌های طراح قبل از انجام تحلیل برای تحلیل‌گر شفاف شود.

شاید یکی از عللی که باعث نگرانی کاربران عادی و اجتناب آنها از ورود به بحث تحلیل می‌شود، این باشد که عمده مباحث مربوط به تحلیل آزمون، تحت عنوان سایکومتریک و با استفاده از ترمینولوژی تخصصی و فرمول‌های آماری و ریاضی در کتب تخصصی مورد بحث قرار می‌گیرند. برای کاهش این مشکل تلاش شده است که در این بخش کتاب، اغلب مفاهیم برای درک بهتر به صورت ساده و با ذکر مثال‌های روزمره بیان شوند. توصیه می‌شود به منظور مطالعه بیشتر و عمیق‌تر در این خصوص به متون اصلی و تخصصی مرتبط مراجعه گردد.

نکته دیگر اینکه تحلیل آزمون در اکثر کتب تنها به آزمون‌های کتبی آن هم از نوع چندگزینه‌ای محدود شده است. در حالی که یکی از نکاتی که باید به آن توجه داشت، این است که تحلیل و ارزشیابی بعد از استفاده از هر نوع ابزاری که به منظور ارزیابی دانشجویان به کار می‌رود، لازم و ضروری است. در این بخش تلاش می‌شود با توجه به تنوع ابزارهایی که در بخش‌های پیشین کتاب معرفی شدند، پوشش مناسبی از نظر تحلیل آنها نیز ارائه شود. برخی از شاخص‌های معرفی شده برای اغلب ابزارها قابل محاسبه هستند و تعدادی از شاخص‌ها برای یک ابزار خاص قابل استفاده هستند. همچنین برخی از شاخص‌ها برای بررسی کل آزمون به کار می‌روند و برخی دیگر اطلاعات خوبی در خصوص تک‌تک سؤالات و ایتم‌ها به دست می‌دهند.

نکته مهم دیگر این که همان‌گونه که برای ازریابی دانشجو نمی‌توان تنها به یک ابزار ارزیابی اکتفا کرد و لازم است که نتایج ابزارهای متعدد مدنظر قرار گیرند، در اینجا هم نمی‌توان تنها با یک شاخص در خصوص کیفیت آزمون یا سؤالات آن اظهار نظر قطعی کرد. معمولاً مجموعه‌ای از شاخص‌ها در کنار هم باید لحاظ شوند تا بتوان به دید مناسبی نسبت به فرایند و پیامدهای ارزیابی دست پیدا کرد. از آن مهم‌تر توجه به این نکته است که اگر مشاهده شد شاخص مربوط به یک سؤال پایین‌تر از حد مطلوب است، تنها راه‌حل، حذف آن سؤال نیست. این امر بسیار معمول است که به طراحان پیشنهاد شود سؤال با ضریب دشواری یا تمیز پایین را حذف کنند. در حالی که شاید در آن آزمون، پایین بودن ضریب دشواری یا تمیز واقعاً اتفاق بد و ناشی از اشتباه نبوده است. همچنین حذف سؤالات این چنینی ممکن است موجب کاهش پایایی گردد که به نوبه خود خطای آزمون را بالا می‌برد. بنابراین قبل از پیدا کردن راه‌حل، در نظر گرفتن شاخص‌های متعدد، تفسیر و تحلیل یافته‌ها و بررسی علل وقوع آنها، برای رفع اشکالات و نیل به اهداف مورد نظر بسیار مفید و کمک‌کننده است.

مسأله دیگر اینکه آنچه به صورت رایج در بسیاری از کتب و مقالات در خصوص تحلیل آزمون گفته می‌شود، بر مبنای نظریه کلاسیک آزمون است. در برخی از موارد هم که به سایر نظریه‌های اندازه‌گیری پرداخته می‌شود، مسائل به صورت غامض و پیچیده طرح می‌شوند که خارج از توان یک کاربر عادی به نظر می‌رسند. در بخش اول این کتاب، مفاهیم و مبانی سه نظریه اندازه‌گیری بیان شد. در این بخش قصد داریم که با ذکر مثال‌های مرتبط با ارزیابی فراگیران، کاربرد هر سه نظریه کلاسیک آزمون، نظریه تعمیم‌پذیری و نظریه سؤال پاسخ را در تحلیل آزمون بررسی کنیم.

فصل ۳۱

توزیع نمرات

در بسیاری از مواقع برای تفسیر نتایج حاصل از آزمون‌ها نمی‌توان صرفاً به نمرات خام دانشجویان اتکا کرد و باید اقداماتی روی آنها انجام داد. این اقدامات باعث می‌شوند که نمرات خام دانشجویان به صورت معنادار درآیند و قابل تفسیر و تحلیل شوند. برخی از این اقدامات در سطح کل آزمون و برخی روی تک‌تک سؤالات یا ایستگاه‌ها (به صورت کلی آیتم‌ها) صورت می‌گیرند.

در این فصل، به مرور شاخص‌هایی خواهیم پراخت که به کل آزمون برمی‌گردند. از جمله این موارد، محاسبه شاخص‌های مرکزی و پراکندگی، نمایش توزیع فراوانی نمرات به شکل جدول و نمودار، محاسبه نمرات استاندارد و همچنین تعیین همبستگی بین نمرات مختلف هستند. روایی و پایایی در دو فصل جداگانه مورد بحث قرار خواهند گرفت.

شاخص‌های مرکزی و پراکندگی

گفته می‌شود که بهترین شاخص مرکزی نمرات، میانگین[1] است. به عبارت دیگر، اگر تنها به یک شاخص دسترسی داشته باشیم که بیشترین اطلاعات را به دست بدهد، آن شاخص میانگین است.

$$\text{میانگین} = \frac{\text{مجموع نمرات}}{\text{تعداد دانشجویان}}$$

اما مشکل اینجا است که میانگین شدیداً تحت تاثیر نمرات خیلی بالا یا خیلی پایین که احتمالاً تعداد کمی هم دارند، قرار می‌گیرد. شاخص دیگر، میانه[2] است که دقیقاً نقطه وسط نمرات را مشخص می‌کند. به این ترتیب که نمرات به ترتیب ردیف می‌شوند و نمره‌ای که در وسط قرار گرفته است، مشخص می‌شود. شاخص مرکزی دیگر، نما[3] است که بیشترین نمره‌ای است که تکرار شده است.

اگرچه شاخص‌های مرکزی اطلاعات خوبی در اختیار می‌گذارند، برای تفسیر نتایج آزمون کافی نیستند و نشان نمی‌دهند پراکندگی نمراتی که دانشجویان کسب کرده‌اند، چطور بوده است. برای توصیف بهتر نتایج آزمون، شاخص‌های پراکندگی استفاده می‌شوند. دامنه[4]، فاصل بین کمترین و بیشترین نمره کسب شده است. واریانس[5] شاخص دیگری است که برای محاسبه آن، فاصله هر نمره از میانگین (نمره انحراف) به توان دو می‌رسد و سپس مجموع نمرات انحراف بر تعداد دانشجوبان تقسیم می‌شود. با جذر گرفتن از واریانس، انحراف معیار[6] به دست می‌آید:

1. Mean
2. Median
3. Mode
4. Range
5. Variance
6. Standard Deviation (SD)

$$\text{واریانس} = \frac{\text{مجموع مجذورات نمرات انحراف}}{\text{تعداد دانشجویان}}$$

$$\text{انحراف معیار} = \sqrt{\text{واریانس}}$$

محاسبه موارد فوق در مواردی که تعداد دانشجویان زیاد است، به راحتی با نرم افزارهایی مانند Excel و SPSS قابل انجام است. برای نشان دادن نحوه محاسبه آنها یک مثال ساده را مرور می‌کنیم. پنج دانشجو در امتحان شرکت کرده‌اند و نمرات آنها به صورت جدول ۱-۳۱ است. میانگین و انحراف معیار نمرات چند است؟

جدول ۱-۳۱ : نمرات ۵ دانشجو در یک آزمون

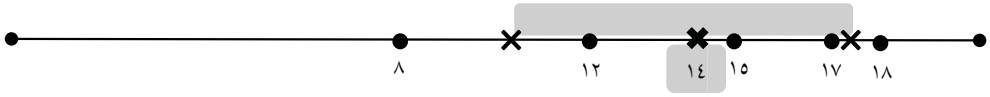
شماره دانشجو	۱	۲	۳	۴	۵
نمره	۸	۱۲	۱۵	۱۷	۱۸

$$\text{میانگین} = \frac{۸+۱۲+۱۵+۱۷+۱۸}{۵} = \frac{۷۰}{۵} = ۱۴$$

$$\text{میانگین} = \frac{(۸-۱۴)^۲+(۱۲-۱۴)^۲+(۱۵-۱۴)^۲+(۱۷-۱۴)^۲+(۱۸-۱۴)^۲}{۵} = \frac{۳۶+۴+۱+۹+۱۶}{۵} = \frac{۶۶}{۵} = ۱۳.۲$$

$$\text{انحراف معیار} = \sqrt{۱۳.۲} = ۳.۶۴$$

موارد فوق یعنی نمرات، میانگین و انحراف معیار نمرات یک آزمون را می‌توان به صورت زیر نشان داد:



جداول و نمودارها

توزیع نمرات شرکت‌کنندگان در آزمون به شکل جدول و نمودار قابل نمایش است.

جدول توزیع فراوانی نمرات

برای تهیه جدول توزیع فراوانی، نمرات دانشجویان به ترتیب ردیف می‌شوند و فراوانی هر نمره، یعنی دفعات تکرار آن در مقابل آن نوشته می‌شود. سپس می‌توان درصد هر یک از نمرات را نیز محاسبه کرد. به جدول ۲ـ۳۱ توجه کنید.

جدول ۲-۳۱ : فراوانی نمرات یک آزمون

نمره	۸	۹/۵	۱۱	۱۱/۵	۱۲	۱۳	۱۳/۵	۱۴	۱۵	۱۵/۵	۱۶	۱۷	۱۷/۵
فراوانی	۱	۱	۱	۱	۳	۲	۲	۱	۱	۱	۲	۳	۱

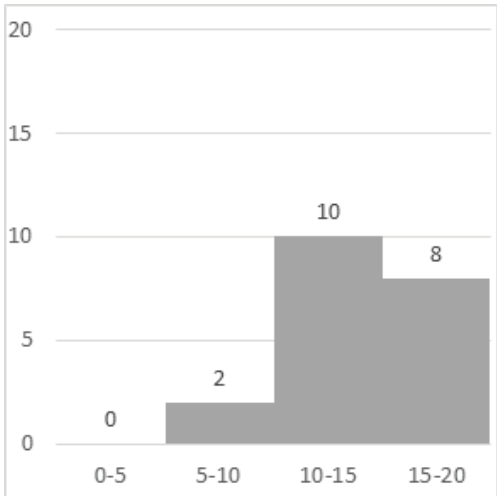
این حالت، فهرستی طولانی به دست می‌دهد که چندان کاربردی نیست. حالت کاربردی‌تر این جدول، مخصوصاً زمانی که تعداد دانشجویان زیاد است، این است که نمرات دسته‌بندی شوند و توزیع فراوانی نمرات هر دسته گزارش شود. مثلاً می‌توان نمرات ۰ تا ۲۰ را به صورت جدول ۳-۳۱ به چهار دسته تقسیم کرد.

جدول ۳-۳۱ : فراوانی و درصد نمرات آزمون با شرکت ۲۰ دانشجو

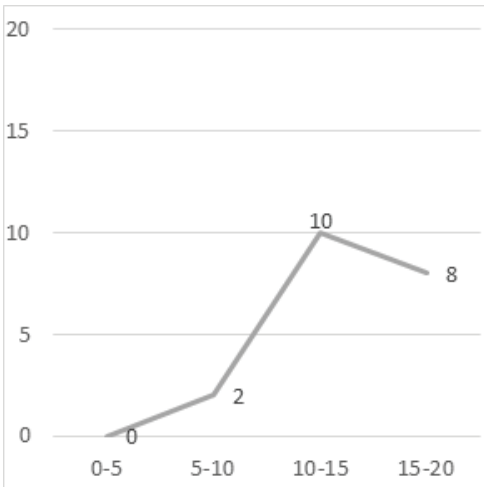
دسته	فروانی	درصد
۰ تا ۵	۰	۰
۵ تا ۱۰	۲	٪۱۰
۱۰ تا ۱۵	۱۰	٪۵۰
۱۵ تا ۲۰	۸	٪۴۰
۰ تا ۲۰	۲۰	٪۱۰۰

نمودار هیستوگرام و چند ضلعی

توزیع فراوانی نمرات دانشجویان را می‌توان علاوه بر جدول به شکل نمودار نیز نشان داد. هیستوگرام به شکل چند مستطیل به هم چسبیده است و توزیع فراوانی متغیرهای «پیوسته»[1] مانند نمرات را نمایش می‌دهد. محور افقی، فاصله دسته‌ها و محورعمودی، فراوانی نمرات هر دسته است (شکل ۱-۳۱). اگر در همین نمودار، وسط طبقات را به یکدیگر متصل کنیم، نمودار چندضلعی رسم می‌شود (شکل ۲-۳۱).



شکل ۱-۳۱: نمودار فراوانی نمرات



شکل ۲-۳۱: نمودار چند ضلعی

1. Continuose

در آزمون‌های هنجاری معمولاً اکثر نمرات نزدیک نمره میانگین هستند و هرچه از میانگین دورتر می‌شوند، تعداد آنها کمتر می‌شود. بنابراین، توزیع نمرات، مشابه توزیع نرمال است که در آن دو قسمت منحنی، نسبت به خط وسط قرینه هستند و میانگین و نما و میانه بر هم منطبق می‌باشند. در این حالت، از نمره میانگین تا فاصله یک انحراف معیار از آن، ۳۴/۱۳ درصد نمرات قرار می‌گیرند. پس بین منفی یک انحراف معیار تا مثبت یک انحراف معیار، ۶۸/۲۶ درصد نمرات قرار دارند. این مقادیر برای مثبت–منفی دو و مثبت–منفی سه انحراف معیار در شکل ۳–۳۱ نشان داده شده‌اند.

در آزمون‌های معیاری اکثر فراگیران به اکثر سؤالات جواب می‌دهند و نمرات بالا زیاد است. بنابراین، نمودار توزیع نمرات آزمون‌های معیاری، چولگی منفی[1] دارد یعنی زنگوله آن به سمت راست متمایل است. (شکل ۴–۳۱).

شکل ۳–۳۱: نمودار توزیع نمرات و انحراف معیار آنها

شکل ۴–۳۱: توزیع نمرات با چولگی منفی، نرمال و مثبت

1. Negative skewed

نمودار ستونی و دایره‌ای

برای نمایش متغیرهای «گسسته»[1] می‌توان از نمودار ستونی[2] یا دایره‌ای[3] استفاده کرد. به عنوان مثال، اگر در یک امتحان با شرکت ۲۰ دانشجو، دو نفر در امتحان رد شده باشند، میزان قبولی و ردی دانشجویان را می‌توان به صورت شکل‌های ۵-۳۱ و ۶-۳۱ نشان داد:

شکل ۵-۳۱: نمودار ستونی نمرات

شکل ۶-۳۱: نمودار دایره‌ای نمرات

نمرات استاندارد

محاسبه نمرات استاندارد[4] این کاربرد را دارد که موقعیت نمره یک دانشجو در مقایسه با موقعیت سایر شرکت‌کنندگان تعیین می‌شود. به عنوان مثال دانشجویی را در نظر بگیرید که در آزمون آناتومی نمره ۱۷ و در آزمون بیوشیمی نمره ۱۵ را کسب کرده است. درست است که نمره او در بیوشیمی کمتر است اما نمی‌توان گفت که واقعاً در آزمون آناتومی عملکرد بهتری داشته است زیرا این احتمال وجود دارد که آزمون بیوشیمی سخت‌تر بوده باشد. همچنین نمی‌توان دریافت که وضعیت او نسبت به سایر دانشجویان چطور است. برای اینکه بتوان این قیاس‌ها را انجام داد، نمرات خام را به نمرات استاندارد تبدیل می‌کنیم.

نمره Z توزیعی است که میانگین آن صفر و انحراف معیار آن ۱ است. برای محاسبه نمره Z از این فرمول استفاده می‌شود:

$$\text{Z score} = \frac{\text{میانگین نمرات} - \text{نمره دانشجو}}{\text{انحراف معیار نمرات}}$$

اگر میانگین و انحراف معیار آزمون آناتومی به ترتیب ۱۳ و ۲ و میانگین و انحراف معیار آزمون بیوشیمی به ترتیب ۱۶ و ۱ باشد، نمره Z دانشجوی فوق در آزمون آناتومی و بیوشیمی به شکل زیر قابل محاسبه است:

1. Discrete
2. Bar chart
3. Pie chart
4. Standard score

$$Z\ score = \frac{۱۵-۱۶}{۱} - \frac{-۱}{۱} = -۱ \qquad Z\ score = \frac{۱۷-۱۳}{۲} - \frac{+۴}{۲} = +۲$$

همان‌طور که مشخص است، نمره دانشجو در آزمون آناتومی به اندازه دو انحراف معیار بالاتر از میانگین است در حالی که در آزمون بیوشیمی به اندازه یک انحراف معیار پایین‌تر از متوسط کلاس است. پس توانستیم نمره خام دانشجو را به شکلی دربیاوریم که وضعیت او را در هر یک از دروس با سایر دانشجویان کلاس مقایسه کند. برای مقایسه وضعیت او در دو درس، محاسبه نمره T تفسیر آسان‌تری به دست می‌دهد. نمره T توزیعی است که میانگین آن ۵۰ و انحراف معیار آن ۱۰ است:

$$T\ score = ۱۰\ Z + ۵۰$$

نمره T این دانشجو به ترتیب در درس آناتومی و بیوشیمی به صورت زیر است:

$$T\ score = (۱۰\times(+۲)) + ۵۰ = ۷۰ \qquad T\ score = (۱۰\times(-۱)) + ۵۰ = ۴۰$$

نمرات ۷۰ و ۴۰ به صورت مستقیم قابل مقایسه با یکدیگر هستند. از آنجا که انحراف معیار نمره استاندارد T، ۱۰ است، می‌توان به راحتی گفت که در آزمون آناتومی ۲ انحراف معیار بالای میانگین و در آزمون بیوشیمی یک انحراف معیار پایین میانگین است که این نتیجه از روی نمره Z نیز به دست آمد.

ضریب همبستگی

یکی از مواردی که در تفسیر نتایج آزمون کمک‌کننده است و اطلاعات مفیدی در اختیار مسؤولان برگزاری آزمون می‌گذارد، همبستگی[1] بین متغیرهای مختلف آزمون‌ها است. بسته به این که نوع متغیرها چیست، از فرمول‌های مختلف برای تعیین رابطه بین آنها استفاده می‌شود. در اینجا در مورد ضریب همبستگی پیرسون[2] که کاربرد بیشتری در تعیین روایی و پایایی ابزارها دارد و همچنین ضریب همبستگی «دورشته‌ای نقطه‌ای»[3] که در بحث ضریب تمیز بحث مورد استفاده قرار خواهد گرفت، توضیحاتی ارائه می‌شود.

ضریب همبستگی پیرسون

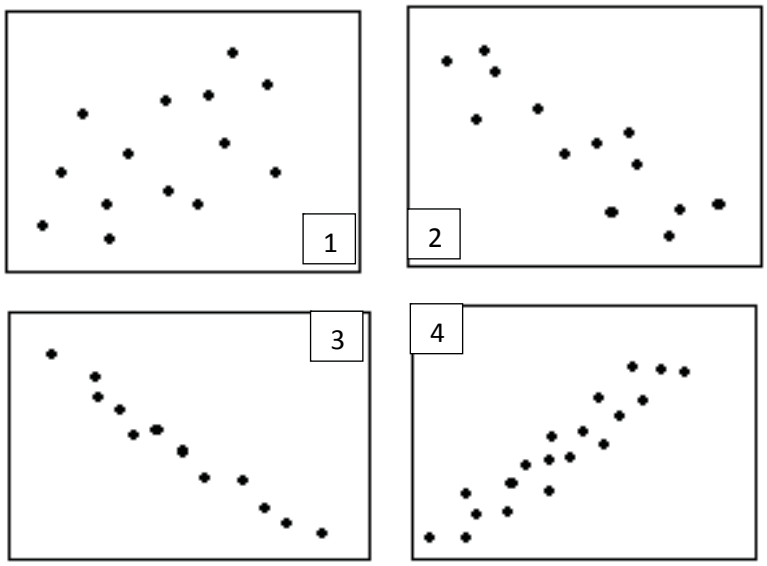
این ضریب بسیار متداول است و برای تعیین رابطه بین دو متغیر «پیوسته» به کار می‌رود. به عنوان مثال، به منظور سنجش همبستگی بین نمرات آزمون جامع علوم پایه یک دوره از دانشجویان با نمرات آزمون جامع پیش‌کاروزی همان دانشجویان از ضریب همبستگی پیرسون استفاده می‌شود. می‌توان نمرات دو آزمون را در یک نمودار پراکندگی نشان داد؛ به گونه‌ای که (به عنوان مثال) نمرات علوم پایه در محور عمودی و نمرات پیش‌کارورزی در محور افقی قرار گیرند (شکل ۷-۱۹).

اندازه همبستگی، میزان نزدیکی نمرات دو آزمون به یکدیگر را در این نمودار نشان می‌دهد. این ضریب که با r نشان داده می‌شود، می‌تواند مقادیر منفی یک تا مثبت یک را به خود اختصاص دهد. ضریب مثبت به معنای وجود رابطه مستقیم بین دو متغیر است و بیانگر این است که دو متغیر در یک راستا حرکت می‌کنند. به عبارت دیگر، اگر یکی از متغیرها افزایش (یا کاهش) یابد، متغیر دیگر نیز افزایش (یا کاهش) می‌یابد. ضریب منفی به معنای وجود رابطه معکوس است. یعنی اگر یک متغیر افزایش یابد، متغیر دیگر کاهش می‌یابد و بالعکس. زمانی که ضریب همبستگی دقیقاً برابر مثبت یک است، به این معناست که نمرات روی نمودار پراکندگی دقیقاً روی خطی با شیب مثبت یک قرار گرفته‌اند. هنگامی که ضریب معادل

1. Correlation
2. Pearson
3. point biserial correlation

صفر است، نشان می‌دهد که بین دو متغیر رابطه خطی وجود ندارد. توجه به این نکته حائز اهمیت است که صفر بودن ضریب همبستگی تنها عدم وجود رابطه خطی بین دو متغیر را نشان می‌دهد و نمی‌توان نتیجه گرفت که لزوماً دو متغیر از یکدیگر مستقل هستند. (شکل ۸-۳۱).

شکل ۷-۳۱: نمودار پراکندگی نمرات دانشجویان در دو آزمون



شکل ۸-۳۱: شماره ۱) رابطه ضعیف و مثبت، شماره ۲) رابطه ضعیف و منفی، شماره ۳) رابطه قوی و منفی، شماره ۴) رابطه قوی و مثبت

همبستگی بین دو متغیر، لزوماً بیانگر رابطه علت و معلولی بین متغیرها نیست و تنها نمایانگر این است که تغییر در یک متغیر موجب چه تغییری در متغیر دیگر می‌شود. به این معنا که افزایش یا کاهش یک متغیر چه تاثیری بر افزایش یا کاهش دیگری دارد. بنابراین، یکی از کاربردهای آن پیش‌بینی تغییرات متغیر مورد نظر است. هر چه قدر مطلق ضریب بزرگتر باشد، شدت رابطه بیشتر است و حاکی از آن است که نمرات نزدیک یک خط متمرکز شده‌اند.

برای آشنایی بیشتر با مفهوم این ضریب به نحوه محاسبه آن در مثال ذکر شده توجه کنید. نمرات پنج دانشجو در دو آزمون جامع علوم پایه و پیش کارورزی به صورت جدول ۴-۳۱ است:

جدول ٤-۳۱ : نمرات پنج دانشجو در آزمون جامع علوم پایه و پیش‌کارورزی

شماره دانشجو	نمره آزمون جامع علوم پایه	نمره آزمون جامع پیش کارورزی
۱	۱۰۰	۹۸
۲	۱۲۰	۱۰۶
۳	۱۳۰	۱۳۰
۴	۱۴۰	۱۲۰
۵	۱۵۰	۱۴۵

شماره دانشجو	۱	۲	۳	۴	۵	جمع	مجذور جمع
نمره آزمون جامع علوم پایه	۱۰۰	۱۲۰	۱۳۰	۱۲۵	۱۴۰	۶۱۵	۳۷۸۲۲۵
نمره آزمون جامع پیش کارورزی	۹۸	۱۰۶	۱۳۰	۱۲۰	۱۴۵	۵۹۹	۳۵۸۸۰۱
حاصل ضرب دو دسته نمرات	۹۸۰۰	۱۲۷۲۰	۱۶۹۰۰	۱۵۰۰۰	۲۰۳۰۰	۷۴۷۲۰	-
مجذور نمره آزمون جامع علوم پایه	۱۰۰۰۰	۱۴۴۰۰	۱۶۹۰۰	۱۵۶۲۵	۱۹۶۰۰	۷۶۵۲۵	-
مجذور نمره آزمون جامع پیش کارورزی	۹۶۰۴	۱۱۲۳۶	۱۶۹۰۰	۱۴۴۰۰	۲۱۰۲۵	۷۳۱۶۵	-

برای محاسبه ضریب همبستگی ابتدا باید حاصل ضرب دو دسته نمرات، مجذور نمره آزمون جامع علوم پایه، مجذور نمره آزمون جامع پیش کارورزی و مجذور جمع نمرات را به شرح زیر محاسبه کنیم:

$$\text{ضریب همبستگی پیرسون} = \frac{\text{(مجموع مجذورات دسته دوم} \times \text{مجموع نمرات دسته اول)} - \text{(مجموع حاصل ضرب دو دسته نمرات} \times \text{تعداد دانشجویان)}}{\sqrt{\text{[مجذور حاصل جمع دسته دوم} - \text{(مجموع مجذورات دسته دوم} \times \text{تعداد)]}\ \text{[مجذور حاصل جمع دسته اول} - \text{(مجموع مجذورات دسته اول} \times \text{تعداد)]}}}$$

$$\text{ضریب همبستگی پیرسون} = \frac{(5 \times 74720) - (615 \times 599)}{\sqrt{[(5 \times 76525) - 378225]\ [5 \times 73165)) - 358801]}} = \frac{5215}{\sqrt{30905600}} = \frac{5215}{5560} = 0.93$$

نتیجه نشان می‌دهد که همبستگی بین نمرات آزمون جامع علوم پایه و آزمون جامع پیش‌کارورزی در این مثال، مثبت و بالاست.

همبستگی دو رشته‌ای نقطه‌ای

همان‌طور که ذکر شد، ضریب همبستگی بین دو متغیر بنا به نوع متغیرها به صورت‌های مختلفی محاسبه می‌شود.

روشی که برای محاسبه همبستگی بین یک متغیر دوحالته[1] و یک متغیر پیوسته به کار می‌رود، همبستگی دورشته‌ای نقطه‌ای[2] است که فرمول آن به صورت زیر است:

از این ضریب همبستگی برای تعیین ضریب تمیز سؤال استفاده می‌شود (با در نظر گرفتن نمره کل آزمون متغیر پیوسته و نمره سؤال متغیر دوحالته) که در جای خود مفصلاً بحث خواهد شد.

$$\frac{\text{تعداد دانشجویانی که واجد متغیر دو حالته بوده‌اند}}{\text{تعداد دانشجویانی که واجد متغیر دو حالته نبوده‌اند}} \sqrt{\frac{\text{میانگین متغیر پیوسته در دانشجویانی که واجد متغیر دو حالته بوده‌اند} - \text{میانگین متغیر پیوسته در همه دانشجویان}}{\text{انحراف معیار متغیر پیوسته}}} =$$

رگرسیون خطی

در نمودار پراکندگی دو متغیر نسبت به هم، خطوط زیادی را می‌توان رسم کرد اما برخی از آنها رابطه بین دو متغیر را بهتر پیش‌بینی می‌کنند. با استفاده از رگرسیون بهترین خط و معادله برای پیش‌بینی رابطه دو متغیر به دست می‌آید. هدف تحلیل رگرسیون، پیش‌بینی تغییرات یک یا چند متغیر وابسته (ملاک) با توجه به تغییرات متغیرهای مستقل (پیش‌بین) است.

این معادله به صورت ساده به شکل زیر نوشته می‌شود که در آن، Y و X به ترتیب متغیرهای وابسته و مستقل، a شیب خط و b عرض از مبدا هستند:

$$y=aX+b$$

با جایگزین کردن متغیر X با اندازه‌های مورد نظر می‌توان اندازه متغیر Y را پیش‌بینی کرد.

یکی از کاربردهای رگرسیون خطی، استفاده از آن به عنوان یکی از روش‌های تعیین استاندارد آزمون است که در بخش هفتم کتاب توضیح داده شد. همچنین کاربرد دیگر آن در بحث متریک‌های آزمون OSCE است که در فصل بعدی همین بخش مورد بحث قرار خواهد گرفت.

1. Dichotomous
2. Point biserial correlation

فصل ۳۲

تحلیل آیتم

در فصل قبلی به مرور شاخص‌هایی پرداختیم که در سطح کل آزمون محاسبه می‌شدند اما برای تحلیل جامع‌تر به شاخص‌هایی نیاز داریم که اطلاعاتی خوبی در مورد تک‌تک آیتم‌های آزمون نیز در اختیار ما بگذارند. منظور از آیتم، سؤالات امتحان چندگزینه‌ای، شفاهی، تشریحی و حتی ایستگاه‌های یک OSCE است. برخی از شاخص‌هایی که در این فصل مرور می‌شوند، برای تمام انواع آزمون‌ها قابل محاسبه هستند و بعضی از آنها اختصاصی یک آزمون هستند.

در این فصل شاخص‌هایی که بر اساس نظریه کلاسیک آزمون به دست می‌آیند، مورد بحث قرار می‌گیرند. شاخص‌های مربوط به نظریه سؤال پاسخ در فصل آخر این بخش توضیح داده می‌شوند.

ضریب جذب گزینه‌ها

همان‌طور که در بخش دوم کتاب عنوان شد، در آزمون‌های چندگزینه‌ای یکی از مواردی که می‌تواند اطلاعات خوبی در مورد سؤال در اختیار طراح قرار دهد، محاسبه درصد جذب هر گزینه است. گزینه‌ای که توسط هیچ‌کدام از دانشجویان انتخاب نشود، مناسب نیست. هنگام طراحی سؤال چنانچه استادان گزینه‌های مناسبی پیدا نکنند، به مواردی روی می‌آورند که گزینه‌های انحرافی ضعیفی هستند و دانشجویان به سرعت نادرستی آنها را تشخیص می‌دهند. به جدول ۱-۳۲ زیر توجه کنید:

جدول ۱-۳۲ : توزیع پاسخ‌های ۱۲۰ دانشجو به دو سؤال از یک آزمون چندگزینه‌ای

		تعداد دانشجویانی که گزینه ... را انتخاب کرده‌اند.				
شماره سؤال	پاسخ صحیح	الف	ب	ج	د	نزده
۱	د	۰	۱۸	۲۲	۸۰	۰
۲	الف	۲۰	۲۵	۳۰	۳۲	۱۳

در سؤال اول، هیچ کس گزینه الف را انتخاب نکرده است. این نشان می‌دهد که اشتباه بودن آن به طرز واضحی برای همه دانشجویان مشخص بوده است. بهتر است برای استفاده در آزمون‌های بعدی گزینه الف با گزینه مناسب‌تری جایگزین شود. چنانچه تعداد کمی از دانشجویان (معمولاً کمتر از ۵ درصد) جذب گزینه‌ای شوند، به این معنا است که گزینه انحرافی جذاب نبوده است که به آن گزینه غیرعملکردی[1] می‌گویند

1. Non-functioning

تارنت و همکاران ۲۰۰۹

در این مطالعه، در دانشگاهی در هنگ‌کنگ تمام آزمون‌های چندگزینه‌ای به عمل آمده در رشته پرستاری بین سال های ۲۰۰۱ تا ۲۰۰۵ جمع‌آوری شد (۱۲۱ آزمون). از بین آنها آزمون‌هایی انتخاب شدند که حداقل تعداد ۵۰ سؤال و پایایی بالای ۷۰ داشتند. نهایتاً هفت آزمون شامل ۵۱۴ سؤال چهارگزینه‌ای که مجموعاً ۲۰۵۶ گزینه داشتند، مورد بررسی قرار گرفتند. تمام این امتحانات بر اساس بلوپرینت و اهداف دوره طراحی شده بودند و قبل از اجرا توسط پانلی از استادان مرور شده بودند. در تحلیل آزمون، فراوانی گزینه‌های غیرعملکردی و ضریب تمیز سؤالات محاسبه شد.

بر اساس نتایج به دست آمده، ۵۴۱ گزینه (۳۵/۱ درصد)، غیرعملکردی بودند. ۴۷۲ گزینه (۳۰/۶ درصد)، ضریب تمیز مثبت داشتند. ۱۰/۲ درصد گزینه‌ها آنقدر غیرجذاب بودند که توسط هیچ دانشجویی انتخاب نشدند. تنها در ۱۳/۸ درصد موارد، هر سه گزینه انحرافی به صورت قابل‌قبول عمل کرده بودند. به صورت متوسط هر سؤال بین ۱/۳۵ تا ۱/۷۴ گزینه (یعنی کمتر از دو گزینه) عملکردی داشت. به صورت کلی، سؤالاتی که گزینه عملکردی بیشتری داشتند، بهتر قادر به تمایز بین دانشجویان قوی و ضعیف بودند.

ضریب دشواری آیتم

ضریب دشواری[1] که به صورت P نشان داده می‌شود و به آن p value نیز گفته می‌شود، برای یک آیتم (به عنوان مثال، یک سؤال در آزمون چندگزینه‌ای یا یک ایستگاه در OSCE) تعریف می‌شود. ضریب دشواری، درصد شرکت‌کنندگانی است که به آیتم مورد نظر جواب درست داده‌اند و می‌تواند مقداری بین صفر تا یک داشته باشد. به این ترتیب، هر چقدر تعداد دانشجویانی که به آیتم پاسخ صحیح داده باشند، بیشتر باشد، منطقاً آن آیتم آسان‌تر است در حالی که بنا به تعریف، ضریب دشواری آن بزرگتر است.

امروزه با دسترس بودن نرم‌افزارهای متنوع، محاسبه ضریب دشواری به صورت غیردستی صورت می‌گیرد. در واقع، اکثر نرم افزارهایی که برای تصحیح پاسخنامه‌ها به کار می‌روند، دارای افزونه‌ای به این منظور هستند و همزمان با اعلام نمرات دانشجویان، قادر هستند گزارشی از ضریب دشواری سؤالات ارائه دهند. هرچند گاهی استاد قصد دارد خودش ضریب شواری سؤالات را برای یک آزمون کلاسی محاسبه کند. در هر دو حالت، روش محاسبه تفاوتی ندارد.

برای محاسبه ضریب دشواری سؤالات در آزمون چندگزینه‌ای، باید تعداد کل افرادی که به هر سؤال پاسخ صحیح داده‌اند، مشخص کنیم و بر تعداد کل دانشجویان تقسیم کنیم:

$$\text{ضریب دشواری سوال} = \frac{\text{دانشجویانی که به سوال پاسخ صحیح داده‌اند}}{\text{تعداد کل دانشجویان}}$$

به مثال زیر که در خصوص داده‌های جدول ۱-۳۲ است، توجه کنید:

ضریب دشواری هر سؤال مطابق فرمول بالا محاسبه شده است.

$$\text{ضریب دشواری سوال یک} = \frac{۸۰}{۱۲۰} = ۰/۶۶ \qquad \text{ضریب دشواری سوال دو} = \frac{۲۰}{۱۲۰} = ۰/۱۶$$

در این میان، سؤال شماره دو سخت‌تر از سؤال شماره یک است.

در آزمون OSCE نیز می‌توان حدنصاب قبولی را در سطح ایستگاه تعیین کرد و به این ترتیب مشخص کرد که در هر ایستگاه چند نفر قبول شده‌اند. به عنوان مثال، در یک آزمون OSCE از بین ۱۰۰ دانشجو ۷۵ نفر توانسته‌اند در ایستگاه سونداژ قبول شوند پس ضریب دشواری ایستگاه ۷۵ درصد است. همان‌طور که مشخص است، به سادگی می‌توان گفت درصد قبولی ایستگاه، معادل ضریب دشواری آن است.

1. Difficulty index

برای محاسبه دستی ضریب دشواری، در صورتی که تعداد دانشجویان زیاد باشد (بیش از ۴۰ نفر)، سخت است که پاسخنامه همه دانشجویان بررسی شود. به همین دلیل، دانشجویان را به دو دسته ضعیف و قوی تقسیم می‌کنند. برای این کار دانشجویان بر اساس نمره کل ردیف می‌شوند. سپس، برگه‌های یک سوم بالایی (گروه قوی) و یک سوم پایینی (گروه ضعیف) جدا می‌شوند. ضریب دشواری به صورت فرمول زیر قابل محاسبه است:

$$\text{ضریب دشواری سوال} = \frac{\text{تعداد دانشجویان ضعیف که درست جواب داده‌اند+ تعداد دانشجویان قوی که درست جواب داده‌اند}}{\text{تعداد دانشجویان دو گروه}}$$

به عنوان مثال، در یک آزمون دانشجویان بر اساس نمره کل به دو گروه ۴۰ نفره قوی و ضعیف تقسیم شدند. در گروه قوی ۳۸ دانشجو به سؤال شماره ۲۰ پاسخ صحیح داده‌اند و در گروه ضعیف ۱۴ نفر به این سؤال شماره پاسخ صحیح داده‌اند. ضریب دشواری این سؤال به صورت زیر محاسبه می‌شود:

$$\text{ضریب دشواری سوال} = \frac{۳۰+۴۰}{۸۰} = \frac{۵۲}{۸۰} = ۰/۶۵$$

در آزمون‌های معیارمحور، به جای تقسیم‌بندی دانشجویان به دو گروه ضعیف و قوی، آنها را به دو گروه رد و قبول تقسیم می‌کنیم اما نحوه محاسبه ضریب دشواری برای این آزمون تغییری نمی‌کند:

$$\text{ضریب دشواری سوال} = \frac{\text{تعداد دانشجویان رد که درست جواب داده‌اند + تعداد دانشجویان قبول که درست جواب داده‌اند}}{\text{تعداد دانشجویان دو گروه}}$$

به مثال زیر توجه کنید:

در یک آزمون چندگزینه‌ای پایان ترم (که معیارمحور است)، ۲۰ دانشجو قبول و ۵ دانشجو رد شده‌اند. تعداد کسانی که به سؤال شماره ۱۰ جواب درست داده‌اند، به ترتیب در گروه قبول و رد، ۹ و ۱ نفر بوده است. ضریب دشواری این سؤال به صورت زیر حساب می‌شود:

$$\text{ضریب دشواری سوال} = \frac{۹+۱}{۲۵} = \frac{۱۰}{۲۵} = ۰/۴$$

به مثال زیر در خصوص آزمون OSCE توجه کنید. در یک آزمون OSCE از بین ۱۰۰ دانشجو ۹۵ نفر توانسته‌اند در کل آزمون قبول شوند. میزان قبولی در ایستگاه تعبیه سونداژ در گروه قبول، ۷۳ و در گروه رد، ۲ نفر بوده است. ضریب دشواری ایستگاه به صورت زیر قابل محاسبه است:

$$\text{ضریب دشواری ایستگاه} = \frac{۷۳+۲}{۱۰۰} = \frac{۷۵}{۱۰۰} = ۰/۷۵$$

آنچه تا اینجا ذکر شد، برای آزمون‌هایی کاربرد دارد که جواب دوارزشی دارند مانند سؤالات چندگزینه‌ای. به عبارت دیگر، دانشجویان در مواجهه با هر سؤال یا به آن پاسخ صحیح داده‌اند و تمام نمره را دریافت کرده‌اند یا پاسخ اشتباه داده‌اند و صفر شده‌اند. بنابراین همین روش برای تعیین ضریب دشواری در یک آزمون با سؤالات درست–نادرست و جورکردنی گسترده نیز می‌تواند به کار رود.

در آزمون‌هایی که پاسخ به صورت صفر و یک نیست، قضیه کمی متفاوت است. آزمون تشریحی، سؤالات شفاهی و OSCE از این دست سؤالات هستند که نمره‌ای که می‌تواند به پاسخ آنها تعلق بگیرد، به صورت یک محدوده نمره است.

محاسبه ضریب دشواری این سؤالات به صورت محاسبه میانگین نمرات دانشجویان در آن سؤال است که بر نمره کل سؤال (یعنی نمره‌ای که دانشجویان می‌توانستند به صورت بالقوه کسب کنند)، تقسیم می‌شود:

$$\text{ضریب دشواری سوال} = \frac{\text{میانگین نمرات دانشجویان}}{\text{نمره سوال}}$$

دو ایستگاه OSCE یا دو سؤال تشریحی را در نظر بگیرید که حداکثر نمره در آنها ۱۰ است. میانگین نمرات آنها را محاسبه می‌کنیم که ۷/۵ و ۳ به دست می‌آید. بنابراین، ضریب دشواری دو ایستگاه/دو سؤال به ترتیب ۰/۷۵ و ۰/۳ است. همان‌طور که مشخص است ایستگاه با ضریب دشواری ۰/۳ برای دانشجویان سخت‌تر بوده است.

در صورتی که نمرات دانشجویان را به صورت دسته‌بندی شده داشته باشیم، می‌توان از فرمول زیر که در واقع تغییریافته فرمول بالاست، استفاده کرد:

$$\text{ضریب دشواری سوال} = \frac{\text{میانگین نمرات دانشجویان}}{\text{نمره سوال}} = \frac{\text{مجموع نمرات دانشجویان}}{\text{نمره سوال} \times \text{تعداد دانشجویان}}$$

به عنوان مثال، در یک آزمون تشریحی با شرکت ۵۰ دانشجو، یک سؤال تشریحی سه نمره دارد و مقادیری که برای پاسخ‌های مختلف دانشجویان در آن می‌توان در نظر گرفت، شامل ۰، ۱، ۲ و ۳ است. با بررسی برگه‌های دانشجویان، مشخص می‌شود که به ترتیب ۱۰، ۲۵، ۱۰ و ۵ دانشجو نمرات فوق را کسب کرده‌اند.

$$\text{ضریب دشواری سوال} = \frac{(۱۰\times۰) + (۲۵\times۱) + (۱۰\times۳) + (۵\times۳)}{۵۰\times۳} = \frac{۰ + ۲۵ + ۳۰ + ۱۵}{۱۵۰} = \frac{۷۰}{۱۵۰} = ۰/۴۶$$

در این گونه سؤالات نیز اگر به علت تعداد زیاد دانشجویان آنان را به دو گروه قوی و ضعیف تقسیم می‌کنیم، می‌توان از فرمول زیر استفاده کرد:

$$\text{ضریب دشواری سوال} = \frac{\text{میانگین نمرات دانشجویان ضعیف} + \text{میانگین نمرات دانشجویان قوی}}{\text{نمره سوال} \times \text{تعداد دانشجویان دو گروه}}$$

به مثال جدول ۲-۳۲ توجه کنید:

جدول ۲-۳۲ : توزیع نمرات دانشجویان در یک ایستگاه OSCE با چک‌لیست ۱۰ نمره‌ای

نمره	۰	۱	۲	۳	٤	٥	٦	۷	۸	۹	۱۰
تعداد دانشجویان گروه قوی	۰	۰	۱	۰	۴	۵	۸	۸	۵	۵	۴
تعداد دانشجویان گروه ضعیف	۱	۳	۵	۷	۵	۸	۴	۵	۲	۰	۰

مجموع نمرات گروه قوی:

$$= (۰\times۰) + (۰\times۱) + (۱\times۲) + (۰\times۳) + (۴\times۴) + (۵\times۵) + (۸\times۶) + (۸\times۷) + (۵\times۸) + (۵\times۹) + (۴\times۵)$$

$$= ۰+۰+۲+۰+۱۶+۲۵+۴۸+۵۶+۴۰+۴۵+۲۰ = ۲۷۲$$

مجموع نمرات گروه ضعیف:

$$= (۱\times۰) + (۳\times۱) + (۵\times۲) + (۷\times۳) + (۵\times۴) + (۸\times۵) + (۴\times۶) + (۵\times۷) + (۲\times۸) + (۰\times۹) + (۰\times۵)$$

$$= ۰+۳+۱۰+۲۱+۲۰+۴۰+۲۴+۳۵+۱۶+۰+۰ = ۱۶۹$$

ضریب دشواری ایستگاه:

$$\text{ضریب دشواری سوال} = \frac{۱۶۹+۲۷۲}{۸۰\times۱۰} = \frac{۴۴۱}{۸۰۰} = ۰/۵۵$$

اولین مسأله‌ای که در تفسیر ضریب دشواری باید مدنظر قرار بگیرد، نوع آزمون از لحاظ هدفی است که دنبال می‌کند. باید مشخص شود که آیا آزمون مبتنی بر معیار بوده است یا نتایج آن به طور هنجاری تفسیر می‌شوند.

در آزمون هنجاری، بهتر است نمرات پراکندگی خوبی داشته باشند یعنی واریانس آنها بالا باشد. زیرا اگر نمرات در یک محدوده باریک تجمع پیدا کنند، یعنی نمرات دانشجویان نزدیک هم یا مشابه شود، نمی‌توان تمایز بین دانشجویان قائل شد و رتبه‌بندی درستی انجام داد. واریانس سؤال با ضریب دشواری (P) رابطه‌ای دارد که به صورت زیر است:

$$\text{واریانس سوال} = P \times (۱ - P)$$

طبق فرمول فوق، اگر ضریب دشواری سؤال ۱ یا ۰ باشد یعنی سؤال بسیار آسان یا بسیار سخت باشد، واریانس به کمترین مقدار می‌رسد و صفر می‌شود. هنگامی واریانس به حداکثر مقدار خود می‌رسد که ضریب دشواری برابر ۰/۵ باشد. در این صورت واریانس سؤال ۰/۲۵ می شود. بنابراین می‌توان گفت که در آزمون هنجاری بیشتر سؤالاتی مورد نیاز هستند که واریانس بزرگتری داشته باشند یعنی میزان دشواری آنها متوسط باشد.

در مورد سؤالات چندگزینه‌ای از آنجا که یکی از نگرانی‌های رایج، پاسخ دادن دانشجویان بر اساس حدس و گمان است، تلاش شده است تا این دغدغه لحاظ شود و برای جلوگیری از آن گفته می‌شود بهتر است ضریب دشواری آزمون هنجاری چندگزینه‌ای کمی بالاتر از سطح متوسط باشد. برای به دست آوردن عدد دقیق‌تر، باید احتمال پاسخ حدسی را برآورد کرد. احتمال اینکه دانشجویان با حدس به پاسخ سؤال چهارگزینه‌ای برسند، ۲۵ درصد است. بنابراین میزان پایه به جای صفر، ۰/۲۵ در نظر گرفته می‌شود. در این حالت، حدوسط یعنی نقطه میانی بین ۰/۲۵ و ۱ محاسبه می‌شود که ۰/۳۷۵ است. به عبارت دیگر، ضریب دشواری مطلوب آزمون هنجاری چهارگزینه‌ای ۰/۶۲۵ (۰/۲۵+۰/۳۷۵) خواهد بود. به صورت کلی در آزمون‌های هنجاری، ضریب دشواری بین ۰/۳ تا ۰/۷ بیشترین اطلاعات را در مورد تفاوت دانشجویان به دست می‌دهد (سیف ۱۳۹۰). هر چند که عوامل دیگری نیز تعیین‌کننده هستند.

در آزمون معیاری، مانند ارزیابی پایان نیمسال، به طور معمول انتظار می‌رود که اگر دانشجویان به صورت فعال در دوره مشارکت کرده باشند و فعالیت‌های یاددهی–یادگیری را انجام داده باشند، مطالب را فرا بگیرند و به اهداف دوره برسند. از طرفی سؤالات نیز به گونه‌ای طراحی می‌شوند که دانشجویانی که در دوره شرکت کرده‌اند و منابع را مطالعه نموده‌اند، بتوانند به آنها پاسخ دهند. به همین دلیل، اکثر دانشجویان به اکثر سؤالات جواب صحیح می‌دهند. به عبارت دیگر، عجیب

نیست که در آزمون معیاری، ضریب دشواری سؤالات بزرگ و نزدیک ۱ باشد.

پس این باور که ضریب دشواری باید متوسط و بین ۰/۳ تا ۰/۷ باشد، هرچند رایج است اما دقیق نیست. از آنجا که در ارزیابی فراگیران علوم پزشکی، جز چند نمونه محدود، اکثراً با آزمون‌های معیاری سر و کار داریم، توجه به این نکته قبل از تفسیر ضریب دشواری ضروری است.

ضریب تمیز آیتم

ضریب تمیز[1] که به صورت D نشان داده می‌شود، مانند ضریب دشواری برای یک آیتم تعریف می‌شود. ضریب تمیز، مشخص می‌کند که یک آیتم تا چه میزان قادر به ایجاد تمایز و افتراق بین دانشجویان ضعیف و قوی است. هرچقدر ضریب تمیز بزرگتر باشد، به این معنا است که قدرت آیتم در جدا کردن دانشجویان قوی از ضعیف بیشتر بوده است. ضریب تمیز می‌تواند مقداری بین منفی یک تا مثبت یک داشته باشد.

به منظور محاسبه ضریب تمیز در سؤالات بسته پاسخ، شرکت‌کنندگان باید بر اساس نمره‌ای که در کل آزمون کسب کرده‌اند، به دو دسته پایین (ضعیف) و بالا (قوی) تقسیم شوند. سپس باید مشخص شود که چه تعداد از هر دسته به آیتم مورد نظر جواب درست داده‌اند. تفاضل این دو مقدار، ضریب تمیز را مشخص می‌کند:

$$\text{ضریب تمیز سوال} = \frac{\text{تعداد دانشجویان ضعیف که درست جواب داده‌اند} - \text{تعداد دانشجویان قوی که درست جواب داده‌اند}}{\text{تعداد دانشجویان یک گروه}}$$

به مثال زیر توجه کنید:

در یک آزمون چهارگزینه‌ای، دانشجویان بر اساس نمره کل به دو گروه ۴۰ نفره قوی و ضعیف تقسیم شدند که به ترتیب ۳۸ و ۱۴ نفر ایشان به سؤال ۲۰ پاسخ صحیح داده‌اند. ضریب تمیز این سؤال به صورت زیر محاسبه می‌شود:

$$\text{ضریب دشواری سوال} = \frac{۳۸-۱۴}{۴۰} = \frac{۲۴}{۴۰} = ۰/۶$$

روش دوم محاسبه ضریب تمیز در آزمون‌های چندگزینه‌ای، استفاده از ضریب همبستگی نمره سؤال با نمره کل آزمون است. در واقع، سؤالی ضریب تمیز بالا دارد که توسط دانشجویان گروه قوی به خوبی جواب داده شده باشد و در دانشجویان گروه ضعیف نمره پایینی داشته باشد. از آنجا که گروه‌های قوی و ضعیف بر اساس نمره کل مشخص می‌شوند، همبستگی سؤال با نمره کل، نشان‌دهنده قدرت آن در تمایز بین دانشجویان ضعیف و قوی است. در فصل قبل ذکر شد که ضریب همبستگی بین دو متغیر بنا به نوع متغیرها به اشکال گوناگون محاسبه می‌شود. در آزمون چندگزینه‌ای، نمره هر سؤال به صورت صفر یا یک، یعنی یک متغیر دوحالته است. در حالی که نمره کل آزمون، یک متغیر پیوسته (مثلاً از صفر تا ۲۰) است. روشی که برای محاسبه همبستگی بین متغیر دوحالتی و متغیر پیوسته به کار می‌رود، همبستگی دورشته‌ای نقطه‌ای است که فرمول آن به صورت زیر است:

$$\sqrt{\frac{\text{تعداد دانشجویانی که به سوال پاسخ درست داده‌اند}}{\text{تعداد دانشجویانی که به سوال پاسخ غلط داده‌اند}}} = \frac{\text{میانگین نمره کل دانشجویانی که به سوال درست پاسخ داده‌اند} - \text{میانگین نمره کل همه دانشجویان}}{\text{انحراف معیار نمره کل}}$$

1. Discrimination index

برای توضیح بیشتر، مثال جدول ۳-۳۲ را در نظر بگیرید:

جدول ۳-۳۲ : نحوه پاسخ‌دهی دانشجویان به یک سؤال امتحان و نمره کل آزمون

شماره دانشجو	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
نمره سؤال	۰	۰	۱	۱	۱	۰	۱	۱	۱	۰
نمره کل	۱۴	۱۱/۵	۱۹	۱۹	۱۷	۱۵	۱۸/۵	۱۷	۱۵	۱۰

اگر ضریب تمیز سؤال مورد نظر را به همان روش اول حساب کنیم، عدد زیر به دست می‌آید:

$$\text{ضریب دشواری سوال} = \frac{۵-۱}{۵} = \frac{۴}{۵} = ۰/۸$$

طبق روش ضریب همبستگی دورشته‌ای نقطه‌ای، ضریب تمیز به صورت زیر قابل محاسبه است: میانگین نمره کل برای همه دانشجویان ۱۵/۶ و برای آن دسته از دانشجویان که به سؤال مورد نظر پاسخ صحیح داده‌اند، ۱۷/۵۸ است. انحراف معیار آزمون نیز ۲/۹۵ به دست می‌آید. بنابراین:

$$\text{ضریب تمیز سوال} = \frac{۱۷/۵۸-۱۵/۶}{۲/۹۵}\sqrt{\frac{۶}{۴}} = ۰/۸۲$$

دو روش فوق برای آزمون‌های هنجاری مناسب‌تر هستند. در آزمون‌های معیارمحور، منطقی‌تر است که تقسیم‌بندی دانشجویان به دو گروه ضعیف و قوی به منظور تعیین ضریب تمیز بر اساس قبولی یا ردی آنها در آزمون باشد؛ نه نمره کل آنها در آزمون. دانشجویان که حدنصاب قبولی را کسب کنند، در گروه قبول قرار می‌گیرند و آنهایی که نتوانند حداقل نمره را کسب کنند، در آزمون رد می‌شوند. پس ضریب تمیز در آزمون معیاری به این شکل محاسبه می‌شود:

$$\text{ضریب تمیز سوال} = \frac{\text{تعداد دانشجویان قبول که به سوال درست جواب داده‌اند}}{\text{تعداد دانشجویان قبول}} - \frac{\text{تعداد دانشجویان رد که به سوال درست جواب داده‌اند}}{\text{تعداد دانشجویان رد}}$$

در مثال جدول ۴-۳۲ ضریب تمیز یک سؤال چهارگزینه‌ای را با فرض یک آزمون معیارمحور به دست می‌آوریم. در ابتدا تعداد دانشجویانی که قبول و رد هستند را تعیین می‌کنیم (هرچند حدنصاب قبولی به روش‌های گوناگون تعیین می‌شود و بحث مفصل در خصوص آن در بخش هفتم کتاب ارائه شده است، در اینجا فرض می‌کنیم که مانند بسیاری از آزمون‌های معمول، حداقل نمره قبولی آزمون ۱۲ از ۲۰ می‌باشد).

جدول ٤-۳۲ : نمره دانشجویان در یک امتحان پایان ترم و یک سؤال از آن

شماره دانشجو	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
نمره سؤال	۰	۰	۱	۱	۱	۰	۱	۱	۱	۰
نمره کل	۱۴	۱۱/۵	۱۹	۱۹	۱۷	۱۵	۱۸/۵	۱۷	۱۵	۱۰

بر اساس داده‌های جدول ۴-۳۲ مشخص است که هشت نفر در آزمون قبول شده‌اند و دو نفر موفق به کسب حدنصاب نشده‌اند. از بین دانشجویان گروه قبول، شش نفر به سؤال پاسخ درست داده‌اند و از بین دانشجویان گروه رد، هیچ‌کس نتوانسته است به سؤال جواب دهد. بنابراین ضریب تمیز سؤال مورد نظر در آزمون فوق که معیارمحور است، به صورت زیر است:

$$\text{ضریب تمیز سوال} = \frac{۶}{۸} - \frac{۰}{۲} = ۰/۷۵$$

اکنون مثال دیگری را در نظر بگیرید که مربوط به یک OSCE معیار محور است. نتیجه عملکرد دانشجو هم در ایستگاه هم در کل آزمون به صورت رد و قبول (متغیر دوحالته) در نظر گرفته شده است. در یک OSCE از بین ۱۰۰ دانشجو ۹۵ نفر توانسته‌اند در کل آزمون قبول شوند. میزان قبولی در ایستگاه تعبیه سونداژ در گروه قبول، ۷۳ و در گروه رد، ۲ نفر بوده است.

$$\text{ضریب تمیز ایستگاه} = \frac{۷۳}{۹۵} - \frac{۲}{۵} = ۰/۷۶ - ۰/۴۰ = ۰/۳۶$$

محاسبه ضریب تمیز در آزمون‌های تشریحی یا OSCE که در آنها پاسخ سؤال یا ایستگاه به صورت صفر و یک نیست، به این ترتیب است که میانگین نمره دانشجویان در سؤال یا ایستگاه مورد استفاده قرار می‌گیرد.

چنانچه هدف آزمون معیارمحور باشد یعنی نتیجه کلی آزمون به صورت رد و قبول بیان شده باشد، از تفاضل نمره دو گروه رد و قبول استفاده می‌کنیم. یک ایستگاه OSCE در نظر بگیرید که حداکثر نمره آن ۱۰ است. میانگین نمرات گروه قبول و رد به ترتیب ۵/۵ و ۴ است. بنابراین:

$$\text{ضریب تمیز ایستگاه} = \frac{۵/۵-۴}{۱۰} - \frac{۱/۵}{۱۰} = ۰/۱۵$$

حال اگر آزمون هنجاری باشد، از تفاضل نمره دو گروه قوی و ضعیف (بالا و پایین) به صورت زیر استفاده می‌کنیم:

$$\text{ضریب تمیز سوال} = \frac{\text{میانگین نمرات دانشجویان ضعیف} - \text{میانگین نمرات دانشجویان قوی}}{\text{نمره سوال}}$$

به عنوان مثال، یک ایستگاه OSCE در نظر بگیرید که حداکثر نمره آن ۱۰ است. میانگین نمرات آنها را محاسبه می‌کنیم که در گروه قوی ۷/۵ و و در گرووه ضعیف ۳ به دست می‌آید. بنابراین:

$$\text{ضریب تمیز ایستگاه} = \frac{۷/۵-۳}{۱۰} - \frac{۴/۵}{۱۰} = ۰/۴۵$$

چنانچه نمرات دانشجویان دو گروه قوی و ضعیف را به صورت دسته‌بندی شده داریم، می‌توان فرمول بالا را به شکل زیر تغییر داد:

$$\text{ضریب تمیز سوال} = \frac{\text{مجموع نمرات دانشجویان ضعیف} - \text{مجموع نمرات دانشجویان قوی}}{\text{نمره سوال} \times \text{تعداد دانشجویان یک گروه}}$$

به مثال جدول ۵-۳۲ توجه کنید:

جدول ۵-۳۲ : توزیع نمرات دانشجویان در یک ایستگاه OSCE با چک‌لیست ۱۰ نمره‌ای

نمره	۰	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
تعداد دانشجویان گروه قوی	۰	۰	۱	۰	۴	۵	۸	۸	۵	۵	۴
تعداد دانشجویان گروه ضعیف	۱	۳	۵	۷	۵	۸	۴	۵	۲	۰	۰

مجموع نمرات گروه قوی:

$$= (۰\times۰) + (۰\times۱) + (۱\times۲) + (۰\times۳) + (۴\times۴) + (۵\times۵) + (۸\times۶) + (۸\times۷) + (۵\times۸) + (۵\times۹) + (۴\times۵)$$

$$= ۰+۰+۲+۰+۱۶+۲۵+۴۸+۵۶+۴۰+۴۵+۲۰ = ۲۷۲$$

مجموع نمرات گروه ضعیف:

$$= (۱\times۰) + (۳\times۱) + (۵\times۲) + (۷\times۳) + (۵\times۴) + (۸\times۵) + (۴\times۶) + (۵\times۷) + (۲\times۸) + (۰\times۹) + (۰\times۵)$$

$$= ۰+۳+۱۰+۲۱+۲۰+۴۰+۲۴+۳۵+۱۶+۰+۰ = ۱۶۹$$

ضریب دشواری ایستگاه:

$$\text{ضریب دشواری سوال} = \frac{۲۷۲+۱۶۹}{۴۰\times۱۰} = \frac{۴۴۱}{۴۰۰} = ۰/۲۵$$

در خصوص تفسیر ضریب تمیز باید گفت که هرچه‌قدر ضریب تمیز سؤال بزرگ‌تر و به یک نزدیک‌تر باشد، به این معنا است که سؤال دانشجویان قوی و ضعیف را بهتر از یکدیگر جدا می‌کند. ضریب تمیز معادل صفر به این معنا است که سؤال اصلاً قادر نیست دانشجویان دو گروه را از هم تفکیک کند. ضریب تمیز می‌تواند مقادیر منفی نیز به خود بگیرد که این حالت به این معنا است که دانشجویان گروه پایین، بهتر به سؤال جواب داده‌اند. در آزمون‌های پیشرفت تحصیلی این وضعیت مطلوب نیست و باید به دنبال علت آن بود. یکی از اولین مواردی که در سؤال با ضریب تمیز منفی به ذهن می‌رسد، این احتمال است که کلید سؤال اشتباه اعلام شده باشد. احتمال دیگر، وجود نکته انحرافی یا خطای ساختاری در سؤال است که موجب شده دانشجویان قوی برداشت اشتباهی داشته باشند و غلط پاسخ دهند. همچنین می‌تواند نشانه این موضوع باشد که دانشجویان طی دوره مبحث مورد نظر را به صورت اشتباه یاد گرفته‌اند.

در تفسیر ضریب تمیز همانند تفسیر ضریب دشواری مهم است که هدف آزمون در نظر گرفته شود. در یک آزمون هنجاری، سؤالات باید قادر باشند بین دانشجویان تمایز ایجاد کنند. اگر سؤالات قادر به افتراق دانشجویان رقابت‌کننده نباشند، نمرات شرکت‌کنندگان مشابه و نزدیک هم خواهد شد و انتخاب نفرات برتر را با دشواری مواجه می‌سازد. پس لازم است سؤالاتی با ضریب تمیز بالا یعنی نزدیک به یک داشته باشیم. در حالی که در آزمون‌های معیاری، مورد انتظار است که اکثر فراگیران به اکثر سؤالات به درستی جواب دهند. بنابراین، دور از انتظار نیست که ضریب تمیز سؤالات چندان بالا نباشد.

شاخص‌های ایستگاه OSCE

در مبحث قبلی چگونگی محاسبه ضرایب دشواری و تمیز برای ایستگاه‌های OSCE ذکر شد. علاوه بر این دو شاخص، موارد دیگری وجود دارند که بعد از برگزاری آزمون اطلاعات خوبی در مورد تک‌تک ایستگاه‌ها در اختیار ممتحنان می‌گذارند:

یکی از این شاخص‌ها، درصد قبولی در ایستگاه است که در بحث ضریب دشواری به آن اشاره شد.

شاخص دوم، آلفای کرونباخ در سطح ایستگاه است که بر اساس آیتم‌های چک‌لیست ایستگاه محاسبه می‌شود و ثبات درونی یک ایستگاه را نشان می‌دهد.

شاخص سوم، آلفای کرونباخ در سطح کل آزمون است که هرچند اطلاعات در مورد ثبات درونی کل OSCE به دست می‌دهد، با محاسبه «آلفا در صورت حذف آیتم»[1] می‌توان به اطلاعاتی در خصوص ایستگاه دست یافت. به این معنا که با حذف تک‌تک ایستگاه‌ها مجدداً آلفای کل OSCE محاسبه می‌شود.

شاخص چهارم، توافق بین داوران در برخی از OSCEهایی است که بیش از یک آزمونگر در هر ایستگاه مستقر می‌شوند.

شاخص‌های دوم، سوم و چهارم در فصل پایایی به صورت مفصل بیان خواهند شد. در اینجا ضمن اینکه بار دیگر تاکید می‌کنیم یک شاخص به تنهایی برای تصمیم‌گیری در مورد کیفیت آزمون یا سؤالات آن کافی نیست، به بررسی دو شاخص دیگر یعنی ضریب R^2 و تمایز بین درجات[2] می‌پردازیم که برای تحلیل کیفیت هر ایستگاه به کار می‌روند.

برای هر دوی این متریک‌ها لازم است علاوه بر چک‌لیستی که به صورت معمول در ایستگاه تکمیل می‌شود، نمره‌دهی گلوبال نیز برای ایستگاه ثبت شود و معادله رگرسیونی محاسبه شود که در آن نمره چک‌لیست و گلوبال به ترتیب به عنوان متغیر وابسته و مستقل در نظر گرفته می‌شوند. در خصوص این دو نمره در بخش مربوط به OSCE (بخش پنجم کتاب) مفصلاً توضیح داده شد. یادآوری این نکته ضروری است که برای نمره‌دهی به صورت گلوبال باید دو مسأله را در نظر داشت:

اولاً آزمونگر برای دادن نمره گلوبال باید یک قضاوت کیفی و ذهنی مستقل از چک‌لیست انجام دهد. به عبارت دیگر، نباید نمره چک‌لیست را عیناً به نمره گلوبال ترجمه و تبدیل کند.

دوم این که آزمونگر برای دادن نمره گلوبال باید سطح و مقطع دانشجویان شرکت‌کننده را در نظر بگیرد. عملکردی که یک کارآموز در بخیه زدن از خود نشان می‌دهد، ممکن است برای کارآموزان در سطح «عالی» باشد اما ممکن است برای یک دستیار در سطح «قابل قبول» در نظر گرفته شود. بنابراین مهم است در آزمون OSCE که برای کارآموزان برگزار می‌شود، آزمونگران در این خصوص توجیه شوند.

ضریب R^2

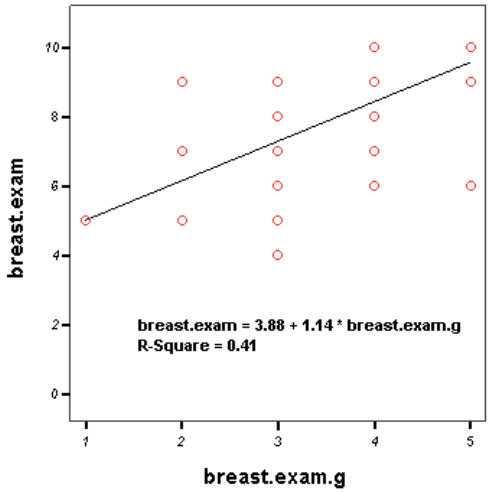
در صورت استفاده از چک‌لیست و نمره گلوبال در ایستگاه‌ها، برای هر دانشجو در هر ایستگاه یک نمره چک‌لیست و یک نمره گلوبال داریم که به ترتیب به عنوان متغیرهای وابسته و مستقل در نظر گرفته می‌شوند. به صورت منطقی در نمرات گلوبال بالا، انتظار داریم نمرات چک‌لیست هم بالا باشند و برعکس.

اگر نمودار پراکندگی نمرات دانشجویان را رسم کنیم، با استفاده از رگرسیون می‌توان خطی را رسم کرد که رابطه بین دو نمره را به بهترین شکل پیش‌بینی می‌کند. ضریب R^2 میزان تغییر در متغیر وابسته را نسبت به تغییراتی که در متغیر مستقل ایجاد شده است، نشان می‌دهد. اگر مقدار ضریب R^2 بالای ۰/۵ باشد، نشان‌دهنده رابطه منطقی بین نمرات گلوبال و چک‌لیست است اما باید دقت شود که اگر آزمونگر حاضر در ایستگاه برای دادن نمره گلوبال، نمره چک‌لیست را ترجمه

1. Alpha if item deleted
2. Intergrade discrimination

کرده باشد، مقدار این ضریب به صورت غیرطبیعی بالا در خواهد آمد. (پل وفولر ۲۰۱۰)

به عنوان مثال، ضریب معادل ۰/۶ مشخص می‌کند که ۶۰ درصد تغییرات در نمرات گلوبال دانشجویان به دلیل تغییرات در نمره چک‌لیست آنها است و به عبارت دیگر با تفاوت‌های موجود در نمرات چک‌لیست توجیه می‌شود که خوب در نظر گرفته می‌شود. اما اگر به عنوان مثال، ضریبی برابر ۰/۴ برای ایستگاهی به دست آید، باید آن را تحلیل کنیم. برای این کار، بهتر است نمودار پراکندگی[1] نمرات و خط رگرسیون ایستگاه را ببینیم و رابطه دو متغیر را به صورت دقیق بررسی کنیم. در اینجا ایستگاه معاینه سینه را به عنوان مثال در نظر می‌گیریم که نمودار توزیع نمرات آن در شکل ۱-۳۲ نشان داده شده است و R^2 آن ۰/۴۱ محاسبه شده است. در این ایستگاه نمره چک‌لیست مقادیری بین صفر تا ۱۰ را به خود اختصاص داده است و برای ارزیابی گلوبال از لیکرت پنج‌تایی استفاده شده است (یک رد، دو مرزی، سه قابل‌قبول، چهار خوب و پنج فراتر از انتظار).



شکل ۱-۳۲: نمودار توزیع نمرات چک‌لیست و گلوبال در ایستگاه معاینه سینه

همان‌طور که مشاهده می‌شود، نمرات چک‌لیست دانشجویان در هر یک از رده‌های ارزیابی گلوبال طیف وسیعی داشته است. به عنوان مثال، دانشجویانی که عملکرد آنها در ارزیابی گلوبال، قابل قبول (۳) برآورد شده است، نمرات چک‌لیست متنوعی داشته‌اند که عمدتاً از ۵۰ تا ۹۰ متغیر بوده است. به عبارت دیگر برخی از دانشجویان علی رغم اینکه نمره چک‌لیست خوبی گرفته‌اند، در ارزیابی گلوبال رضایت‌بخش نبوده‌اند. این مسأله می‌تواند نشان دهنده مشکلی در ارزیاب یا چک‌لیست باشد. در قدم بعدی چک‌لیست ایستگاه را بررسی می‌کنیم (جدول ۶-۳۲)

این‌طور به نظر می‌رسد که از نظر طراح چک‌لیست، هدف اصلی این ایستگاه معاینه بوده است زیرا هشت نمره از ۱۰ نمره را به خود اختصاص داده است؛ در حالی که احتمالاً از نظر آزمونگر، هدف نهایی و اصلی از این ایستگاه، یافتن توده بوده است بنابراین هنگام نمره‌دهی به صورت ذهنی وزن بیشتری به آن داده است. این مسأله باعث شده است تا دانشجویانی که مراحل معاینه را انجام داده‌اند اما قادر به یافتن توده نبوده‌اند، قسمت عمده نمره چک‌لیست را دریافت کنند در حالی که در ارزیابی گلوبال نمره بالایی نگرفته‌اند. اگر قرار است این ایستگاه در آزمون بعدی استفاده شود، این مسأله باید حل شود یا با تغییر چک‌لیست به نحوی که اهمیت یافتن توده در آیتم ها منعکس شود یا با توجیه آزمونگر که هدف اصلی این ایستگاه معاینه است و نه یافتن توده.

1. scatter

جدول ۶-۳۲ : چک‌لیست ایستگاه معاینه سینه (نمره چک‌لیست از ۱۰)

ردیف	آیتم	نمره	نمره دانشجو
۱	استفاده از سه انگشت دوم، سوم و چهارم در حالت نیمه خم شده (semi-flexed)	۱	
۲	انجام معاینه به شکل نظام‌مند به یکی از اشکال زیر: • vertical strip pattern • circular • wedge	۲	
۳	معاینه در هر نقطه از مسیر به صورت small concentric circles	۱	
۴	معاینه در هر نقطه به صورت سطحی و عمقی	۱	
۵	معاینه با دو دست	۱	
۶	معاینه هر ۴ ربع پستان	۱	
۷	لمس نوک پستان و فشار دادن از نظر وجود ترشح	۱	
۸	نتیجه معاینه ۱ (وجود توده)	۱	
۹	بیان خصوصیات توده (حاشیه توده، چسبندگی، سایز، قوام)	۱	

شاخص تمایز بین درجات

این شاخص معادل شیب خط رگرسیون است و نشان می‌دهد که به طور متوسط افزایش چند واحد در نمره چک‌لیست دانشجویان منجر به ارتقا یک واحد در مقیاس گلوبال می‌گردد. هرچند که مقدار ایده‌آلی برای این شاخص وجود ندارد، یک‌دهم نمره کل چک‌لیست به عنوان حد مطلوب این شاخص ذکر شده است (پل و فولر ۲۰۱۰). به عنوان مثال اگر نمره کامل چک‌لیست ۱۰ باشد، مقدار مطلوب این شاخص یک است.

مقدار پایین این شاخص معمولاً با وجود مشکل در شاخص‌های دیگر همراه است؛ مانند پایین بودن R^2 که رابطه ضعیف بین نمرات چک‌لیست و گلوبال را نشان می‌دهد یا واریانس بالای بین آزمونگران که نشان‌دهنده عدم توافق بین آنها است. مقدار بالای این شاخص می‌تواند با حدنصاب قبولی پایین همراه باشد.

فصل ۳۳

روایی

تعریف روایی و انواع آن در بخش اول کتاب مورد بحث قرار گرفت. همچنین در بخش‌های مختلف کتاب دیدیم که چه نکاتی هنگام استفاده از ابزارهای مختلف ارزیابی باید مورد توجه قرار گیرد که روایی آزمونی که برگزار می‌کنیم، تضمین شود. در اینجا به معرفی راه‌هایی می‌پردازیم که بعد از برگزاری آزمون در نظر می‌گیریم تا از چگونگی و میزان روایی آن اطلاع و اطمینان پیدا کنیم.

روایی صوری

در آزمون‌ها به صورت معمول بر روایی صوری و بررسی آن تاکید چندانی نمی‌شود. در علوم پزشکی بیشتر کاربردی بودن آزمون و مطابقت آن با شرایط کار واقعی فراگیر در این رده قرار می‌گیرد که روش سنجش آن شبیه روایی محتوایی است.

روایی محتوایی

طبق تعریف، هدف از بررسی روایی محتوایی آزمون این است که ببینیم آزمون تا چه میزان با اهداف دوره در یک راستا بوده است. به این منظور، سؤالات آزمون توسط متخصصان مورد بررسی قرار می‌گیرند تا میزان تطابق آنها را با اهداف دوره مشخص نمایند. با توجه به اینکه هدف از بررسی روایی محتوایی آزمون، میزان تطابق آن با اهداف دوره است، آزمون در اختیار متخصصان قرار می‌گیرد و از آنها درخواست می‌شود تا با خواندن تک‌تک سؤالات قضاوت کنند که سؤال تا چه اندازه هدف دوره را برآورده می‌کند. البته برای پاسخ دادن به چنین سؤالی مهم است که اهداف دوره قبلاً به صورت دقیق مشخص شده باشد، بلوپرینت آزمون تدوین شده باشد و متخصصان نیز از آن مطلع باشند.

در یک حالت ساده، هر داور برای هر سؤال نظر خود را به صورت زیر در قالب یک لیکرت پنج‌تایی (کاملاً مخالفم تا کاملاً موافقم) اعلام می‌کند: «به نظر من این سؤال به خوبی توانسته است اهداف دوره را مورد سنجش قرار دهد». سپس نظر افراد مختلف در مورد یک سؤال جمع‌آوری شده و میانگین آن محاسبه می‌شود. به این ترتیب مشخص می‌شود که آیا سؤال واجد روایی بوده است یا باید به طراح بازخورد داده شود که در آن تجدید نظر کند.

نکته حائز اهمیت این است که بررسی روایی سؤال تنها مختص آزمون‌های چندگزینه‌ای نیست. به عنوان مثال، در OSCE می‌توان از متخصصان خواست با بررسی محتوای ایستگاه‌ها شامل سناریو و چک‌لیست مشخص کنند که آیا ایستگاه توانسته به خوبی اهداف دوره را مورد سنجش قرار دهد یا خیر. همچنین این نکته در مورد آزمون‌های مبتنی بر محیط کار واقعی نیز صادق است. همان‌طور که اولین مرحله در طراحی لاگ‌بوک و DOPS و mini-CEX و ...، تدوین بلوپرینت و مشخص کردن

حیطه‌ها یا پروسیجرهای مورد نظر بود، در بررسی روایی نیز میزان مطابقت آنچه سنجیده شده با بلوپرینت باید مدنظر قرار گیرد.

حالت دیگری برای اخذ نظر داوران وجود دارد که به روش لاشه[1] معروف است و به این صورت است که هر داور میزان ارزش و ضرورت هر سؤال را در قالب پاسخ‌های «ضروری»، «مفید» یا «غیرضروری» مشخص می‌کند. «ضریب روایی محتوایی»[2] هر سؤال از فرمول زیر به دست می‌آید:

$$\text{ضریب روایی محتوایی} = \frac{\text{تعداد داورانی که سوال را ضروری تشخیص داده‌اند}}{\text{نصف تعداد داوران}} - ۱$$

این ضریب می‌تواند مقادیر منفی یک تا مثبت یک را به خود اختصاص دهد. اگر عدد به دست آمده بالای صفر، یعنی مثبت، باشد به این معنا است که نیمی از داوران سؤال را ضروری تشخیص داده‌اند و اگر منفی باشد، یعنی بیشتر داوران سؤال را غیرضروری دانسته‌اند. در نهایت، ضریب‌های مربوط به سؤالات مختلف با یکدیگر میانگین گرفته می‌شود تا «شاخص روایی محتوایی»[3] که مربوط به کل آزمون است، به دست آید.

علاوه بر ضرورت سؤال که در روش لاشه مورد تاکید قرار می‌گیرد و شاید بیشتر مناسب آزمون‌های معیاری باشد، پژوهشگران دیگر برای محاسبه روایی محتوایی موارد دیگری را نیز مدنظر قرار داده‌اند. به عنوان مثال هر داور باید هر سؤال را از چهار منظر «ارتباط»[4]، «شفافیت»[5]، «سادگی»[6] و «ابهام»[7] بررسی کند و در هریک از این جنبه‌ها نمره‌ای از یک تا چهار به سؤال اختصاص دهد.

با توجه قضاوتی بودن ماهیت روایی محتوایی آزمون مهم است که بدانیم از نظر چه کسانی و با چه ویژگی‌هایی استفاده کنیم. معمول است که برای تعیین روایی محتوایی از نظر افراد خبره و متخصصان استفاده شود. یعنی کسانی که بر محتوای دوره و موضوع آزمون تسلط دارند و همچنین با سطح و ویژگی‌های شرکت‌کنندگان آشنا هستند. در عین حال می‌توان از نظر خود شرکت‌کنندگان در آزمون نیز در مورد روایی محتوایی آزمون استفاده کرد (جدول ۱-۳۳ و ۲-۳۳).

مسأله دیگر تعداد کسانی است که در نظرخواهی شرکت می‌کنند. هرچند می‌توان برای یک آزمون کلاسی از نظر یک نفر که احتمالاً همان مدرس دوره بوده است، برای این منظور استفاده کرد، در آزمون‌های مهم و سطح بالا باید نظر تعداد بیشتری از متخصصان لحاظ شود.

در هرحال، توجه به این مسأله حائز اهمیت است که مبنای بررسی روایی محتوایی آزمون، قضاوت است و روش آماری برای به دست آوردن میزان واقعی آن وجود ندارد. هرچند روش‌های ریاضی برای کمّی کردن و نشان دادن مفهوم آن در قالب توافق بین متخصصان مورد استفاده قرار می‌گیرند، آنچه در مورد روایی محتوایی یک آزمون گفته می‌شود، نهایی و قطعی و قابل اثبات نیست. همچنین از آنجا که قضاوت افراد مختلف با یکدیگر فرق می‌کند، ممکن است گروه‌های مختلف در مورد روایی محتوایی یک آزمون نظرات گوناگونی ابراز کنند.

همچنین توجه به این نکته ضروری است که تمام روش‌های فوق تنها یک آیتم (سؤال، ایستگاه یا پروسیجر) را مبنای قضاوت قرار می‌دهند. در حالی که احتمال دارد آیتم‌ها مبتنی بر اهداف دوره باشند اما نمونه‌گیری از محتوای دوره به خوبی صورت نگرفته نباشد. به عبارت دیگر آزمون پوشش و توزیع منطقی از اهداف دوره نداشته باشد. به عنوان مثال، در آزمون کتبی اکثر سؤالات مربوط به دو فصل ابتدای کتاب باشند یا در OSCE جامع پیش‌کارورزی اکثر ایستگاه‌ها مربوط به

1. Lawshe
2. Content Validity Ratio (CVR)
3. Content Validity Index (CVI)
4. Relevance
5. Clarity
6. Simplicity
7. Ambiguity

رشته جراحی باشند یا در DOPS از بین تمام پروسیجرهای مربوط به دوره تنها یک پروسیجر برای ارزیابی انتخاب شده باشد. بنابراین، به منظور بررسی روایی محتوایی نمی‌توان صرفاً به بررسی تک‌تک سؤالات و آیتم‌ها بسنده کرده بلکه مطابقت کل آزمون با بلوپرینت دوره ضروری است. به عبارت دیگر، از متخصصان خواسته می‌شود در مورد این مسأله قضاوت کنند که طراحی آزمون تا چه حد بر اساس بلوپرینت و اهداف دوره بوده است.

جدول ۱-۳۳ : چند سؤال مرتبط با روایی محتوایی آزمون در نظرخواهی از اعضای هیات علمی

آزمون	گویه
آزمون کتبی و شفاهی	سؤالات زیادی در مورد یک موضوع خاص پرسیده شد (در مورد کل آزمون). مواردی وجود داشت که پرسیده نشد (در مورد کل آزمون). این سؤال مطرح‌کننده یک هدف مهم دوره است. این سؤال همان چیزی را می‌سنجد که در بلوپرینت دوره مقرر بود مورد ارزیابی قرار گیرد. سطح این سؤال با سطح شرکت‌کنندگان مطابقت دارد. این سؤال قائل به افتراق بین فراگیران خوب و بد است.
OSCE	موارد فوق این ایستگاه شبیه شرایط واقعی کار فراگیران است. اگر فراگیری نمره حدنصاب این ایستگاه را کسب کند، در مواجه با بیمار نیز توانمندی مورد نظر را داراست.
DOPS و mini-CEX	تعریف سطوح عملکرد مورد انتظار واضح بود. تعریف سطوح عملکرد مورد انتظار واقعی بود. این آزمون سطوح توانمندی در پروسیجر ... را اندازه می‌گیرد. این آزمون در تشخیص نقاط ضعف توانمندی ... فراگیران مفید است. استفاده از این آزمون می‌تواند توانمندی .... فراگیران را بهبود بخشد.

جدول ۲-۳۳ : چند سؤال مرتبط با روایی محتوایی آزمون در نظرخواهی از فراگیران

آزمون	گویه
آزمون کتبی و شفاهی	سؤالات زیادی در مورد یک حیطه خاص پرسیده شد (در مورد کل آزمون). مواردی وجود داشت که علی‌رغم انتظار من پرسیده نشد (در مورد کل آزمون). محتوای سؤال در اهداف دوره وجود داشت. محتوای سؤال آموزش داده شده بود. محتوای سؤال با کار روزمره من ارتباط داشت.
OSCE	موارد فوق وظیفه خواسته شده در این زمان قابل اجرا بود. عملکرد بیمارنما قابل باور بود.
DOPS و mini-CEX	تعریف سطوح عملکرد مورد انتظار واضح بود. حیطه‌های مورد ارزیابی واقعی به درستی انتخاب شده بودند. این آزمون در تشخیص نقاط ضعف من کمک‌کننده بود.

روایی معیاری

روایی معیاری به صورت معمول به صورت روایی پیش‌بینی و روایی همزمان مطرح می‌شود. هرچند که برخی از لحاظ مفهومی بین این دو تمایز قائل می‌شوند، روش بررسی این دو نوع روایی تفاوت چندانی ندارد. به همین دلیل آنچه در اینجا توضیح می‌دهیم کلاً برای بررسی روایی معیاری به کار می‌رود.

برای تعیین روایی معیاری وجود دو آزمون ضروری است. هدف این است که دریابیم نتایج دو آزمون تا چه اندازه در یک راستا هستند. منظور از نتایج آزمون می‌تواند رتبه دانشجویان باشد یا نتیجه‌ای که از لحاظ رد و قبول کسب کرده‌اند:

هنگامی که رتبه دانشجویان در دو آزمون مدنظر قرار می‌گیرد، به این معنا است که نمرات خام هر یک از دانشجویان در هر یک از آزمون‌ها باید وجود داشته باشد. در این حالت، ضریب همبستگی بین نمرات حاصل از دو آزمون محاسبه می‌شود. از آنجا که معمولاً نمرات به شکل یک متغیر پیوسته گزارش می‌شوند، از ضریب همبستگی پیرسون استفاده می‌شود که در فصل اول این بخش مورد بحث قرار گرفت. اگر ضریب به دست آمده مثبت باشد، به این معنا است که دو آزمون در راستای اهداف مشترکی حرکت می‌کنند. هرچقدر عدد به دست آمده به مثبت یک نزدیک‌تر باشد، یعنی روایی بیشتری دیده می‌شود. مثال‌های متعددی از سنجش این نوع روایی در دسترس است. به عنوان مثال، بررسی همبستگی نمرات کنکور سراسری با آزمون جامع علوم پایه، همبستگی نمرات آزمون جامع علوم پایه با آزمون جامع پیش‌کارورزی، همبستگی نمرات آزمون کتبی پیش‌کارورزی با آزمون OSCE پیش کارورزی. باید توجه داشت که ضریب همبستگی پیرسون، میزان یکسان بودن رتبه افراد را در دو آزمون بررسی می‌کند. به عنوان مثال، اگر همبستگی آزمون جامع علوم پایه با آزمون جامع پیش‌کارورزی بالا باشد، به این معنا است که اگر دانشجویی در آزمون جامع علوم پایه رتبه‌ای را کسب کند، به احتمال زیاد در آزمون جامع پیش‌کارورزی هم همان حدود رتبه را تکرار خواهد کرد.

با توجه به مسأله فوق، این نکته حائز اهمیت است که ضریب همبستگی پیرسون بیشتر برای آزمون‌های هنجاری مناسب است که با رتبه فراگیران سر و کار دارند، نه آزمون‌های معیاری که باید در خصوص وضعیت رد و قبول دانشجویان تصمیم‌گیری کنند. هنگامی که تصمیم رد/قبول در آزمون‌ها اهمیت دارد، نمرات خام دانشجویان به دو حالت قبول (یک) و رد (صفر) تبدیل می‌شود. به عنوان مثال می‌خواهیم ببینم دانشجویانی که در آزمون جامع علوم پایه قبول می‌شوند، از لحاظ وضعیت قبولی در آزمون جامع پیش‌کارورزی چگونه هستند. روایی معیاری که در اینجا مورد بررسی قرار می‌گیرد، روایی حدتسلط نیز نامیده می‌شود و از طریق محاسبه ضریب همبستگی بین دو آزمون تعیین می‌شود. در اینجا نمی‌توان از ضریب همبستگی پیرسون استفاده کرد چون دو متغیر در واقع پیوسته بوده‌اند و ما آنها را به متغیر دوحالته (رد/قبول) تبدیل کرده‌ایم. از آنجا که ضریب همبستگی پیرسون برای متغیرهای پیوسته مناسب است، برای تعیین روایی معیاری حدتسلط به روش‌های دیگری نیاز است. این همبستگی می‌تواند از طریق ضریب همبستگی چهارخانه‌ای[1] محاسبه شود که توضیح آن از حوصله این بحث خارج است. یک روش ساده که در اینجا با ذکر یک مثال توضیح داده می‌شود، بررسی میزان توافق بین نتایج دو آزمون است. البته برای بررسی میزان توافق، روش‌های بهتری از جمله ضریب کاپا و «ضریب همبستگی درون گروهی»[2] وجود دارد که مخصوصاً برای محاسبه «پایایی بین آزمونگران»[3] مورد استفاده قرار می‌گیرند و در فصل پایایی ارائه خواهند شد.

فرض کنید نمرات دو آزمون جامع یک ورودی از دانشجویان را که شامل ۱۱۰ دانشجو هستند، بررسی کرده‌ایم و میزان رد و قبول آنها را استخراج کرده‌ایم که مطابق جدول ۳-۳۳ به دست آمده است (بدیهی است دانشجویی که در امتحان جامع علوم پایه قبول نشود، نمی‌تواند مراحل بعدی را ادامه دهد و در امتحان جامع پیش‌کاروزی شرکت نماید.

1. Tetrachoric
2. Intra-class correlation (ICC)
3. Inter-rater reliability (IRR)

در این مثال همه دانشجویان در نوبت دوم امتحان علوم پایه نهایتاً قبول شده‌اند اما ما روایی معیاری آزمون اول ایشان را محاسبه می‌کنیم).

تعداد کسانی که در دو آزمون یک نتیجه کسب کرده‌اند، یعنی در هر دو قبول یا در هر دو رد شده‌اند، محاسبه می‌شود. در مثال فوق، ۶ نفر در هر دو آزمون رد و ۱۰۰ نفر در هر دو آزمون قبول شده‌اند. به عبارت دیگر ۱۰۶ نفر از ۱۱۰ نفر نتیجه یکسان کسب کرده‌اند. یعنی نتایج دو آزمون در ۹۶/۳ موارد یکسان بوده است که عدد قابل توجه و بالایی است و نشان می‌دهد دانشجویی که بار اول در امتحان علوم پایه رد می‌شود به احتمال زیاد در امتحان پیش‌کارورزی هم رد می‌شود و دانشجویی که در امتحان علوم پایه قبول می‌شود به احتمال زیاد در امتحان پیش‌کارورزی نیز قبول می‌شود.

جدول ۳-۳۳: میزان ردی و قبولی ۱۱۰ دانشجو در آزمون‌های جامع علوم پایه و پیش‌کارورزی

	قبول در آزمون جامع پیش‌کارورزی	رد در آزمون جامع پیش‌کارورزی
قبول در آزمون جامع علوم پایه	۱۰۰ نفر	۳ نفر
رد در آزمون جامع علوم پایه	۱ نفر	۶ نفر

روایی سازه

مفهوم روایی سازه شاید بیشتر برای پرسشنامه‌ها کاربرد داشته باشد و در بحث آزمون‌ها و ارزیابی فراگیران چندان مورد توجه قرارنگیرد. هر چند که روش‌های تعیین روایی سازه در این دو حوزه با یکدیگر تفاوت چندانی ندارند، در اینجا تأکید اصلی را بر روش‌هایی می‌گذاریم که بیشتر برای تعیین روایی سازه آزمون‌ها به کار گرفته می‌شوند.

ضریب همبستگی: یکی از روش‌های تعیین روایی سازه مشابه روش تعیین روایی معیاری است. یعنی ضریب همبستگی آزمون با یک آزمون دیگر محاسبه می‌شود. هرچه عدد به دست آمده بزرگ‌تر باشد، می‌توان نتیجه گرفت که دو آزمون سازه‌های مشابهی را مورد ارزیابی قرار داده‌اند. برعکس اگر ضریب همبستگی کوچک باشد، سازه‌های مورد سنجش متفاوت هستند. به عنوان مثال، پیشتر ذکر شد که می‌توان ضریب همبستگی آزمون کتبی و OSCE پیش کارورزی را برای بررسی روایی معیاری (همزمان) تعیین کرد. اما هنگامی که عدد حاصله چندان بزرگ نباشد، به عنوان مثال حدود ۰/۴ به دست آید، یکی از علل در توضیح این پدیده می‌تواند اشاره به این امر باشد که اصولاً سازه‌های مورد ارزیابی در دو آزمون با یکدیگر متفاوت هستند. یکی به بررسی دانش نظری می‌پردازد و دیگری مهارت‌های عملی دانشجویان را ارزیابی می‌کند.

تمایز سال/رده: هرچند که در نگاه اول به نظر می‌رسد این روش بیشتر برای آزمون‌های دبستانی کاربرد دارد اما در واقع یکی از روش‌های متداول برای بررسی روایی آزمون‌هایی مانند DOPS و mini-CEX است. اگر یک گروه آموزشی تصمیم بگیرد برای دستیاران در سال‌های مختلف از یک پروسیجر خاص به صورت DOPS ارزیابی کند، می‌توان نمرات دستیاران در سال‌های مختلف را با هم مقایسه کرد. اگر این طور باشد که دستیاران سال بالاتر نمرات بهتری کسب کرده باشند، یعنی این توانمندی در طول دوره پیشرفت نشان داده است و ارزیابی واجد روایی سازه است.

همسانی درونی: همسانی درونی[1] که در قسمت مربوط به پایایی مفصلاً توضیح داده می‌شود، به این معنا است که کل یک آزمون تا چه اندازه با قسمت‌های مختلف خود یا سؤالات مختلف خود در یک راستا است. یعنی همبستگی اجزای تشکیل‌دهنده آزمون با کل آزمون محاسبه می‌شود. به عنوان مثال آزمون جامع پیش‌کارورزی را در نظر بگیرید که سؤالات آن از رشته‌های گوناگون جراحی، داخلی، زنان و ... طرح شده‌اند. همسانی درونی تعیین می‌کند که آیا نمره

1. Internal consistency

فراگیر در قسمت جراحی با سایر قسمت‌های آزمون یا با کل آزمون در یک راستا است یا خیر. فرض بر این است که اجزا باید همبستگی خوبی با کل آزمون داشته باشند زیرا روی هم قرار است مجموعه‌ای از اهداف مشخص را ارزیابی کنند. در عین حال میزان همبستگی اگر خیلی زیاد باشد، نشانه چندان خوبی نیست زیرا این احتمال را مطرح می‌کند که اجزا در حال ارزیابی موارد یکسان هستند. مثلاً آزمون OSCE را در نظر بگیرید که همسانی درونی ایستگاه‌های آن با یکدیگر و با کل آزمون خیلی بالاست. با مراجعه به متن ایستگاه‌ها در می‌یابیم که تمام ایستگاه‌ها از حیطه شرح حال انتخاب شده‌اند و مهارت‌هایی مانند معاینه، تفسیر، ارائه برنامه درمانی و پروسیجر که در بلوپرینت بوده‌اند، در نظر گرفته نشده‌اند. منطقی است که یک دانشجو در شرح حال گرفتن از بیماران مختلف، مهارت کم و بیش مشابه‌ای از خود بروز می‌دهد؛ در حالی که وقتی در یک ایستگاه باید معاینه شکم انجام دهد و در ایستگاه دیگر باید برای بیمار با ضایعه پوستی نسخه بنویسد، عملکرد متنوعتری از خود نشان خواهد داد.

ضریب آلفای کرونباخ: یکی از روش‌های متداول برای سنجش همسانی درونی، ضریب آلفای کرونباخ است که بیشتر به عنوان روش سنجش پایایی مشهور است و در فصل پایایی توضیح داده خواهد شد.

تحلیل عاملی: تحلیل عاملی[1] یکی از متداول‌ترین روش‌ها برای تعیین روایی سازه پرسشنامه‌ها است اما شاید برای تعیین روایی سازه آزمون‌هایی که ما با آنها سر و کار داریم چندان پرکاربرد نباشد. با تحلیل عاملی که از طریق نرم‌افزارهای آماری صورت می‌گیرد، می‌توان مشخص کرد که آزمون از چه سازه‌هایی تشکیل شده است. در واقع سؤالات آزمون دسته‌بندی می‌شوند و مواردی که ماهیت مشابه دارند، با یکدیگر یک عامل را تشکیل می‌دهند که در حقیقت یک سازه مشخص را ارزیابی می‌کند.

1. Factor analysis

فصل ۳۴

پایایی

پایایی در نظریه کلاسیک آزمون

تعریف پایایی و انواع آن در بخش اول کتاب مورد بحث قرار گرفت. همچنین در بخش‌های مختلف کتاب بحث شد که هنگام استفاده از ابزارهای مختلف ارزیابی چه نکاتی باید رعایت شود تا آزمونی که برگزار می‌کنیم، پایایی بهتری داشته باشد. در اینجا به معرفی روش‌های تعیین پایایی می‌پردازیم تا بعد از برگزاری آزمون از میزان پایایی آزمون اطلاع پیدا کنیم. اما قبل از شروع بحث، چند نکته را در مورد مفهوم پایایی تاکید می‌کنیم که اگرچه که قبلاً ذکر شده‌اند اما به دلیل اهمیت، تکرار آنها ضروری است:

- اول اینکه در گذشته نه چندان دور، اغلب مفهوم پایایی و عینی بودن آزمون معادل یکدیگر به کار می‌رفتند و در بسیاری از موارد عینی بودن آزمون، معیاری برای پایایی بالا محسوب می‌شد. در حالی که تحقیقات اخیر این موضوع را تایید نمی‌کنند. به عنوان مثال، در مورد OSCE این نظر وجود داشت که حسن عمده آن نسبت به سایر ابزارهای سنجش مهارت‌های بالینی، در چک‌لیست‌های عینی آن است که تصور می‌شد پایایی ابزار را افزایش می‌دهند. در حالی که بر اساس نتایج مطالعات مشخص شده است که پایایی خوب OSCE متاثر از عوامل دیگری است؛ مانند تعداد بیشتر ایستگاه که منجر به نمونه‌گیری خوب از محتوای بالینی مورد نظر می‌شود. در واقع عینی بودن، ساختارمند بودن و چک‌لیست‌های استاندارد اثر کمتری روی پایایی OSCE دارند.
- دوم اینکه در متون برای تعریف پایایی به میزان ثبات نتایج و میزان یکسان بودن نمرات در تکرار دفعات اندازه‌گیری تاکید شده است. در حالی که شاید تعریف جامع‌تر پایایی، میزان تاثیر خطای تصادفی[1] در نتایج مشاهده شده باشد. هر چند که همیشه تلاش می‌شود خطای اندازه‌گیری کاهش پیدا کند، هر روش اندازه‌گیری در ذات خود واجد خطا است و نتیجه‌ای که به دست می‌دهد همیشه با درجاتی از خطا همراه است. به این ترتیب هیچ وقت نمی‌توانیم بگوییم نمره‌ای که دانشجو در آزمون کسب کرده است، دقیقاً معادل نمره واقعی او است اما با تعیین پایایی آزمون می‌توانیم بگوییم میزان تاثیر خطای تصادفی در نمره او چقدر بوده است.

$$\text{پایایی} = \frac{\text{واریانس نمرات واقعی}}{\text{واریانس نمرات مشاهده شده}}$$

- منابع خطا در هر آزمون متعدد هستند و از خود امتحان، دانشجو و ممتحن ناشی می‌شوند. در نظریه کلاسیک در هر لحظه تنها اثر یکی از این منابع قابل بررسی است. نحوه تاثیر همه منابع با هم در قسمت بعدی، یعنی پایایی از دیدگاه نظریه تعمیم‌پذیری، بحث خواهد شد.

اکنون به معرفی روش‌های تعیین پایایی می‌پردازیم. روش‌های تعیین پایایی متنوع هستند و هرچند اصول کلی آنها در حوزه

1. Random

ارزیابی فراگیر همانند سایر حیطه‌های اندازه‌گیری است، اما به علت شرایط خاص حاکم بر آزمون‌ها از برخی از روش‌ها بیشتر استفاده می‌شود که در اینجا آنها را بیشتر مورد تاکید قرار می‌دهیم. به صورت مشخص‌تر دو روش بازآزمایی و آزمون‌های هم ارز به دلایلی که ذکر خواهد شد، کاربرد کمتری در حوزه آموزش دارند و روش‌های دو نیمه کردن، کودر–ریچاردسون و آلفای کرونباخ که تحت عنوان کلی همسانی درونی شناخته می‌شوند، برای سنجش پایایی ابزارهای ارزیابی فراگیر بیشتر استفاده می‌شوند.

بازآزمایی

از آنجا که یکی از تعاریف رایج پایایی، میزان یکسان بودن نتایج در تکرار اندازه‌گیری است، استفاده از روش بازآزمایی[1] برای سنجش پایایی بسیار معروف و متداول است. در این روش، آزمون (یا پرسشنامه) به عده‌ای از شرکت‌کنندگان ارائه می‌شود تا به آن پاسخ دهند. پس از گذشت دو هفته تا یک ماه مجدداً از همین افراد خواسته می‌شود تا آن را تکمیل کنند. نتایج حاصل از دو نوبت محاسبه می‌شود و ضریب همبستگی بین آنها ضریب پایایی ابزار است.

روش بازآزمایی به صورت کلی مشکلاتی دارد و در حیطه ارزیابی فراگیر به این مشکلات اضافه هم می‌شود که موجب می‌گردد استفاده از آن برای تعیین پایایی امتحانات چندان رایج نباشد و بیشتر برای سنجش پایایی پرسشنامه‌ها و آزمون‌هایی مانند هوش و شخصیت کاربرد داشته باشد: اولاً اینکه اساساً تکرار آزمون برای فراگیران و مدرسان مسأله‌ای غیراجرایی است. هر آزمون در یک زمان مشخص برگزار می‌شود و معمولاً دلیلی ندارد که همان آزمون برای همان دانشجویان با فاصله زمانی تکرار شود. حتی اگر برای سنجش پایایی، آزمون تکرار شود، مشکلاتی وجود دارد:

- شرایط برگزاری دو آزمون لزوماً یکسان نیست. عواملی مانند دما و نور محیط، خستگی، اضطراب، انگیزه و علاقه فراگیران بر نمرات آنها تاثیر می‌گذارند و باعث تفاوت نتایج دو نوبت و مخدوش شدن پایایی می‌گردند.
- سؤالات در دو نوبت یکسان هستند. بنابراین نمی‌توان نتایج کسب شده را به نمونه دیگری از سؤالاتی که ممکن بود ارائه شوند، تعمیم داد. واضح است که در یک امتحان تعداد سؤالاتی که می‌توان طرح کرد، محدود است و چون همیشه تعدادی از سؤالات ممکن انتخاب می‌شوند، همیشه این مسأله وجود دارد که آیا اگر نمونه دیگری از سؤالات انتخاب شده بودند، باز هم نتایج همین بود یا خیر.
- سؤالات در دو نوبت یکسان هستند. بنابراین، شرکت‌کنندگان بعد از نوبت اول با سؤالات آشنا می‌شوند که ممکن است جواب را یاد بگیرند و در نوبت دوم نمرات بهتری کسب کنند. به این ترتیب عددی که برای پایایی گزارش خواهد شد، مخدوش است.
- مطالب فراگرفته شده در طی زمان ممکن است در ذهن فراگیر کمرنگ شوند و موجب شود در نوبت دوم نمره کمتری کسب کند. این مسأله در مورد ویژگی‌های ثابت‌تر مانند هوش و شخصیت چندان نگرانی ایجاد نمی‌کند اما در خصوص آزمون‌های پیشرفت تحصیلی مشکل جدی دارد.

آزمون‌های هم‌ارز یا موازی

در روش آزمون‌های هم‌ارز[2] یا موازی[3]، دو آزمون با سؤالات غیریکسان اما مشابه طراحی می‌شوند و از یک عده افراد یکسان درخواست می‌شود که هر دو را تکمیل کنند. دو آزمون معمولاً در صورتی هم‌ارز اطلاق می‌شوند که پس از تحلیل نمرات، میانگین و واریانس آنها برابر باشد. ضریب همبستگی بین نتایج دو آزمون به عنوان ضریب پایایی آنها محسوب می‌شود.

همان‌طور که مشخص است مشکل فاصله زمانی که در روش بازآزمایی وجود داشت و منجر به کاهش قابلیت اجرا و کمرنگ شدن مطالب فراگرفته در ذهن فراگیران می‌شد، در اینجا مسأله‌ساز نیست اما همچنان مشکلاتی مانند آشنایی با سؤالات و یکسان نبودن شرایط برگزاری دو آزمون وجود دارد. در کنار اینها باید در نظر داشت که طراحی دو آزمون مشابه برای مدرسان آسان نیست.

1. Test-retest
2. Equivalent
3. Parallel

کودر-ریچاردسون

روش کودر-ریچاردسون[1] مبتنی بر یک نوبت اجرای آزمون است و بر خلاف روش بازآزمایی و آزمون‌های هم‌ارز که بر یکسان بودن نتایج در تکرار اندازه‌گیری تاکید داشتند، بیشتر بر همسانی درونی اجزای آزمون تاکید می‌کند.

همسانی درونی در واقع مشخص می‌کند که اجزای یک آزمون تا چه حد هم‌راستا با کل توانمندی مورد نظر هستند. به صورت ساده، اجزای یک آزمون همان سؤالات آزمون هستند. همچنین می‌توان آزمون‌هایی را در نظر گرفت که از چند بخش مختلف تشکیل شده‌اند و هر بخش شامل چندین سؤال است. همسانی درونی، دقیقاً معادل پایایی نیست اما معمولاً به عنوان شاخصی از پایایی آزمون در نظر گرفته می‌شود. ضریب همسانی درونی می‌تواند مقادیر بین ۰ تا ۱ رابه خود اختصاص دهد که هر چه بیشتر باشد، به معنای این است که همسانی درونی آزمون بیشتر است. معمولاً برای آزمون‌های مهم و سطح بالا، ضریب باید بالای ۰/۸ باشد تا قابل قبول در نظر گرفته شد. اما باید توجه داشت که اگر ضریب خیلی بالا بود، چندان نیز مایه خوشحالی نیست. زیرا می‌تواند ناشی از این موضوع باشد که سؤالات شبیه یکدیگر هستند و موضوعات مشابه و تکراری را مورد سنجش قرار می‌دهند.

روش کودر-ریچاردسون صرفاً برای تعیین پایایی آزمون‌هایی کاربرد دارد که پاسخ‌های دوحالته (صفر و یک) دارند؛ مانند سؤال چندگزینه‌ای و نمی‌توان از آن برای سؤالات تشریحی یا در مقیاس لیکرت استفاده نمود.

فرمول کودر-ریچاردسون به دو صورت K20 و K21 نوشته می‌شود. ابتدا فرمول K20 را بررسی می‌کنیم:

$$\text{پایایی} = \frac{\text{تعداد سوال}}{\text{تعداد سوال} - ۱} \left(۱ - \frac{\text{مجموع حاصل ضرب ضریب دشواری هر سوال در یک منهای آن برای همه سوالات}}{\text{واریانس آزمون}} \right)$$

به مثال زیر توجه کنید. یک آزمون چندگزینه‌ای با ۱۰ سؤال برگزار شده است که واریانس آن چهار و ضریب دشواری سؤالات آن به صورت جدول ۱-۳۴ است:

جدول ۱-۳٤ : ضرایب دشواری مربوط به ۱۰ سؤال

شماره سؤال	ضریب دشواری	یک منهای ضریب دشواری
۱	۰/۳	۰/۷
۲	۰/۷	۰/۳
۳	۱	۰
۴	۰/۸۵	۰/۱۵
۵	۰/۷	۰/۳
۶	۰/۸۵	۰/۱۵
۷	۰/۷۵	۰/۲۵
۸	۰/۵	۰/۵
۹	۰/۵	۰/۵
۱۰	۰/۱۵	۰/۸۵

1. Kuder-Richardson

برای محاسبه ضریب پایایی، مقادیر زیر را محاسبه می‌کنیم و در فرمول جایگزین می‌کنیم:

$$\text{پایایی} = \frac{۱۰}{۹}\left[۱-\frac{۰/۲۱+۰/۲۱+۰+۰/۱۲+۰/۲۱+۰/۱۲+۱۰/۲۸+۰/۲۵+۰/۲۵+۰/۱۲}{۴}\right] = ۱/۱ \times (۱-۰/۴۱) = ۰/۶۵$$

اگر فرض شود که ضریب دشواری همه سؤالات با هم برابر هستند، فرمول به نام KR21 شناخته می‌شود و به صورت زیر در می‌آید:

$$\text{پایایی} = \frac{\text{تعداد سوال}}{\text{تعداد سوال} - ۱}\left(۱-\frac{\text{میانگین دشواری سوالات} \times \text{تعداد سوال} \times (\text{یک منهای میانگین دشواری سوالات})}{\text{واریانس آزمون}}\right)$$

در فرمول فوق، به جای محاسبه ضریب دشواری تک تک سؤالات، از میانگین دشواری سؤالات استفاده می‌شود که به این صورت محاسبه می‌شود: میانگین نمرات دانشجویان تقسیم بر نمره کل.

به مثال زیر توجه کنید. در یک آزمون چندگزینه‌ای ۲۰ سؤالی، میانگین کلاس ۱۴ و انحراف معیار نمرات سه است. برای به دست آوردن پایایی آزمون به طریق زیر عمل می‌کنیم:

$$\text{میانگین دشواری سوالات} = \frac{۱۴}{۲۰} = ۰/۷$$

$$\text{پایایی} = \frac{۱۰۰}{۹۹}\left(۱-\frac{۲۰ \times ۰/۷ \times ۰/۳}{۹}\right) ۰/۰۱ \times ۰/۵۳ = ۰/۵۴$$

مزیت فرمول KR21 در سادگی آن است. همان‌طور که مشخص است برای تعیین پایایی با این روش، نیازی نیست که داده‌های مربوط به تک‌تک سؤالات را داشته باشیم. تنها با داشتن میانگین و واریانس کل آزمون و تعداد سؤالات می‌توان ضریب پایایی را تعیین کرد. البته این نکته را باید مدنظر قرار داد که ضریب به دست آمده از این فرمول دقیقاً برابر ضریب به دست آمده از فرمول KR20 نیست. در واقع، هرچه ضریب دشواری سؤالات، تفاوت بیشتری با یکدیگر داشته باشند، تفاوت بین دو فرمول بیشتر خواهد بود. به طوری که ضریب حاصل از فرمول KR20 دقیق‌تر و از لحاظ مقدار بزرگتر است.

آلفای کرونباخ

در روش آلفای کرونباخ[1] نیز آزمون یک بار اجرا می‌شود و همسانی درونی سؤالات آزمون مورد بررسی قرار می‌گیرند. فرمول آلفای کرونباخ که در واقع گسترش یافته فرمول کودر–ریچاردسون است، به صورت معمول هم برای تعیین پایایی سؤالات چندگزینه‌ای کاربرد دارد که پاسخ آنها متغیر دوحالته (صفر یا یک) است، هم برای پرسش‌های لیکرت که مثلاً در مقیاس ۱ تا ۵ تنظیم شده‌اند و هم برای سنجش همسانی درونی OSCE قابل استفاده است.

فرمول آلفای کرونباخ به صورت زیر است:

$$\text{پایایی} = \frac{\text{تعداد سوال}}{\text{تعداد سوال} - ۱}\left(۱-\frac{\text{مجموع واریانس سوالات}}{\text{واریانس آزمون}}\right)$$

به مثال جدول ۲–۳۴ توجه کنید. یک آزمون OSCE با ۱۰ ایستگاه ده‌نمره‌ای برای پنج دانشجو برگزار شده است. می‌خواهیم با استفاده از روش آلفای کرونباخ، پایایی کل آزمون را حساب کنیم. برای این کار باید واریانس هر ایستگاه را حساب کنیم.

1. Cronbach

جدول ۲-۳۴ : نمرات پنج دانشجو در یک OSCE با ۱۰ ایستگاه

	ایستگاه ۱	ایستگاه ۲	ایستگاه ۳	ایستگاه ۴	ایستگاه ۵	ایستگاه ۶	ایستگاه ۷	ایستگاه ۸	ایستگاه ۹	ایستگاه ۱۰
دانشجوی ۱	۱۰	۸	۸	۹	۵	۶	۷	۱۰	۷	۵
دانشجوی ۲	۶	۶	۳	۳	۰	۲	۶	۷	۵	۰
دانشجوی ۳	۸	۴	۵	۵	۲	۴	۴	۸	۵	۰
دانشجوی ۴	۹	۸	۹	۷	۷	۶	۸	۷	۶	۵
دانشجوی ۵	۵	۵	۷	۷	۴	۴	۴	۶	۷	۴
واریانس	۳/۴۴	۲/۵۶	۴/۶۴	۴/۱۶	۵/۸۴	۲/۲۴	۲/۵۶	۱/۸۴	۰/۸	۵/۳۶

و سپس آنها را با هم جمع کنیم تا مجموع واریانس ایستگاه‌ها به دست آید (۳۳/۴۴). همچنین واریانس کل آزمون را حساب می‌کنیم (۲۱۳/۸۴) و در فرمول جایگزین می‌نماییم:

$$\text{پایایی} = \frac{۱۰}{۹}\left(۱-\frac{۳۳/۴۴}{۲۱۳/۸۴}\right) = ۱/۱ \times ۰/۸۴ = ۰/۹۲$$

البته همان‌طور که در بخش مربوط به OSCE ذکر شد، آلفای کرونباخ OSCE را در دو سطح می‌توان محاسبه کرد:

- آلفای کرونباخ در سطح کل آزمون اطلاعاتی در مورد همسانی درونی کل OSCE به دست می‌دهد. روش محاسبه آن مانند مثال بالا است که توضیح داده شد. یعنی هر ایستگاه به صورت یک سؤال با دامنه نمره مشخصی در نظر گرفته می‌شود.
- آلفای کرونباخ در سطح هر ایستگاه همسانی درونی یک ایستگاه را نشان می‌دهد و کیفیت چک‌لیست را مشخص می‌کند. یعنی هر یک از آیتم‌های چک‌لیست به عنوان یک متغیر در نظر گرفته می‌شوند. به مثال جدول ۳-۳۴ توجه کنید. نمرات دانشجویان در ایستگاه اول آزمون فوق به صورت زیر است. برای به دست آوردن آلفای کرونباخ ایستگاه به روش زیر عمل می‌کنیم. ابتدا واریانس هر آیتم و سپس مجموع آنها را حساب می‌کنیم. واریانس کل ایستگاه را نیز به دست می‌آوریم و در فرمول می‌گذاریم:

جدول ۳-۳۴ : نمرات پنج دانشجو در ۱۰ آیتم یک ایستگاه OSCE

	آیتم۱	آیتم۲	آیتم۳	آیتم۴	آیتم۵	آیتم۶	آیتم۷	آیتم۸	آیتم۹	آیتم۱۰
دانشجوی ۱	۱	۱	۱	۱	۱	۱	۱	۱	۱	۱
دانشجوی ۲	۱	۰	۰	۱	۱	۱	۱	۰	۱	۰
دانشجوی ۳	۱	۱	۱	۰	۱	۱	۱	۱	۰	۱
دانشجوی ۴	۱	۱	۱	۱	۱	۱	۱	۰	۱	۱
دانشجوی ۵	۱	۱	۱	۰	۰	۰	۱	۰	۱	۰
واریانس	**آیتم۱**	**آیتم۲**	**آیتم۳**	**آیتم۴**	**آیتم۵**	**آیتم۶**	**آیتم۷**	**آیتم۸**	**آیتم۹**	**آیتم۱۰**
	۰	۰/۱۶	۰/۱۶	۰/۲۴	۰/۱۶	۰/۱۶	۰	۰/۲۴	۰/۱۶	۰/۲۴

مجموع واریانس ایتم‌ها ۱/۵۲ و واریانس کل ایستگاه ۳/۴۴ به دست می‌آید.

$$\text{پایایی ایستگاه} = \frac{۱۰}{۹}\left(۱-\frac{۱/۵۲}{۳/۴۴}\right) = ۱/۱ \times ۱/۵۵ = ۰/۶۱$$

از آنجا که در مثال فوق، پاسخ‌ها به صورت دوحالته هستند، می‌توان به جای آلفای کرونباخ از فرمول کودر–ریچاردسون نیز استفاده کرد.

پیش‌تر اشاره شد که علاوه بر اینکه منظور از همسانی درونی می‌تواند هم راستا بودن سؤالات یک آزمون باشد، همچنین می‌تواند به همسانی اجزای مختلف یک آزمون برگردد. به عنوان مثال، آزمون جامع علوم پایه پزشکی از چندین درس مختلف تشکیل شده است. در این مواقع برای به دست آوردن پایایی، از فرمول آلفای کرونباخ به شکل زیر می‌توان استفاده کرد:

$$\text{پایایی} = \frac{\text{تعداد اجزا}}{\text{۱– تعداد اجزا}}\left(۱-\frac{\text{مجموع واریانس اجزا}}{\text{واریانس کل آزمون}}\right)$$

آزمون آناتومی تنه از سه بخش توراکس، شکم و لگن تشکیل شده است که واریانس آنها به ترتیب ۱، ۰/۶۴ و۱/۴۴ است. با توجه به واریانس کل آزمون که ۴ است، آلفای کرونباخ آزمون به صورت زیر حساب می‌شود:

$$\text{پایایی} = \frac{۳}{۲}\left(۱-\frac{۳/۰۸}{۴}\right) = ۱/۵ \times ۰/۲۳ = ۰/۳۴$$

همان‌گونه که مشخص است، هر چه تعداد سؤالات آزمون بیشتر باشد، آلفای کرونباخ افزایش خواهد یافت. با داشتن ضریب پایایی یک آزمون می‌توان پیش‌بینی کرد با افزودن یا کاستن تعدادی از سؤالات ضریب پایایی جدید چند خواهد بود. برای این کار از فرمول اسپیرمن–براون[1] استفاده می‌شود که کاربرد اصلی آن در تخمین پایایی یک آزمون بعد از تغییر تعداد سؤالات آن است. فرمول آن به صورت زیر است:

$$\text{پایایی کل آزمون} = \frac{\text{پایانی فعلی} \times \text{نسبت تعداد کل سوالات جدید به تعداد سوالات فعلی}}{[\text{پایانی فعلی} \times (\text{۱– نسبت تعداد سوالات جدید به تعداد سوالات فعلی})] + ۱}$$

به مثال زیر توجه کنید. پایایی یک آزمون ۵۰ سؤالی ۰/۷ است. اگر ۲۰ سؤال اضافه کنیم. پایایی آزمون جدید به صورت فرمول صفحه بعد است:

$$\text{پایایی جدید آزمون} = \frac{۲/۵ \times ۰/۷}{۱+ (۱/۵\times۰/۷)} = \frac{۱/۷۵}{۲/۰۵} = ۰/۸۵$$

آلفای کرونباخ در صورت حذف آیتم نیز شاخص مفید دیگری است که اطلاعاتی در مورد تک‌تک سؤالات به دست می‌دهد. با محاسبه «آلفا در صورت حذف آیتم» می‌توان به اطلاعاتی در خصوص هر سؤال دست یافت. این محاسبه توسط SPSS به راحتی قابل انجام است. به این معنا که هر بار یکی از سؤالات آزمون حذف شده و مجدداً آلفای کل آزمون محاسبه می‌شود. به طور معمول انتظار داریم با حذف یک سؤال و کاهش تعداد سؤالات، آلفا کاهش پیدا کند، اگر مقدار آلفا در صورت حذف آیتم افزایش نشان داد، پی می‌بریم که احتمالاً مشکلی در آن آیتم (سؤال یا ایستگاه) وجود داشته است که

1. Spearman-Brown

با سایر سؤالات و کل آزمون هم‌راستا نبوده است به گونه‌ای که کنار گذاشتن آن موجب بهبود همسانی درونی شده است.

دو نیمه کردن آزمون

روش دو نیمه کردن آزمون[1] نیز بر همسانی درونی سؤالات آزمون تاکید می‌کند و در یک نوبت اجرای آزمون قابل انجام است. روش اجرا به این شکل است که پس از برگزاری آزمون، سؤالات به دو نیمه تقسیم می‌شوند. مبنای تقسیم می‌تواند زوج یا فرد بودن شماره سؤالات باشد. نمرات هر یک از نیمه‌ها جداگانه حساب می‌شود و ضریب همبستگی بین نمرات دو نیمه به عنوان ضریب پایایی هر یک در نظر گرفته می‌شود. برای تعیین ضریب پایایی کل آزمون از فرمول اسپیرمن–براون استفاده می‌شود که قبلاً توضیح داده شد که کاربرد اصلی آن در تخمین پایایی یک آزمون بعد از تغییر تعداد سؤالات آن است. فرمول ساده شده آن به صورت زیر است:

$$\text{پایایی کل آزمون} = \frac{\text{ضریب همبستگی بین دو نیمه} \times ۲}{\text{ضریب همبستگی بین دو نیمه} + ۱}$$

به آزمون جدول ۲–۳۴ که قبلاً پایایی آن را با آلفای کرونباخ محاسبه کرده بودیم، توجه کنید. یک آزمون OSCE با ۱۰ ایستگاه ده نمره‌ای بود که برای پنج دانشجو برگزار شده است. می‌خواهیم با استفاده از روش دو نیمه کردن پایایی کل آزمون را حساب کنیم. برای این کار آزمون را بر اساس سؤالات زوج و فرد آن تقسیم می‌کنیم و سپس ضریب همبستگی بین آن دو را حساب می‌کنیم. در اینجا فرمول ضریب همبستگی پیرسون را به کار می‌بریم. به این منظور، نمره هر دانشجو را در ایستگاه‌های فرد و زوج به تفکیک محاسبه می‌کنیم و هر یک را به مجذور می‌رسانیم. همچنین حاصل‌ضرب ایستگاه‌های زوج و فرد را برای هر دانشجو حساب می‌کنیم. از طرفی مجموع نمرات دانشجویان در ایستگاه‌های زوج را حساب کرده و به توان دو می‌رسانیم. همین کار را برای ایستگاه‌های فرد تکرار می‌کنیم (جدول ۴–۳۴):

همان‌طور که قبلا ذکر شد، کاربرد فرمول اسپیرمن–براون فقط برای دو نیمه کردن آزمون نیست و اصولاً هرگاه ما ضریب پایایی یک آزمون را داشته باشیم و بعد بخواهیم تعدادی از سؤالات آزمون را کم یا زیاد کنیم، از طریق این فرمول می‌توانیم ضریب پایایی جدید را تخمین بزنیم.

جدول ٤–۳٤ : محاسبه پایایی به روش دو نیمه کردن آزمون

	نمره در سؤالات فرد	نمره در سؤالات زوج	مجذور نمره سؤالات فرد	مجذور نمره سؤالات زوج	حاصل‌ضرب دو نیمه
دانشجوی ۱	۳۷	۳۸	۱۳۶۹	۱۴۴۴	۱۴۰۶
دانشجوی ۲	۲۰	۱۸	۴۰۰	۳۲۴	۳۶۰
دانشجوی ۳	۲۴	۲۱	۵۷۶	۴۴۱	۵۰۴
دانشجوی ۴	۳۹	۳۳	۱۵۲۱	۱۰۸۹	۱۲۸۷
دانشجوی ۵	۲۷	۲۶	۷۲۹	۶۷۶	۷۰۲
جمع	۱۴۷	۱۳۶	۴۵۹۵	۳۹۷۴	۴۲۵۹
مجذور جمع	۲۱۶۰۹	۱۸۴۹۶			

1. Split-halves

$$\text{ضریب همبستگی پیرسون} = \frac{(۵ \times ۴۲۵۹) - (۱۴۷ \times ۱۳۶)}{\sqrt{[(۵ \times ۳۹۷۴)) - ۱۸۴۹۶][(۵ \times ۴۵۹۵) - ۲۱۶۰۹]}} = \frac{۲۱۲۹۵ - ۱۹۹۹۲}{\sqrt{۱۳۶۶ \times ۱۳۷۴}} = \frac{۱۳۰۳}{۱۳۶۹} = ۰/۹۵$$

$$\text{پایایی کل آزمون} = \frac{۲ \times ۰/۹۵}{۱ + ۰/۹۵} = ۰/۹۵$$

توافق بین آزمونگران

در امتحان چندگزینه‌ای، جواب هر سؤال به صورت عینی است و تصحیح پاسخ‌های دانشجویان نیازی به بررسی و قضاوت ندارد. به طوری که کار تصحیح می‌تواند توسط ماشین انجام شود. اما در برخی از آزمون‌ها تصمیم‌گیری در مورد نمره‌ای که به هر سؤال تعلق می‌گیرد، منوط به قضاوت آزمونگر است. از آنجا که احتمال زیادی وجود دارد که قضاوت دو نفر با هم تفاوت‌هایی داشته باشد، اگر از چند آزمونگر استفاده شود، بسیار محتمل است که یک دانشجو نمرات متفاوتی کسب کند. به عنوان مثال، آزمون‌های کوته‌پاسخ، تشریحی، شفاهی و حتی OSCE چنین حالتی دارند.

در این موارد که آزمونگر به عنوان منبع خطای اندازه‌گیری شناخته می‌شود، میزان توافق بین آزمونگران در مورد نمره یک دانشجو، به عنوان شاخصی از پایایی آزمون محسوب می‌شود. برای محاسبه آن از چند شاخص می‌توان استفاده کرد که با ذکر مثال به آنها می‌پردازیم.

نتیجه آزمون ممکن است به صورت متغیر اسمی باشد مثلاً در جدول ۵-۳۴ از آزمونگران خواسته شده است وضعیت قبولی یا ردی هر یک از ده دانشجو را پس از یک آزمون شفاهی مشخص کنند. خلاصه نظر دو آزمونگر در جدول ۶-۳۴ نمایش داده شده است.

جدول ۵-۳٤ : نتیجه ارزیابی دو آزمونگر در مورد رد (صفر) و قبول (یک) ۱۰ دانشجو

دانشجو	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰	۱۱	۱۲	۱۳	۱۴	۱۵
آزمونگر ۱	۰	۱	۱	۱	۰	۱	۱	۱	۰	۱	۱	۱	۰	۱	۱
آزمونگر ۲	۱	۱	۱	۱	۱	۰	۰	۱	۰	۱	۱	۱	۰	۰	۰

جدول ۶-۳٤ : خلاصه نظر دو آزمونگر

	قبول توسط آزمونگر ۲	رد توسط آزمونگر ۲	جمع
قبول توسط آزمونگر ۱	۷	۴	۱۱
رد توسط آزمونگر ۱	۲	۲	۴
جمع	۹	۶	۱۵

یک روش ساده برای بررسی میزان توافق بین دو آزمونگر، محاسبه درصد موارد یکسان است. در آزمون فوق دو آزمونگر در خصوص ردی ۲ دانشجو و همچنین قبولی ۷ دانشجو توافق داشته‌اند. به عبارت دیگر در ۹ مورد از ۱۵ مورد یعنی ۶۰ درصد با یکدیگر موافق بوده‌اند پس میزان پایایی بین آزمونگران ۶۰ درصد است. این مقدار چندان خوب نیست و در واقع فقط کمی بیش از توافقی است که صرفاً به صورت شانسی (۵۰ درصد) محتمل بود رخ دهد.

اگر نظرات آزمونگران را به صورت خلاصه به صورت جدول ۷-۳۴ در نظر بگیریم، می‌توان این محاسبه را به این شکل انجام داد:

$$\text{میزان توافق بین آزمونگران} = \frac{a+d}{N}$$

جدول ۷-۳٤ : خلاصه نظر دو آزمونگر

	قبول توسط آزمونگر ۲	رد توسط آزمونگر ۲	جمع
قبول توسط آزمونگر ۱	a	b	a+b
رد توسط آزمونگر ۱	c	d	c+d
جمع	a+c	b+d	N

این روش برای بررسی توافق اگرچه ساده است اما بهترین روش نیست. یکی از دلایلی که در خصوص نامناسب بودن این این نحوه محاسبه ذکر شده است، این است که میزان شانس و تصادف را در بروز توافق بین دو آزمونگر نادیده می‌گیرد.

روش دیگری که برای تعیین توافق بین آزمونگران در خصوص یک متغیر اسمی کاربرد دارد، محاسبه ضریب کاپا است. فرمول‌های متنوعی برای محاسبه کاپا وجود دارند که برای اهداف مختلف اصلاح شده‌اند. در اینجا به نوع معمول آن اشاره می‌کنیم که میزان شانس را نیز در برآورد خود در نظر می‌گیرد:

$$\text{ضریب کاپا} = \frac{\text{توافق مورد انتظار- توافق مشاهده شده}}{\text{توافق مورد انتظار-۱}}$$

برای محاسبه توافق مشاهده‌شده، درصد مواردی که دو نفر به صورت یکسان تصمیم‌گیری کرده‌اند و توافق داشته‌اند، بر تعداد کل موارد تقسیم می‌شود:

$$\text{توافق مشاهده شده} = \frac{a+d}{N}$$

احتمال توافق مورد انتظار (تصادفی) از فرمول زیر به دست می‌آید:

$$\text{توافق مورد انتظار} = \left[\left(\frac{a+b}{N}\right)\times\left(\frac{a+c}{N}\right)\right]+\left[\left(\frac{a+d}{N}\right)\times\left(\frac{b+d}{N}\right)\right]$$

به عبارت دیگر، برای محاسبه توافق مورد انتظار باید ببینیم احتمال اینکه هر دو آزمونگر تصمیم مبتنی بر قبولی و تصمیم مبتنی بر ردی بگیرند، چقدر بوده است. برای این کار مراحل زیر را انجام می‌دهیم:

- برآورد اینکه هر دو آزمونگر تصمیم قبولی بگیرند: ابتدا باید احتمال تصمیم قبولی هر آزمونگر را حساب کنیم و سپس دو مقدار را در یکدیگر ضرب کنیم. در مثال جدول ۷-۳۴ می‌بینیم که آزمونگر اول، یازده مورد قبول و ۴ مورد تصمیم رد داشته است. یعنی او در حدود ۷۳ درصد موارد نتیجه قبول اعلام کرده است. آزمونگر دوم از ۹ دانشجو، ۷ تصمیم قبول داشته است یعنی در ۶۰ درصد موارد تصمیم قبول گرفته است. بنابراین احتمال اینکه دو آزمونگر تصمیم قبولی بگیرند، ۴۴ درصد (۷۳ ضرب در ۶۰) است.
- برآورد اینکه چقدر احتمال دارد هر دو نفر تصمیم مبتنی بر ردی بگیرند: آزمونگر اول در ۲۶ درصد موارد و آزمونگر دوم در

۴۰ درصد موارد تصمیم ردی گرفته‌اند بنابراین احتمال اینکه هر دو تصمیم ردی بگیرند، ۱۰ درصد (۲۶ ضرب در ۴۰) است.

- احتمال توافق تصادفی بین دو آزمونگر از جمع دو مقدار بالا به دست می‌آید. به عبارت دیگر احتمال توافق تصادفی بین دو آزمونگر ۵۴ درصد (۴۴ به اضافه ۱۰) است. بنابراین ضریب کاپا برابر است با:

$$\text{توافق مورد انتظار} = \left[\left(\frac{۷+۴}{۱۵}\times\frac{۷+۲}{۱۵}\right)\right]+\left[\left(\frac{۲+۲}{۱۵}\times\frac{۴+۲}{۱۵}\right)\right] = ۰/۴۶ + ۱۰ = ۰/۵۴$$

$$\text{ضریب کاپا} = \frac{۰/۶-۰/۵۴}{۱-۰/۵۴} = ۰/۱۳$$

هرچند مثال فوق فقط برای نمونه ارائه شده بود و به علت حجم نمونه پایین، قابل تعمیم نیست اما به خوبی نشان می‌دهد که میزان توافق تحت تاثیر شانس قرار می‌گیرد چنانچه مقدار آن پس از کسر احتمال توافق تصادفی، نسبت به روش اول کاهش یافته است.

خطای معیار اندازه‌گیری

همان‌طور که چندین بار گفته شد، خطا در ذات اندازه‌گیری وجود دارد. هنگام اندازه‌گیری طول یک میز افراد مختلف ممکن است به اعداد گوناگونی دست پیدا کنند که هر یک از آنها به درجاتی با خطا همراه است و هیچ‌یک از آنها اندازه حقیقی طول میز نیست. در آزمون نیز که در واقع اندازه‌گیری دانش یا توانمندی دانشجویان است، بروز خطا اجتناب ناپذیر است و موجب می‌شود نمره مشاهده‌شده دانشجو به درجاتی واجد خطا باشد. حال سؤال این است که چه نمره‌ای، نمره واقعی دانشجو است. واقعیت این است که در هیچ اندازه‌گیری نمی‌توان نمره واقعی را مشخص کرد و همیشه خطا در ذات اندازه‌گیری وجود دارد. نمره واقعی دست نیافتنی است و تنها می‌توان حدود آن را تخمین زد. در نظریه کلاسیک، فرض بر این است که اندازه مشاهده شده، از جمع اندازه واقعی و اندازه ناشی از خطای تصادفی تشکیل شده است.

نمره خطا + نمره واقعی= نمره مشاهده شده

فرض کنید که نمره مشاهده شده، نمره خطا و نمره واقعی دانشجویان در یک آزمون به صورت جدول ۸-۳۴ است. برای هر گروه از نمرات زیر می‌توان واریانس محاسبه کرد:

جدول ۸-۳٤ : نمرات مشاهده شده، واقعی و خطای دانشجو در یک آزمونر

شماره دانشجو	نمره مشاهده شده	نمره خطا	نمره واقعی
۱	۱۹/۵	-۰/۵	۲۰
۲	۹	-۱	۱۰
۳	۱۲	۰	۱۲
۴	۱۵	۰/۵	۱۴/۵
۵	۱۷	۲	۱۵
۶	۱۱	-۲	۱۳
۷	۱۴	۰	۱۴
۸	۱۸/۵	۰/۵	۱۸
۹	۱۱/۵	۱/۵	۱۰
۱۰	۱۴	-۱	۱۵
واریانس	۱۰/۴۳	۱/۳	۹/۱۳

همان‌طور که برای نمرات مشاهده‌شده می‌توان توزیع رسم کرد و واریانس محاسبه کرد، می‌توان توزیع و واریانس نمرات خطا را نیز به دست آورد و رابطه آنها را به شکل زیر نشان داد:

واریانس نمره خطا + واریانس نمره واقعی= واریانس نمره مشاهده شده

۱۰/۴۳ =۹/۱۳ + ۱/۳۰

به خاطر داریم که ضریب پایایی، از تقسیم واریانس نمرات واقعی بر واریانس نمرات مشاهده‌شده محاسبه می‌شود، یعنی:

$$\text{پایایی} = \frac{\text{واریانس نمرات واقعی}}{\text{واریانس نمرات مشاهده شده}}$$

اکنون با در نظر داشتن فرمول مربوط به جمع واریانس نمرات خطا و واقعی، یعنی:

واریانس نمره خطا + واریانس نمره واقعی= واریانس نمره مشاهده شده

یا به عبارت دیگر:

واریانس خطا – واریانس نمرات مشاهده شده= واریانس نمره واقعی

می‌توان نوشت:

$$\text{پایایی} = \frac{\text{واریانس نمرات خطا– واریانس نمرات مشاهده شده}}{\text{واریانس نمرات مشاهده شده}} = ۱- \frac{\text{واریانس نمرات خطا}}{\text{واریانس نمرات مشاهده شده}}$$

در همان مثال مربوط به جدول ۴–۲۲، پایایی به صورت زیر است:

$$\text{پایایی} = ۱- \frac{\text{واریانس نمرات خطا}}{\text{واریانس نمرات مشاهده شده}} = ۱- \frac{۱/۳۰}{۱۰/۴۳} = ۰/۸۷$$

درست است که در مثال بالا مقادیری برای خطا به صورت فرضی در نظر گرفتیم اما حقیقت این است اساساً مقادیر خطا در اندازه‌گیری نامشخص است. پس چگونه می‌توان نمره واقعی دانشجو را تخمین زد؟ بر اساس فرمول پایایی که ذکر شد، می‌توان نوشت:

$$\text{پایایی} = ۱- \frac{\text{واریانس نمرات خطا}}{\text{واریانس نمرات مشاهده شده}}$$

یا به عبارت دیگر:

(پایایی–۱) × واریانس نمره مشاهده شده= واریانس نمرات خطا

با گرفتن ریشه دوم، این تساوی به صورت زیر در می‌آید:

$$\text{انحراف معیار نمرات خطا} = \text{انحراف معیار نمرات مشاهده شده} \times \sqrt{1-\text{پایایی}}$$

انحراف معیار نمرات خطا به نام خطای معیار اندازه‌گیری[1] (SEM) معروف است و برای تخمین نمره واقعی از آن استفاده می‌شود. به این ترتیب که ۶۸ درصد احتمال دارد نمره مشاهده شده در فاصله مثبت و منفی یک SEM از نمره واقعی باشد. تقریباً ۹۵ درصد احتمال دارد در فاصله مثبت و منفی دو SEM از نمره واقعی باشد و تقریباً ۹۹ درصد محتمل است که در فاصله مثبت و منفی سه SEM از نمره واقعی باشد.

به مثال زیر توجه کنید. آزمونی با ۲۰ سؤال چندگزینه‌ای برگزار شده است که میانگین آن ۱۵ و انحراف معیار آن ۳ است. می‌خواهیم ببینیم نمره واقعی دانشجویی که در امتحان نمره ۱۲ کسب کرده است، به احتمال ۹۵ درصد چقدر است. برای این کار ابتدا پایایی آزمون را با فرمول کودر–ریچاردسون حساب می‌کنیم:

$$\text{پایایی} = \frac{\text{تعداد سوال}}{\text{تعداد سوال} - ۱} \left(۱ - \frac{\text{میانگین دشواری سوالات} \times \text{تعداد سوال} \times (\text{یک منهای میانگین دشواری سوالات})}{\text{واریانس آزمون}}\right)$$

$$\text{میانگین دشواری سوالات} = \frac{۱۵}{۲۰} = ۰/۷۵$$

$$\text{پایایی} = \frac{۱۰۰}{۹۹}\left(۱ - \frac{۱/۱۰ \times ۰/۷۵ \times ۰/۲۵}{۹}\right) = ۱/۰۱ \times ۰/۵۸ = ۰/۵۹$$

$$\text{خطای معیار اندازه‌گیری} = ۳ \times \sqrt{۱-۰/۵۹} = ۱/۹۲ = ۱/۹۲$$

۹۵ درصد احتمال دارد نمره واقعی دانشجو (یعنی نمره ۱۲) در بازه دو SEM (یعنی ۳/۸۴) بیشتر یا کمتر از نمره مشاهده‌شده وی باشد. به عبارت دیگر می‌توان گفت نمره واقعی دانشجو به احتمال ۹۵ درصد در بازه ۸/۱۶ تا ۱۵/۸۴ است. از آنجا که پایایی آزمون پایین بود، میزان خطای اندازه‌گیری زیاد است و دامنه نمره واقعی که تخمین زدیم، گسترده شده است.

پایایی در نظریه تعمیم‌پذیری

خطاهای آزمون می‌توانند از منابع مختلفی نشأت بگیرند و با دور کردن نمرات مشاهده‌شده از نمرات واقعی دانشجویان، بر پایایی آزمون تاثیر بگذارند. آشناترین منابع خطا، سؤالات امتحان و آزمونگران هستند. به عنوان مثال در یک آزمون چندگزینه‌ای با صد آیتم اگر سؤالات دیگری انتخاب می‌شدند، احتمال داشت نمره دانشجویان دقیقاً همین نمره فعلی نباشد. سؤال این است که این آزمون با همین سؤالات چقدر خطا داشته است و نمره‌ای که دانشجو در آن کسب کرده است، چقدر قابل تعمیم به سؤالات ممکن دیگر است. باز هم به عنوان نمونه در OSCE اگر آزمونگران دیگری در ایستگاه مستقر باشند، احتمالاً نمره دانشجویان متفاوت خواهد بود و باز هم این پرسش مطرح است که تفاوت نمره ناشی از خطای اندازه‌گیری بوده است یا خیر.

1. Standard Error of Measurement (SEM)

همان‌طور که قبلاً گفته شد، نظریه کلاسیک آزمون نمی‌تواند در آن واحد تمام منابع خطا را در نظر بگیرد و سهم هر یک از آنها را در تولید خطای اندازه‌گیری برآورد نماید. به عنوان مثال در یک OSCE با آلفای کرونباخ می‌توان خطای ناشی از آیتم‌ها را برآورد کرد و با بررسی توافق بین آزمونگران می‌توان منبع خطای مربوط به آزمونگران را بررسی کرد.

نظریه تعمیم‌پذیری که در واقع بسط یافته نظریه کلاسیک آزمون است در سال ۱۹۷۲ توسط کرونباخ معرفی شد. به کمک این نظریه می‌توان مقدار اثر هر یک از منابع خطا را برآورد کرد تا تصویر واضح‌تری از خطای اندازه‌گیری به دست آید و تفسیر دقیق‌تری از نمرات قابل ارائه باشد. به این ترتیب متولیان آزمون قادر خواهند بود منابع اصلی خطا را شناسایی کنند و برای برگزاری آزمونی پایاتر برنامه‌ریزی کنند. پایه محاسبات در نظریه تعمیم‌پذیری بر اساس مدل‌های آماری آنالیز واریانس[1] (ANOVA) است.

با توجه به مواردی که در فصل اول کتاب در مورد نظریه تعمیم‌پذیری عنوان شد، در اینجا مفاهیم پایه این نظریه را مرور می‌کنیم و با ذکر چند مثال کاربرد آن را در بحث ارزیابی دانشجویان بررسی می‌نماییم.

رویه و انواع آن

به طور کلی در نظریه تعمیم‌پذیری، هر اندازه‌گیری (مانند آزمون) دارای خصوصیاتی است که به هر یک از آنها یک رویه[2] گفته می‌شود که می‌توانند در اندازه‌گیری مشاهده‌شده تاثیر بگذارند و به نوعی منابع خطای اندازه‌گیری محسوب می‌شوند. به عنوان مثال در آزمون OSCE، تعداد ایستگاه‌ها، موارد بالینی، بیمارنماها، شرایط آزمون، عملکرد ارزیابان و ... هریک به عنوان یک رویه از آزمون در نظر گرفته می‌شوند که می‌توانند موجب شوند نمره دانشجو با نمره واقعی وی متفاوت باشد.

رویه‌ای که مورد اندازه‌گیری قرار گرفته است و به عبارت دیگر Object of Measurement است، به عنوان رویه‌ای که قرار است تمایز در سطح آن اتفاق بیفتد، در نظر گرفته می‌شود و به آن رویه تمیز[3] گفته می‌شود. به عنوان مثال، دانشجو در آزمون OSCE رویه تمیز است. در واقع، این رویه منبع خطا نیست زیرا تفاوتی که بین نمرات دانشجویان مختلف دیده می‌شود، ناشی از خطا نیست و منعکس‌کننده تفاوت واقعی در سطح توانایی آنها است. برعکس، در مورد سایر رویه‌ها مانند ایستگاه یا آزمونگر، می‌خواهیم مشخص کنیم نمره دانشجو تا چه حد به سایر ایستگاه‌ها یا آزمونگرانی که به صورت بالقوه می‌توانستیم داشته باشیم، قابل تعمیم است. به همین دلیل به چنین رویه‌هایی، رویه تعمیم[4] گفته می‌شود. هرچه پراکندگی بر اساس رویه تمیز بیشتر باشد، بهتر و مناسب‌تر است و هر چه واریانس ناشی از رویه‌هایی به جز رویه مورد اندازه‌گیری بیشتر باشد، پایایی آزمون کمتر خواهد بود.

درست است که منابع خطا در یک آزمون متعدد هستند و با استفاده از نظریه تعمیم‌پذیری می‌توان در آن واحد سهم هر یک را در تولید خطای آزمون به صورت جداگانه مشخص کرد اما مهم است که منابع خطای مهم و اصلی از قبل شناسایی شوند و داده‌های مربوط به تک تک آنها حین آزمون جمع‌آوری گردد. در غیر این صورت، نمی‌توان محاسبات را انجام داد. به عنوان مثال، در یک آزمون OSCE رویه‌ها شامل ایستگاه‌ها، آیتم‌های چک‌لیست‌ها، آزمونگران، بیماران استانداردشده، لاین تقسیم‌بندی و ... هستند اما ممکن است ما تنها بخواهیم نقش آزمونگر و ایستگاه را بررسی کنیم. در این حالت صرفاً داشتن نمره کل دانشجو کافی نیست. باید نمره هر دانشجو را به تفکیک ایستگاه و همچنین آزمونگر در مجموعه داده‌های مربوطه وارد کنیم.

1. ANOVA
2. Facet
3. Facet of differentiation
4. Facet of generalization

نمره جهانی

در بحث پایایی همیشه به دنبال برآورد میزان خطا به منظور تخمین حدود نمره واقعی دانشجویان بودیم. در نظریه تعمیم‌پذیری هم مشابه همین مفهوم وجود دارد اما به جای لفظ نمره واقعی از نمره جهانی[1] استفاده می‌شود. در واقع، پایه و اساس مفهومی نظریه تعمیم‌پذیری بر اساس جهان مشاهدات قابل‌قبول[2] است که به این صورت تعریف می‌شود: کل مشاهدات ممکنی که در شرایط اندازه‌گیری خاص برای محقق قابل‌قبول است. درواقع کل رویه‌ها و سطوح آنها برای یک اندازه‌گیری معین این جهان را می‌سازد. به طور خلاصه در نظریه تعمیم‌پذیری برای مجموعه شرایط اندازه‌گیری از اصطلاح جهان[3] و برای کل مجموع افراد مورد اندازه‌گیری از اصطلاح جمعیت[4] استفاده می‌شود.

تصور کنید یک آزمون OSCE با ۱۰ ایستگاه برگزار شده است. یکی از منابع خطای این آزمون ایستگاه‌های آن هستند. به این معنا که اگر طراحان سؤال به جای این ۱۰ ایستگاه، ایستگاه‌های دیگری طراحی کرده بودند، احتمالاً نمرات دانشجویان دقیقاً به همین اندازه نمی‌شد. بنابراین، خطایی در آزمون ما وجود دارد که ناشی از ایستگاه است و باید مقدار آن برآورد شود. فرض کنید به صورت بالقوه برای انجام این آزمون ۵۰ ایستگاه می‌توانستیم طراحی کنیم که در حال حاضر تنها ۱۰ ایستگاه از بین آنها انتخاب شده است. برای ما مهم است بدانیم نمره‌ای که دانشجو در این آزمون ۱۰ ایستگاهی کسب کرده است، تا چه حد قابل تعمیم به تمام ۵۰ ایستگاه فرضی است. آن حالت بالقوه، جهان تعمیم[5] نام دارد و به حالت عملی‌شده، جهان مشاهدات قابل قبول می‌گویند.

در برخی از موارد این دو کاملاً با هم همپوشانی دارند و یکسان هستند. مثلاً هنگامی که در یک آزمون از پنج آزمونگر استفاده می‌شود و کل تعداد استادان گروه مربوطه هم پنج نفر است، آنچه ممکن و عملی شده است، بر آنچه تعمیم می‌دهیم، منطبق است. تعیین این موضوع اهمیت دارد زیرا در مدل آنالیزی که انجام خواهیم داد، موثر است. به این ترتیب که اگر تعداد رویه مورد نظر در جهان ممکن و تعمیم منطبق بر هم باشند (مثلاً همان پنج آزمونگر)، از مدل اثرات ثابت[6] استفاده می‌شود و اگر مانند مثال ایستگاه تعداد رویه در جهان تعمیم بیشتر باشد، از مدل اثرات تصادفی[7] استفاده می‌گردد.

ضریب تعمیم‌پذیری

ضریب تعمیم‌پذیری[8] مقداری بین صفر تا یک دارد و معادل پایایی آزمون در نظر گرفته می‌شود. این ضریب مشخص می‌کند که نمره دانشجو را تا چه حد می‌توان به تمام رویه‌ها و منابع خطای در نظر گرفته شده تعمیم داد. به عبارت دیگر مشخص می‌کند که نمره دانشجو تا چه حد به نمره جهانی او نزدیک است.

نحوه محاسبه ضریب تعمیم‌پذیری به این ترتیب است که ابتدا باید رویه‌های موردنظر را تعیین کرد. هر رویه واجد مقداری به نام جزء واریانس[9] است که از طریق آزمون آماری آنالیز واریانس (ANOVA) محاسبه می‌شود. از اجزای واریانس رویه‌های مختلف در نهایت ضریب تعمیم‌پذیری به دست می‌آید.

فرمول‌های متنوعی برای محاسبه ضریب تعیم‌پذیری یک آزمون وجود دارد که بسته به اینکه کدام منابع خطا را در نظر می‌گیریم و چگونه نمرات را تفسیر می‌کنیم، یکی از آنها را انتخاب می‌کنیم. به عنوان مثال، از لحاظ مفهومی می‌توان دو نوع ضریب تعمیم‌پذیری مطلق[10] و نسبی[11] برای آزمون در نظر گرفت که به هدف برگزاری آزمون بر می‌گردد. اگر

1. Universe score
2. Universe of admissible observations
3. Universe
4. Population
5. universe of generalization
6. Fixed effects
7. Random effects
8. G coefficient
9. Variance component
10. Absolute
11. Relative

قرار است از نمرات برای تصمیم‌گیری به صورت معیارمحور استفاده شود، فرمول ضریب مطلق استفاده می‌شود اما اگر قرار است از نتایج آزمون به صورت هنجارمحور و برای رتبه‌بندی دانشجویان استفاده شود، از فرمول ضریب نسبی استفاده می‌شود. نحوه محاسبه این دو ضریب با هم فرق دارد که بعداً به آن خواهیم پرداخت. فعلاً برای یادآوری و سهولت درک معنای ضریب تعمیم‌پذیری، با مثالی از یک آزمون تک رویه شروع می‌کنیم.

محاسبه ضریب تعمیم‌پذیری در حالت تک‌رویه

در عالم واقع آنچه در یک آزمون به عنوان منابع خطا دخیل است، محدود به یک رویه نیست اما می‌توان در نظر گرفت که در یک آزمون فرضی تنها نقش یک رویه برای ما مهم است و می‌خواهیم تاثیر آن را بسنجیم. اصول کار همانند نظریه کلاسیک آزمون است. یعنی:

$$G = \frac{\sigma^2 \text{ student}}{\sigma^2 \text{ student} + \sigma^2 \text{ error}}$$

یک آزمون شفاهی ساختارمند را در نظر بگیرید که با سه آزمونگر برای ۱۰ دستیار نورولوژی برگزار می‌شود. آزمونگران به صورت مستقل سؤالاتی از هر دستیار می‌پرسند و نمره‌ای بین ۱ تا ۱۰ به وی اختصاص می‌دهند (جدول ۹-۳۴). برای برآورد میزان خطای این آزمون تنها یک رویه یعنی آزمونگر را در نظر می‌گیریم.

اگر برای داده‌های فوق ANOVA را به صورت معمول انجام دهیم، دو جدول برای منابع تغییرات بین گروه[1] (دستیاران) و درون گروه[2] (آزمونگران و خطا) خواهد داد که خلاصه داده‌های آنها در جدول ۱۰-۳۴ آمده است:

جدول ۹-۳٤ : نمرات ده دستیار از سه آزمونگر

شماره دانشجو	آزمونگر یک	آزمونگر دو	آزمونگر سه
۱	۶	۷	۸
۲	۴	۵	۶
۳	۲	۲	۲
۴	۳	۴	۵
۵	۵	۴	۶
۶	۸	۹	۱۰
۷	۵	۷	۹
۸	۶	۷	۸
۹	۴	۶	۸
۱۰	۷	۹	۸

1. Between-Subjects
2. Within-Subjects

جدول ۱۰-۳٤ : نتیجه آزمون ANOVA درون گروه و بین گروه

منبع	جمع مربعات	درجه آزادی	میانگین مربعات
آزمونگر	۲۰	۲	۱۰
خطا	۱۰	۱۸	۰/۵۶
دستیار	۱۱۴	۹	۱۲/۶۷

باید از جدول‌ها مقادیر میانگین مربعات را انتخاب کنیم و سپس به واریانس تبدیل نماییم:

$$\text{واریانس خطا} = \text{میانگین مربعات خطا} = ۰/۵۶$$

$$\text{واریانس دستیاران} = \frac{\text{میانگین مربعات خطا} - \text{میانگین مربعات دستیاران}}{\text{تعداد آزمونگران}} = \frac{۱۲/۶۷ - ۰/۵۶}{۳} = ۴/۴۰$$

$$\text{واریانس آزمونگران} = \frac{\text{میانگین مربعات خطا} - \text{میانگین مربعات دستیاران}}{\text{تعداد دستیاران}} = \frac{۱۰/۰۰ - ۰/۵۶}{۱۰} = ۰/۹۴$$

سپس با جایگزین کردن مقادیر در فرمول ضریب تعمیم‌پذیری، مقدار آن به صورت زیر به دست می‌آید:

$$G = \frac{۴/۰۳۷}{۴/۰۳۷ + ۰/۵۵۶} = ۰/۸۸$$

همان‌گونه که مشخص است آزمون ضریب تعمیم‌پذیری مقدار خوبی دارد و سهم آزمونگران در ایجاد خطا ۱۷ درصد واریانس کل بوده است که زیاد نیست. یعنی آزمونگران توانسته‌اند ارزیابی خود را با خطای اندکی انجام دهند و نمرات دستیاران قابل اطمینان است.

سهم واریانس هر یک از منابع خطا در جدول ۱۱-۳۴ خلاصه شده است. مشخص است که سهم بزرگی از واریانسی که در نمرات وجود دارد، یعنی بیش از هفتاد درصد آن، مربوط به عملکرد دستیاران است. تفاوت این رویه با بقیه رویه‌ها در این است که واریانس آن قابل قبول است چون تفاوت نمره بین دستیاران مختلف ناشی از خطا نیست و به علت تفاوت واقعی در سطح توانمندی آنها است. اساساً به همین دلیل است که رویه دستیار/دانشجو عموماً به رویه تمیز در نظر گرفته می‌شود. اما همان‌طور که قبلاً گفته شد، آزمونگر رویه تعمیم است زیرا می‌خواهیم مشخص کنیم نمره حاصله تا چه حد به سایر آزمونگرانی که به صورت بالقوه می‌توانستیم داشته باشیم، قابل تعمیم است.

راه ساده‌تر برای انجام محاسبات فوق به جای آنکه ابتدا میانگین مربعات را به دست آوریم و در فرمول جایگزین کنیم، این است که مستقیماً اجزای واریانس را استخراج کنیم. به این صورت که در SPSS نحوه چیدمان داده‌ها را به صورت جدول ۱۲-۳۴ تغییر دهیم. این کار از طریق منوی Data قسمت restructure قابل انجام است. سپس در منوی آنالیز، گزینه General Liner Model و سپس Variance components را انتخاب می‌کنیم. نمره را به عنوان متغیر وابسته وارد می‌کنیم و دستیار و آزمونگر را به عنوان فاکتورهای رندوم. جدولی مشابه جدول ۱۱-۳۴ به صورت مستقیم اجزای واریانس را نشان می‌دهد و نیازی به تبدیل نیست.

جدول ۱۱-۳٤ : مقدار واریانس رویه‌های آزمون شفاهی یک رویه

منبع	مقدار واریانس	درصد واریانس
واریانس دستیار	۴/۰۳۷	۰/۷۳
واریانس آزمونگر	۰/۹۴۴	۰/۱۷
واریانس خطا یا واریانس باقی مانده (دستیار × آزمونگر)	۰/۵۵۶	۰/۱۰

جدول ۱۲-۳٤ : نمرات ۱۰ دستیار در آزمون شفاهی از سه آزمونگر

شماره دانشجو	شماره آزمونگر	نمره
۱	۱	۶
۱	۲	۷
۱	۳	۸
۲	۱	۴
۲	۲	۵
۲	۳	۶
۳	۱	۲
۳	۲	۲
۳	۳	۲
۴	۱	۳
۴	۲	۴
۴	۳	۵
۵	۱	۵
۵	۲	۴
۵	۳	۶
۶	۱	۸
۶	۲	۹
۶	۳	۱۰
۷	۱	۵
۷	۲	۷
۷	۳	۹
۸	۱	۶
۸	۲	۷
۸	۳	۸
۹	۱	۴
۹	۲	۶
۹	۳	۸
۱۰	۱	۷
۱۰	۲	۹
۱۰	۳	۸

محاسبه ضریب تعمیم‌پذیری در حالت دو رویه

اکنون که در حالت یک رویه با مبنای محاسبه ضریب تعمیم‌پذیری آشنا شدیم، حالت پیچیده‌تری را در نظر می‌گیریم. یک OSCE را تصور کنید که با سه ایستگاه و دو آزمونگر در هر ایستگاه برگزار شده است. برای سهولت محاسبات، مدل اثرات ثابت را در نظر می‌گیریم. به عبارت دیگر فرض بر این است که جهان تعمیم و جهان مشاهدات ممکن یکی هستند. در واقع همین سه ایستگاه و همین دو آزمونگر موجود بودند. یا اینکه هدف ما این نیست که نتایج آزمون را به موقعیت دیگر و با موارد بیشتری از ایستگاه‌ها و آزمونگران تعمیم دهیم. واریانس رویه‌ها در این حالت در جدول ۱۳-۳۴ آمده است. در مورد اینکه اعداد مربوط به اجزای واریانس در مثال‌های ذکر شده از کجا به دست آمده‌اند، باید گفت همان طور که قبلاً اشاره شد پایه محاسبات در نظریه تعمیم‌پذیری بر اساس ANOVA است. بنابراین محاسبه اجزای واریانس و سپس ضریب تعمیم‌پذیری از طریق نرم‌افزارهایی مانند SPSS قابل اجرا است اما با توجه به دشواری‌های آن، مخصوصا هنگامی که تعداد رویه‌ها زیاد است، برنامه‌های اختصاصی نظریه تعمیم‌پذیری نیز تدوین شده است که از جمله آنها می‌توان به GENOVA و G string اشاره کرد.

جدول ۱۳-۳٤ : واریانس رویه‌های آزمون OSCE در حالت معیاری (مطلق)

منبع	جمع مربعات	درجه آزادی	میانگین مربعات	مقدار واریانس	درصد واریانس
دانشجو (۱۰ نفر)	۳۹۱۵	۹	۴۳۵	۶۰	۴۰
آزمونگر (۲ نفر)	۸۱۵	۱	۸۱۵	۲۰	۱۳
ایستگاه (۳ ایستگاه)	۹۶۰	۲	۴۸۰	۱۵	۱۰
دانشجو × آزمونگر	۵۸۵	۹	۶۵	۱۵	۱۰
دانشجو × ایستگاه	۵۴۰	۱۸	۳۰	۵	۳
ایستگاه × آزمونگر	۳۴۰	۲	۱۷۰	۱۵	۱۰
خطا یا باقی مانده (دانشجو × آزمونگر × ایستگاه)	۳۶۰	۱۸	۲۰	۲۰	۱۳

فرمولی که برای تعیین ضریب تعمیم‌پذیری به کار می‌رود، به این صورت است (ایستگاه با t نشان داده شده است):

$$G = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_r^2 + \sigma_t^2 + \sigma_{s\times r}^2 + \sigma_{s\times t}^2 + \sigma_{r\times t}^2 + \sigma_{residual}^2}$$

یعنی واریانس تمام منابع به همراه واریانس خطا در مخرج کسر قرار می‌گیرند و طبق فرمول زیر ضریب تعمیم‌پذیری آزمون دست می‌آید:

$$G = \frac{۶۰}{۶۰ + ۲۰ + ۱۵ + ۱۵ + ۵ + ۱۵ + ۲۰} = \frac{۶۰}{۱۵۰} = ۰/۴۰$$

قبلاً اشاره شد که در نظر گرفتن هدف آزمون، یعنی هنجاری یا معیاری بودن آن، در محاسبه ضریب تعمیم‌پذیری مهم است. فرمول فوق مربوط به ضریب مطلق است. از آنجا که در حالت معیارمحور، باید نمره دقیق تک‌تک دانشجویان تعیین شود، در تعیین میزان تعمیم‌پذیری نمرات نیز تمام منابع خطا باید لحاظ شوند. در ضریب نسبی مشخص کردن جایگاه

هر دانشجو نسبت به بقیه کفایت می‌کند و تعیین نمره وی مدنظر نیست. در واقع، تفاوتی که آزمونگران ایجاد می‌کنند، تاثیری در رتبه دانشجویان نسبت به یکدیگر ندارد. به همین دلیل منابع خطای کمتری وارد فرمول می‌شوند یعنی واریانس آزمونگر، واریانس تعامل آزمونگر با دانشجو و واریانس تعامل آزمونگر با ایستگاه حذف می‌شوند. این موضوع باعث می‌شود مقدار عددی ضریب تعمیم‌پذیری مطلق از مقدار عددی ضریب تعمیم‌پذیری نسبی کمتر باشد. واریانس رویه‌ها در حالت هنجاری در جدول ۱۴-۴۳ آمده‌اند.

جدول ۱٤-۳٤ : واریانس رویه‌های آزمون OSCE در حالت هنجاری (نسبی)

منبع	جمع مربعات	درجه آزادی	میانگین مربعات	مقدار واریانس	درصد واریانس
دانشجو (۱۰ نفر)	۳۹۱۵	۹	۴۳۵	۶۰	۵۲
ایستگاه (۳ ایستگاه)	۹۶۰	۲	۴۸۰	۱۵	۱۳
آزمونگر : ایستگاه	۱۱۵۵	۳	۸۱۵	۲۰	۱۷
دانشجو × ایستگاه	۵۴۰	۱۸	۳۰	۵	۴
خطا یا باقی مانده (دانشجو × آزمونگر: ایستگاه)	۹۴۵	۲۷	۱۷۰	۱۵	۱۳

ضریب تعمیم‌پذیری نسبی که با Φ (فی) هم نشان داده می‌شود، با فرمول زیر به دست می‌آید:

$$G_{residual} = \phi = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_t^2 + \sigma_{s\times t}^2 + \sigma_{residual}^2}$$

$$G_{residual} = \phi = \frac{۶۰}{۶۰ + ۱۵ + ۵ + ۲۰} = \frac{۶۰}{۱۰۰}$$

مدل‌های اثرات تصادفی و ثابت

در مثال فوق، تمام رویه‌ها ثابت فرض شده بودند. اکنون حالتی را تصور کنید که ایستگاه رویه ثابت است ولی آزمونگر رویه تصادفی است. مثلاً آزمون در سطح کشوری برگزار می‌شود و آزمونگران متفاوت هستند اما امتحان یکسان است. می‌خواهیم ببینیم نمره دانشجویان چقدر به تمام شرایطی که در آنها از آزمونگران دیگری استفاده شده است، قابل‌تعمیم است. در اینجا رویه ایستگاه را که برای همه ثابت است، از مخرج حذف می‌نماییم:

$$G = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_r^2 + \sigma_{s\times r}^2 + \sigma_{s\times t}^2 + \sigma_{residual}^2} = \frac{۶۰}{۱۰۰} = ۰/۴۴$$

سایر ضرایب تعمیم‌پذیری

فرمول‌های متنوع دیگری برای محاسبه ضریب تعمیم‌پذیری وجود دارد که بسته به هدف آزمون می‌توانند مورد استفاده قرار بگیرند. به عنوان نمونه، اگر هر ایستگاه OSCE دو آزمونگر دارد، نمره دانشجو در ایستگاه با گرفتن میانگین

دو نمره حساب می‌شود و در نهایت، نمره کلی دانشجو با گرفتن میانگین نمرات تمام ایستگاه‌ها به دست می‌آید. در اینجا فرمولی برای محاسبه ضریب تعمیم‌پذیری میانگین لازم است. همچنین این مثال‌ها برای حالتی بود که رویه‌ها متقاطع[1] هستند. اما تصور کنید که آزمونگران در ایستگاه‌ها به صورت آشیانه‌ای[2] قرار گرفته باشند. در این حالت، جدول اجزای واریانس به صورت دیگری خواهد بود که از حوصله این کتاب خارج است.

G study و D study

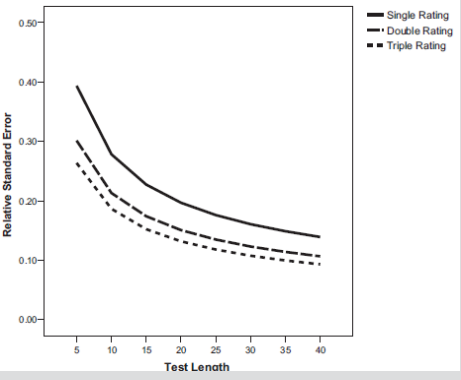
تخمین سهم منابع مختلف خطا در آزمون و تعیین میزان تعمیم‌پذیری نتایج آزمون، یعنی کاری که تا بدینجا انجام دادیم، در واقع به مطالعه تعمیم‌پذیری[3] معروف است اما از قابلیتهای نظریه تعمیم‌پذیری، امکان بررسی پایایی آزمون در شرایط مختلف فرضی است. به آنالیزی که برای بررسی تاثیر شرایط مختلف استفاده میشوند، مطالعه تصمیم‌گیری[4] میگویند. در D study تلاش می‌شود تا با استفاده از یافته‌های G study طرح‌های دیگری برای آزمون پیشنهاد شود که مقدار خطای آزمون به حداقل برسد و تعمیم‌پذیری آزمون افزایش یابد. موارد زیر مثال‌هایی از کاربرد نظریه تعمیم‌پذیری را نشان می‌دهند.

الوس د لیما و همکاران [1] ۲۰۰۷

در این مطالعه مشخصات ابزار mini-CEX در ۱۷ مرکز کاردیولوژی آرژانتین مورد سنجش قرار گرفت. پایایی ابزار به کمک نظریه تعمیم‌پذیری محاسبه شد و نشان داده شد که برای رسیدن به پایایی قابل قبول لازم است هر دستیار حداقل ۱۰ بار مورد ارزیابی قرار گیرد.

کلازر و همکاران [2] ۲۰۰۸

این مطالعه با استفاده از داده‌های OSCE در مرحله دوم USMLE صورت گرفت که در آن بخشی از نمره هر ایستگاه مربوط به توانمندی دانشجویان در مکتوب‌سازی یافته‌ها، تشخیص و ارائه طرح درمان است. در حالت معمول نمره‌دهی به پاسخ [2] دانشجویان توسط یک مصحح و بر مبنای مقیاس ۱ تا ۹ انجام می‌شود. در این مطالعه، هر پاسخ توسط دو مصحح نمره‌دهی شد و از میانگین آنها نمره نهایی به دست آمد. یافته‌ها نشان داد که واریانس مربوط به ایستگاه‌ها سهم بیشتری نسبت به واریانس مربوط به مصحح دارد و همچنین تصحیح توسط دو آزمونگر ضریب تعمیم‌پذیری را افزایش می‌دهد. پژوهشگران همچنین با D study نشان دادند افزایش تعداد مصححان از یک نفر به دو نفر بر دقت اندازه‌گیری بیشتر موثر است تا افزایش تعداد ایستگاه‌ها. در عین حال استفاده از سه مصحح تغییر چندان زیادی در میزان خطا نمی‌دهد. شکل زیر نمودار خطای معیار اندازه‌گیری آزمون بر حسب تعداد آزمونگران را نشان می‌دهد.



1. Alves de Lima et al
2 Clauser et al

1. Crossed
2. Nested
3. Generalizability study (G study)
4. Decision study (D study)

بالدوین و همکاران [1] ۲۰۰۹

این پژوهش نیز روی امتحان OSCE در مرحله دوم USMLE انجام شد. همان طور که ذکر شد در حالت معمول، پاسخ دانشجویان توسط یک مصحح و بر مبنای مقیاس ۱ تا ۹ تصحیح می‌شود. در سال ۲۰۰۷ مداخله‌ای به این صورت انجام شد که روبریکی برای تصحیح در اختیار مصححان قرار گرفت. در سال ۲۰۰۸ مداخله دیگری در پیش گرفته شد به این ترتیب که روبریک اختصاصی برای هر ایستگاه تدوین شد.

پایایی آزمون توسط نظریه تعمیم‌پذیری محاسبه شد. یافته‌ها نشان داد روبریک و آموزش مصححان موجب بهبود پایایی می‌شود.

نکته قابل توجه این است که طی این سه سال، واریانس مربوط به دانشجویان زیاد شده است. همان‌طور که قبلاً گفته شد، واریانس دانشجویان، خطا محسوب نمی‌شود و ناشی از تفاوت واقعی بین سطح دانشجویان است اما نویسندگان علت افزایش آن را به استفاده از روبریک مرتبط دانسته‌اند. به این دلیل که تا قبل از آن ارزیابی مصححان به سمت وسط مقیاس متمایل بود اما با استفاده از روبریک منطقی‌تر شد و نمرات به حالت واقعی نزدیک شدند.

1. A Baldwin et al.

منابع

1. Baldwin SG, Harik P, Keller LA, Clauser BE, Baldwin P, Rebbecchi TA. Assessing the impact of modifications to the documentation component's scoring rubric and rater training on USMLE integrated clinical encounter scores. Acad Med 2009;84(10 Suppl):S97-100.
2. Clauser BE, Harik P, Margolis MJ, Mee J, Swygert K, Rebbecchi T. The generalizability of documentation scores from the USMLE Step 2 Clinical Skills examination. Acad Med 2008;83(10 Suppl):S41-4.
3. Alves de Lima A, Barrero C, Baratta S, et al. Validity, reliability, feasibility and satisfaction of the Mini-Clinical Evaluation Exercise (Mini-CEX) for cardiology residency training. Med Teach 2007;29(8):785-90.
4. Brennan R. 2001. Generalizability theory. New York: Springer Verlag.
5. Bloch R, Norman G. Generalizability theory for the perplexed: a practical introduction and guide: AMEE Guide No. 68. Med Teach 2012;34(11):960-92.
6. Schoonheim-Klein M, Muijtjens A, Habets L, et al. On the reliability of a dental OSCE, using SEM: effect of different days. Eur J Dent Educ 2008;12(3):131-7.
7. Streiner DL, Norman GR. Health Measurement Scales: A practical guide to their development and use. Oxford University Press, USA; 4 edition, 2008
8. Tavakol M, Dennick R. Post-examination interpretation of objective test data: monitoring and improving the quality of high-stakes examinations: AMEE Guide No. 66. Med Teach 2012;34(3):e161-75.

فصل ۳۵

تحلیل آزمون در نظریه سؤال پاسخ

الگوی پاسخ به سؤال

در این فصل کتاب فقط تحلیل سؤالات چندگزینه‌ای بر مبنای نظریه سؤال پاسخ مطرح می‌شود. در حالی که این نظریه برای تحلیل سایر انواع آزمون‌ها از جمله OSCE نیز به کار می‌رود. همچنین از آنجا که برای درک راحت‌تر موضوع، این بحث بسیار ساده شده و کاربردی ارائه می‌شود، برای مفاهیم تخصصی‌تر به متون اصلی مراجعه کنید.

برای آغاز بحث در مورد محاسبه شاخص‌ها بر اساس نظریه سؤال پاسخ به ذکر یک مثال می‌پردازیم. تصور کنید پنج دانشجو در یک آزمون با پنج سؤال چهارگزینه‌ای شرکت کرده‌اند و الگوی پاسخ‌دهی آنها به صورت جدول ۱-۳۵ است:

جدول ۱-۳۵ : توزیع پاسخ‌های پنج دانشجو به پنج سؤال از یک آزمون چهارگزینه‌ای

	سؤال ۱	سؤال ۲	سؤال ۳	سؤال ۴	سؤال ۵	میانگین
دانشجوی ۱	۱	۱	۱	۱	۱	۱
دانشجوی ۲	۰	۱	۱	۱	۱	۰/۸
دانشجوی ۳	۰	۰	۱	۱	۱	۰/۶
دانشجوی ۴	۰	۰	۰	۱	۱	۰/۴
دانشجوی ۵	۰	۰	۰	۰	۱	۰/۲
دشواری سؤال	۰/۸	۰/۶	۰/۴	۰/۲	۰	

در این مثال، دانشجوی اول را که به تمام سؤالات به درستی پاسخ داده است، به صورت «فرضی»[1] واجد توانایی صد درصد در نظر می‌گیریم. دانشجوی شماره دو، هشتاد درصد توانمندی مورد نظر را دارد و همین طور توانمندی دانشجوی پنج، بیست درصد در نظر گرفته می‌شود.

همین مسأله را می‌توان برای تک تک آیتم‌ها نیز محاسبه کرد. مشخص است که سؤال یک سخت‌ترین سؤال بوده است. زیرا فقط یک نفر آن هم با بالاترین سطح توانمندی توانسته است به آن پاسخ دهد. به عبارت دیگر، هشتاد درصد دانشجویان نتوانسته‌اند به این سؤال جواب درست دهند. هرچه به سؤال آخر نزدیک می‌شویم، سؤالات به صورت «فرضی» آسان‌تر به نظر

1. Tentative

می‌رسند. سؤال پنج آسان‌ترین سؤال بوده است و همه دانشجویان از آن نمره گرفته‌اند یعنی دشواری آن صفر بوده است (توجه کنید مفهومی که اینجا برای دشواری به کار می‌رود، متضاد مفهوم ضریب دشواری در نظریه کلاسیک است).

همچنین مشخص است که دانشجویان ضعیف‌تر تعداد سؤالات کمتری را پاسخ داده‌اند و سؤالاتی را پاسخ داده‌اند که آسان‌تر هستند. به این الگوی پاسخ‌دهی ایده‌آل که در آن دانشجویان به ترتیب از قوی تا ضعیف و سؤالات به ترتیب از سخت به آسان مرتب شده‌اند، الگوی گاتمن[1] گفته می‌شود. این الگو در واقعیت کم اتفاق می‌افتد و به اصطلاح overfit است. مسأله این است که اگر نمره کل دانشجویان را بر اساس جمع نمره آنها در تک‌تک سؤالات حساب کنیم، احتمالاً برآورد کاملاً صحیحی از میزان توانایی آنها به دست نیاورده‌ایم. برای درک بهتر موضوع، وضعیت دانشجوی شماره شش را در جدول ۲–۳۵ در نظر بگیرید.

جدول ۲–۳۵ : توزیع پاسخ‌های شش دانشجو به پنج سؤال از یک آزمون چهارگزینه‌ای

	سؤال ۱	سؤال ۲	سؤال ۳	سؤال ۴	سؤال ۵	میانگین
دانشجوی ۱	۱	۱	۱	۱	۱	۱
دانشجوی ۲	۰	۱	۱	۱	۱	۰/۸
دانشجوی ۳	۰	۰	۱	۱	۱	۰/۶
دانشجوی ۴	۰	۰	۰	۱	۱	۰/۴
دانشجوی ۵	۰	۰	۰	۰	۱	۰/۲
دانشجوی ۶	۱	۱	۰	۰	۰	۰/۴

همان‌طور که مشاهده می‌شود، دانشجوی شماره شش هم مانند دانشجوی شماره چهار تنها به دو سؤال پاسخ داده است و توانمندی هر دوی آنها ۴۰ درصد گزارش شده است. در واقع اگر به صورت معمول و بر مبنای نظریه کلاسیک آزمون نمرات را محاسبه کنیم، نمره این دو دانشجو با هم برابر می‌شود. اما واقعیت این است که دانشجوی شماره چهار به دو سؤال آسان پاسخ داده است و دانشجوی شماره شش توانسته است جواب دو سؤال سخت را به درستی بدهد. بنابراین نمی‌توان سطح دو دانشجو را یکسان فرض کرد و این نقص نظریه کلاسیک آزمون است. در ادامه خواهیم دید که نظریه سؤال پاسخ چگونه با این موضوع برخورد می‌کند. همین مسأله ممکن است در مورد سؤال نیز پیش آید. نحوه پاسخ‌دهی دانشجویان به سؤال شش را در جدول ۳–۳۵ مشاهده می‌کنیم.

جدول ۳–۳۵ : توزیع پاسخ‌های پنج دانشجو به شش سؤال از یک آزمون چهارگزینه‌ای

	سؤال ۱	سؤال ۲	سؤال ۳	سؤال ۴	سؤال ۵	سؤال ۶	میانگین
دانشجوی ۱	۱	۱	۱	۱	۱	۰	۰/۸۳
دانشجوی ۲	۰	۱	۱	۱	۱	۰	۰/۶۷
دانشجوی ۳	۰	۰	۱	۱	۱	۰	۰/۵۰
دانشجوی ۴	۰	۰	۰	۱	۱	۰	۰/۳۳
دانشجوی ۵	۰	۰	۰	۰	۱	۱	۰/۳۳
دشواری سؤال	۰/۸	۰/۶	۰/۴	۰/۲	۰	۰/۸	

1. Guttman pattern

در این مثال، هر یک از دو سؤال یک و شش تنها توسط یک دانشجو پاسخ داده شده‌اند و سطح دشواری یکسانی معادل ۰/۸ دارند اما با توجه به اینکه سؤال یک توسط دانشجو با سطح توانمندی ۰/۸۳ و سؤال ۶ توسط دانشجو با سطح توانمندی ۰/۳۳ پاسخ داده شده‌اند، در واقع باید پذیرفت که سطح دو سؤال یکسان نیست.

تا اینجا مشابه محاسباتی بود که در نظریه کلاسیک آزمون برای تحلیل سؤال انجام می‌شد که به نظر می‌رسد هرچند ساده و قابل درک است، اما کاملاً درست نیست. برای اینکه برآورد مناسبی از ویژگی سؤال داشته باشیم، باید سطح توانمندی دانشجو و سطح دشواری سؤال را در تعامل با یکدیگر در نظر بگیریم و نه به صورت منفک و جدا. این کار از طریق نظریه سؤال پاسخ انجام می‌شود. به این منظور از این پس سطح توانایی دانشجو را با θ و سطح دشواری سؤال را با D نشان می‌دهیم. توجه کنید که از این پس دیگر آن حالت «فرضی» مدنظر نیست.

به صورت نظری میزان توانایی افراد یعنی θ می‌تواند از منفی بی‌نهایت تا مثبت بی‌نهایت باشد اما به دلیل محدویت‌های اجرایی معمولاً آن را به صورت ۳- تا ۳+ (یا ۴- تا ۴+ یا ۵- تا ۵+) بیان می‌کنند. بنابراین θ دارای صفر قراردادی و نوع مقیاس آن فاصله‌ای[1] است.

دانشجویان که سطوح توانایی مختلفی دارند، در برابر یک سؤال مشخص عملکرد گوناگونی دارند. به عبارت دیگر در هر سؤال، برای هر سطحی از θ می‌توان احتمال پاسخگویی در نظر گرفت که آن را با $p(\theta)$ نشان می‌دهند و به صورت زیر محاسبه می‌شود:

$$\text{Probabability} = \frac{۱}{۱+\frac{۱}{\exp^{(\theta-D)}}}$$

می‌دانیم که:

$$\exp^{0} = ۱$$

$$\exp^{1} = ۲/۷۱۸۲$$

$$\exp^{2} = ۷/۳۸۹۰$$

$$\exp^{3} = ۲۰/۰۸۵۵$$

با برگشت به جدول ۱-۳۵ دانشجو و با استفاده از فرمول فوق می‌توانیم احتمال پاسخگویی هر سؤال را برای هر دانشجو حساب کنیم. به عنوان مثال، احتمال پاسخگویی دانشجوی شماره یک (با θ معادل ۱) به سؤال پنج با d معادل صفر (چون هیچ‌کس به آن جواب نداده است)، به صورت زیر است:

$$\text{Probabability} = \frac{۱}{۱+\frac{۱}{\exp^{(۱-۰)}}} = \frac{۱}{۱+\frac{۱}{۲/۷}} = \frac{۱}{۱/۳۷} = ۰/۷۳$$

این مسأله به این معنا است که دانشجوی یک با توانایی معادل یک ۷۳ درصد احتمال دارد که بتواند به سؤال پاسخ بگوید. یعنی یا سؤال آسان است و یا دانشجو قوی است. در هر حال، سطح توانایی این دانشجو قطعا بالاتر از سطح این سؤال است. در مثال دیگر احتمال پاسخگویی دانشجوی دوم به سؤال اول را در نظر می گیریم:

1. Interval

$$\text{Probabability} = \frac{۱}{۱+\frac{۱}{\exp^{(۰/۸-۰/۸)}}} = \frac{۱}{۱+\frac{۱}{۱}} = ۰/۵$$

مشخص است که سطح توانایی دانشجو با سطح دشواری سؤال یکسان است. هنگامی که دانشجو ۵۰ درصد شانس این را دارد که به سؤال درست جواب دهد، یعنی نسبت به سؤال برتری ندارد و برعکس. با محاسبه تمام احتمالات به جدول ۴-۳۵ می‌رسیم. همان طور که مشخص است، احتمال پاسخگویی پنج دانشجو به پنج سؤال به صورت جداگانه محاسبه شده است که مجموعاً ۲۵ حالت را ایجاد کرده است:

جدول ٤-۳۵ : احتمال پاسخگویی هر دانشجو به هر یک از سؤالات

	سؤال ۱	سؤال ۲	سؤال ۳	سؤال ۴	سؤال ۵
دانشجوی ۱	۰/۵۵	۰/۶۰	۰/۶۵	۰/۶۹	۰/۷۳
دانشجوی ۲	۰/۵۰	۰/۵۵	۰/۶۰	۰/۶۵	۰/۶۹
دانشجوی ۳	۰/۴۵	۰/۵۰	۰/۵۵	۰/۶۰	۰/۶۵
دانشجوی ۴	۰/۴۰	۰/۴۵	۰/۵۰	۰/۵۵	۰/۶۰
دانشجوی ۵	۰/۳۵	۰/۴۰	۰/۴۵	۰/۵۰	۰/۵۵

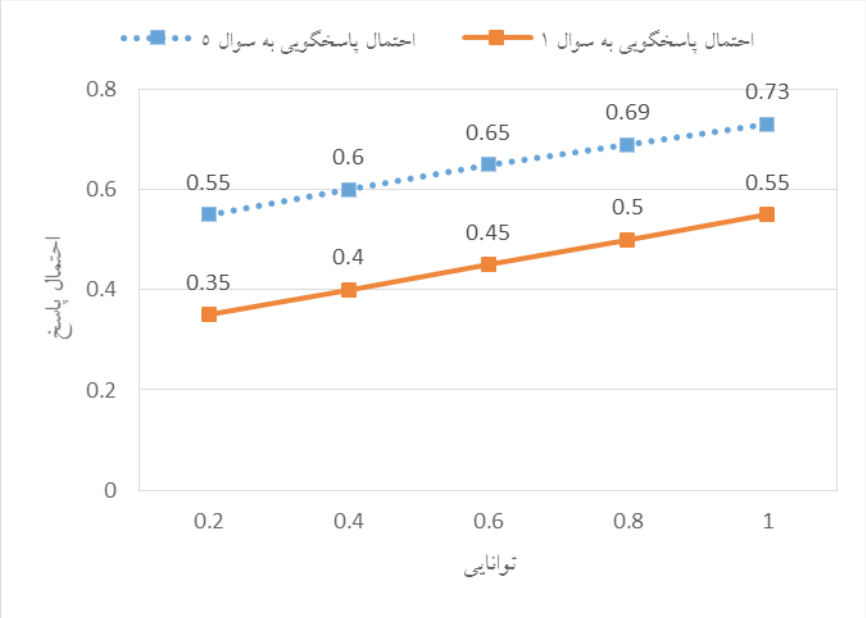
اکنون دانشجوی شماره پنج را در نظر بگیرید. احتمال پاسخگویی به سؤالات یک تا چهار برای این دانشجو از ۰/۳۵ تا ۰/۵۰ متغیر است اما در جدول ۱-۳۵ دیدیم که او در واقع به هیچ کدام از این چهار سؤال جواب نداده بود. از این پدیده تحت عنوان عدم مطابقت داده و مدل نام می‌برند. داده‌ها و مدل ممکن است دقیقاً بر هم منطبق نشوند. به عبارت دیگر، اگرچه احتمال می‌دهیم و انتظار داریم که این دانشجو به این چهار سؤال پاسخ درست بدهد، او به هیچ کدام از این سؤالات پاسخ درست نداده است.

خم ویژه سؤال

بعد از اینکه از طریق معادله مذکور احتمال پاسخگویی به یک سؤال توسط دانشجویان با سطوح مختلف توانایی مشخص شد، می‌توان رابطه بین این دو را در نمودار نشان داد. اگر بخواهیم از روی داده‌های جدول ۴-۳۵ احتمال پاسخگویی پنج دانشجو را به سؤال یک و پنج روی نمودار رسم کنیم، شکلی مشابه شکل ۱-۳۵ خواهیم داشت.

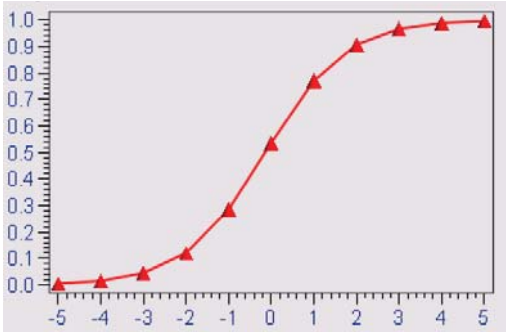
توجه داشته باشید که در اینجا نمودار تنها بر اساس پاسخگویی پنج دانشجو رسم شده است که تعداد کمی است و بنابراین نمودار دقیقی به دست نمی‌آید. همچنین سطوح توانایی مورد استفاده محدود هستند. مثلاً از آنجا که دانشجویی با سطح توانایی صفر یا منفی در نظر گرفته نشده است، نمی‌توان مشخص کرد که احتمال پاسخگویی آنها چقدر بوده است. در عمل تعداد دانشجویان بیشتر از این است و طیف متنوع‌تری را (به عنوان مثال از ۳- تا ۳+ به جای ۰/۲ تا ۱) شامل می‌شوند. افزودن این اطلاعات می‌تواند به دقیق‌تر شدن شکل کمک کند. به این نمودار که در واقع برای هر سؤال به صورت جداگانه رسم می‌شود و میزان احتمال پاسخگویی به سؤال را برای سطوح مختلف توانایی توانایی نشان می‌دهد، خم ویژه سؤال[1] می‌گویند.

1. Item Characteristic Curve (ICC)



شکل ۱-۳۵: احتمال پاسخگویی دانشجویان به دو سؤال به تفکیک سطوح توانایی

خم ویژه سؤال شکلی s مانند دارد و محور افقی آن میزان توانایی دانشجویان مختلف را نشان می‌دهد که از منفی پنج تا مثبت پنج در نظر گرفته شده است (شکل ۲-۳۵). توجه داشته باشید که این شکل فرضی است و نه بر اساس داده‌های واقعی. در عمل ممکن است هیچ دانشجویی با سطح توانایی منفی پنج وجود نداشته باشد اما به صورت فرضی و برای بررسی عملکرد سؤال، می‌خواهیم ببینیم اگر چنین دانشجویی وجود داشت، احتمال پاسخگویی او به این سؤال چقدر بود. در شکل می‌بینیم که این احتمال نزدیک به صفر است. برای دانشجو با سطح توانایی صفر که به عبارت دیگر دانشجوی متوسط است، احتمال پاسخگویی به این سؤال ۰/۵ است.



شکل ۲-۳۵: خم ویژه سؤال

پارامترهای سؤال

پارامتر دشواری: پارامتر دشواری که به آن پارامتر آستانه[1] نیز گفته می‌شود و با b نشان داده می‌شود، مشخص می‌کند که یک سؤال چقدر آسان یا چقدر سخت است. شکل ۳–۳۵ خم ویژه چندین سؤال را در مدل یک پارامتری (به زودی توضیح داده می‌شود که مدل چیست) نشان می‌دهد. دقت کنید که در این مدل خم هیچ یک از سؤالات خم سؤال دیگری را قطع نمی‌کند. اگر جایگاه ICC به سمت راست نمودار میل کند به این معنا است که سؤال دشوار است زیرا احتمال پاسخگویی به آن برای اکثر دانشجویان کم است. ولی اگر ICC در سمت چپ قرار گیرد، به این معنا است که دانشجویان با سطوح توانایی پایین هم به آن پاسخ گفته‌اند. در شکل ۴–۳۵ سه سؤال با دشواری‌های مختلف نشان داده شده‌اند.

به صورت کلی دشواری هر سؤال برابر نقطه‌ای روی محور توانایی تعریف می‌شود که در آن احتمال پاسخگویی ۰/۵ باشد. یعنی اگر خطی به موازات ۰/۵ کشیده شود، محل تقاطع آن با خم ویژه سؤال مقدار پارامتر دشواری سؤال خواهد بود (شکل ۴–۳۵).

پارامتر تمیز: این شاخص که با a نشان داده می‌شود، مشخص می‌کند که یک سؤال در افتراق دانشجویان قوی از ضعیف تا چه حد موثر است. مدل یک پارامتری تنها پارامتر دشواری را لحاظ می‌کند و پارامتر تمیز را ثابت در نظر می‌گیرد. مدل دوپارامتری هم شاخص دشواری و هم شاخص تمیز را در نظر می‌گیرد. شکل ۵–۳۵ نمونه یک مدل دوپارامتری را برای چندین سؤال نشان می‌دهد. همان‌طور که مشاهده می‌کنید، مشخصه آن تقاطع خم‌های برخی از سؤالات با یکدیگر است.

هرچه شیب خم ویژه سؤال بیشتر باشد، به این معناست که قابلیت تمیز بیشتری دارد. در شکل ۶–۳۵ سؤال یک که تقریباً تخت است، قدرت تمیز زیادی ندارد. احتمال اینکه دانشجو با توانایی مثبت پنج به آن پاسخ درست دهد، ۰/۸۲ است و احتمال اینکه دانشجو با سطح توانایی منفی پنج به آن درست پاسخ دهد، ۰/۴۸ است. یعنی تفاوت بین آنها تنها ۰/۲۶ است. در مورد سؤال شماره دو احتمال پاسخگویی برای قوی‌ترین و ضعیف‌ترین دانشجو به ترتیب برابر یک و صفر است. تفاوت بین این دو مقدار نشان می‌دهد که قدرت تمیز این سؤال بالا است.

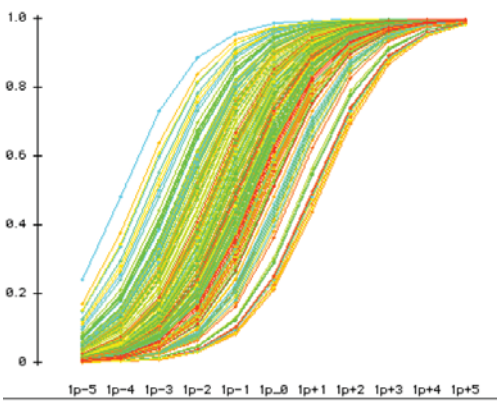
اگر قدرت تمیز سؤالی در حداکثر مقدار ممکن باشد، ICC آن خط مستقیمی است که بر محور افقی عمود شده است. در شکل ۷–۳۵، احتمال پاسخگویی برای دانشجویانی که در سمت راست خط قرار گرفته‌اند، یک و برای دانشجویانی که در سمت چپ خط قرار گرفته‌اند، صفر است. این سؤال در سایر مقادیر θ نمی‌تواند بین دانشجویان افتراق قائل شود.

برعکس، اگر سؤال فاقد قدرت تشخیص باشد، مقدار دشواری آن نامشخص است. به همین دلیل برای همه سطوح θ، احتمال پاسخگویی ۰/۵ در نظر گرفته می‌شود. بنابراین ICC به صورت یک خط افقی متناظر با ۰/۵ روی محور عمودی خواهد بود. مسأله دیگر اینکه همان طور که در نظریه کلاسیک آزمون گفته شد، ضریب تمیز سؤال ممکن است منفی شود. همان دلایل در اینجا نیز صدق می‌کند و سؤال باید مورد بررسی قرار گیرد تا مشکل آن مشخص گردد.

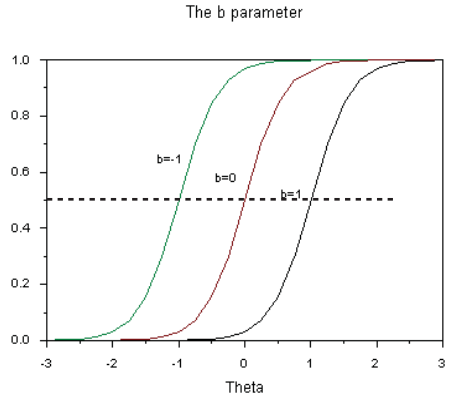
پارامتر حدس: پارامتر بعدی که c یا g خوانده می‌شود، پارامتر حدس زدن[2] است و مشخص می‌کند که چقدر احتمال دارد دانشجو پاسخ صحیح را با حدس انتخاب کند. مقدار حدس زدن وابسته به توانایی نیست. یعنی برای همه دانشجویان با هر سطحی از θ یکسان در نظر گرفته می‌شود. از همین رو است که حتی دانشجویی با سطح توانایی منفی پنج می‌تواند پاسخ برخی از سؤالات را به صورت شانسی درست انتخاب کند. به صورت نظری، مقدار حدس می‌تواند عددی بین صفر تا یک باشد اما به صورت معمول میزان آن بین ۰ تا ۰/۳۵ در نظر گرفته می‌شود.

مدل سه پارامتری شامل هر سه پارامتر A، B و C است. شکل ۸–۳۵ مدل سه پارامتری همان داده‌هایی را نشان می‌دهد که در شکل ۵–۳۵ مدل دوپارامتری آنها رسم شده بود. مشخص است که این دو نمودار بسیار شبیه یکدیگر هستند. با وجود این تفاوت جزئی بین آنها وجود دارد. در شکل ۸–۳۵، نقطه آغازین اغلب آیتم‌ها در محور عمودی (که به آن عرض از مبدا[3] گفته می‌شود) بالاتر از صفر است در حالی‌که در شکل ۵–۳۵ تمام سؤالات از نقطه صفر محور عمودی شروع می‌شوند.

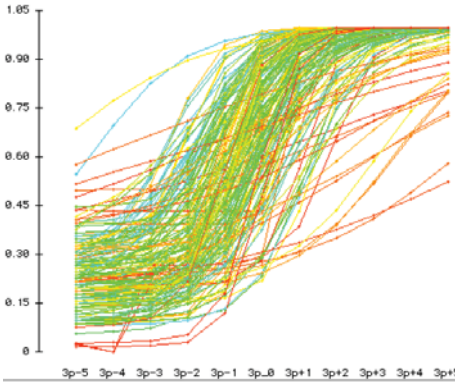
1. Threshold parameter
2. Guessing
3. Intercept



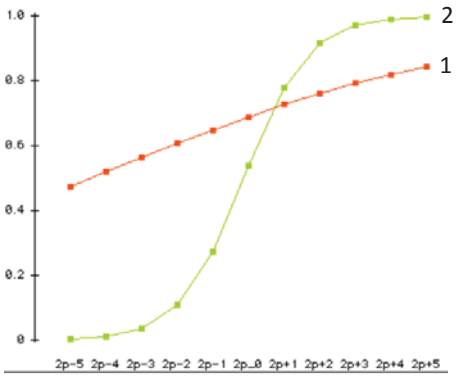
شکل ۳-۳۵: خم ویژه چند سؤال در مدل یک پارامتری



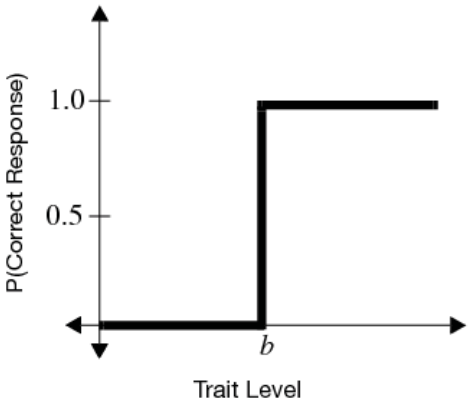
شکل ۴-۳۵: خم ویژه سه سؤال با دشواری متفاوت



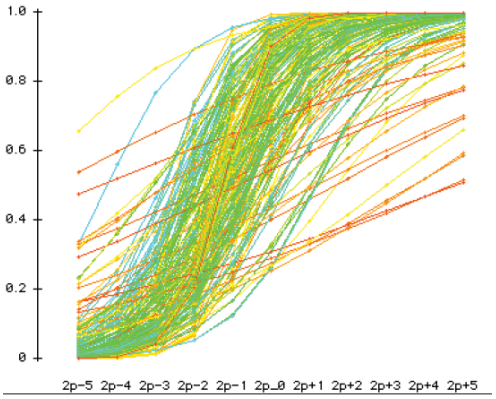
شکل ۵-۳۵: خم ویژه چند سؤال در مدل دوپارامتری



شکل ۶-۳۵: خم ویژه دو سؤال با تمیز و دشواری متفاوت در مدل دوپارامتری



شکل ۷-۳۵: خم ویژه سؤال با قدرت تمیز حداکثری در نقطه b



شکل ۸-۳۵: خم ویژه چند سؤال در مدل سه پارامتری

نکته مهم اینکه برآورد میزان دشواری سؤال در مدل سه پارامتری مانند تعریف قبل نیست و باید مقدار c به صورت زیر در آن لحاظ شود:

$$\text{Probabability} = \frac{۱+C}{۲}$$

مدل‌های نظریه سؤال پاسخ

همان‌طور که گفته شد، مدل‌های IRT می‌توانند یک پارامتری (1PL)، دوپارامتری (2PL) یا سه پارامتری (3PL) باشند. فرمولی که قبلاً در ابتدای این فصل ذکر شد، تنها واجد پارامتر دشواری بود. معادله مدل دوپارامتری که پارامتر تمیز را نیز در نظر می‌گیرد، به این صورت است:

$$\text{Probabability} = \frac{۱}{۱+\frac{۱}{\exp^{a(\theta-b)}}}$$

فرمول مدل سه پارامتری که پارامترهای دشواری، تمیز و حدس را لحاظ می‌کند، به صورت زیر است:

$$\text{Probabability} = c + \frac{۱-c}{۱+\frac{۱}{\exp^{a(\theta-b)}}}$$

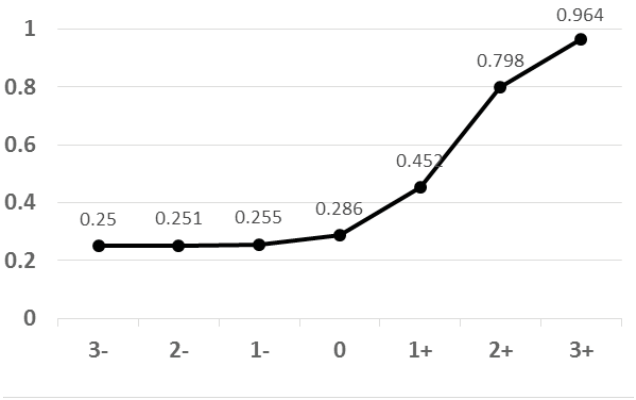
به عنوان مثال، سؤالی را در نظر می‌گیریم که پارامترهای دشواری، تمیز و حدس آن به ترتیب مساوی ۱/۵+، ۲+ و ۰/۲۵ است. طبق فرمول ذکر شده، احتمال پاسخگویی برای دانشجو با توانایی ۲+ به این صورت است:

$$\text{Probabability} = ۰/۲۵ + \frac{۱-۰/۲۵}{۱+\frac{۱}{\exp^{۲(۲-۱/۵)}}} = ۰/۲۵ \frac{۰/۷۵}{۱+\frac{۱}{۲/۷۱}} = ۰/۷۹$$

احتمال پاسخگویی به این سؤال را برای بقیه سطوح دانشجویان به همان ترتیب محاسبه می‌کنیم (جدول ۵–۳۵):
اکنون که توانایی و احتمال پاسخ را داریم، می‌توانیم خم ویژه این سؤال را رسم کنیم (شکل ۹–۳۵):

جدول ۵–۳۵ : احتمال پاسخگویی به یک سؤال در سطوح مختلف توانایی

توانایی (θ)	-۳	-۲	-۱	۰	+۱	+۲	+۳
$a(\theta-b)$	-۹	-۷	-۵	-۳	-۱	۱	۳
$\exp^{a(\theta-b)}$	۸۱۰۴/۰۸	۱۰۹۷/۶۳	۹۴۱/۱۴	۱۲/۰۸	۳/۷۱	۱/۳۶	۱/۰۵
احتمال (p)	۰/۲۵۰	۰/۲۵۱	۰/۵۵۲	۰/۲۸۶	۰/۴۵۲	۰/۷۹۸	۰/۴۶۹



شکل ۹-۳۵: خم ویژه سؤال

نمره کل و خم ویژه آزمون

تا اینجا در مورد یک سؤال صحبت شد اما در واقع هر آزمون از چندین سؤال تشکیل شده است و لازم است محاسبه کنیم که در مجموع احتمال پاسخگویی به همه سؤالات چقدر است. نکته قابل توجه اینکه در نظریه سؤال پاسخ احتمال پاسخگویی به کل آزمون برای هر یک از سطوح توانایی‌ها به صورت جداگانه محاسبه می‌شود.

برای این کار ابتدا باید احتمال پاسخگویی به تک‌تک سؤالات را در همان سطح توانایی حساب کنیم و بعد اعداد به دست آمده را با هم جمع کنیم. در مثال زیر آزمونی در نظر گرفته شده است که دارای چهار سؤال است (جدول ۶-۳۵):

جدول ۶-۳۵ : پارامتر دشواری و تمیز در چهار سؤال

سؤال	پارامتر دشواری	پارامتر تمیز
۱	۰/۵	۱
۲	۲	۲
۳	-۱	۰
۴	-۱/۵	۱

با در نظر گرفتن مدل دوپارامتری، احتمال پاسخگویی به هر سؤال را برای دانشجویان با $\theta=1$ حساب می‌کنیم.

$$\text{Probabability}_{(\theta=1)_1} = \frac{1}{1+\frac{1}{\exp^{1(1-0/6)}}} = \frac{1}{1+0/60} = 0/625$$

$$\text{Probabability}_{(\theta=1)_2} = \frac{1}{1+\frac{1}{\exp^{2(1-2)}}} = \frac{1}{1+7/39} = 0/119$$

$$\text{Probabability} = (\theta=۱)_{۳} \frac{۱}{۱+\frac{۱}{\exp^{۰(۱-(-۱))}}} = \frac{۱}{۱+۱} = ۰/۵$$

$$\text{Probabability} = (\theta=۱)_{۳} \frac{۱}{۱+\frac{۱}{\exp^{۱(۱-(-۱/۵))}}} = \frac{۱}{۱+۰/۰۸} = ۰/۹۲۴$$

همین عملیات را برای سایر سطوح توانایی‌ها تکرار می‌کنیم تا احتمال پاسخگویی به هر سؤال در توانایی‌های مختلف به دست آید. همچنین برای یافتن نمره کل آزمون در هر یک از سطوح توانایی، احتمالات مربوط به چهار سؤال را در آن سطح توانایی با هم جمع می‌کنیم (جدول ۷-۳۵). به عنوان مثال، در سطح توانایی یک، نمره کل آزمون به این صورت است:

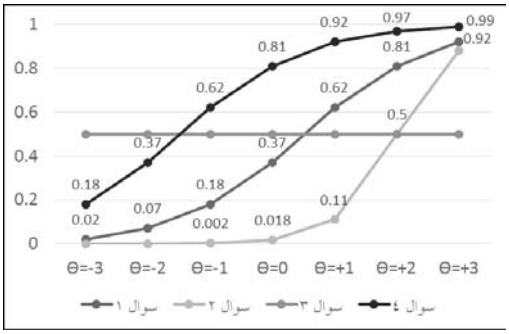
$$\text{Score } (\theta=1)_{test} = ۰/۶۲۵+۰/۱۱۹+۰/۵+۰/۹۲۴= ۲/۱۶۸$$

جدول ۷-۳۵ : احتمال پاسخگویی به هر سوال و کل آزمون در هر یک از سطح توانمندی

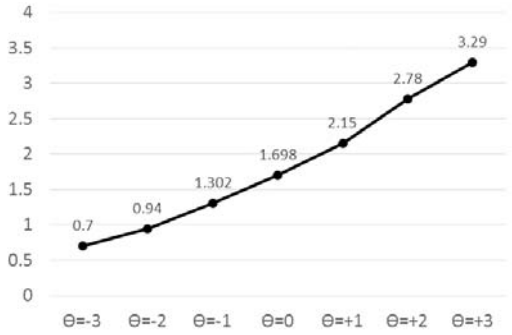
توانایی	سؤال ۱	سؤال ۲	سؤال ۳	سؤال ۴	کل آزمون
θ=-۳	۰/۰۲	۰/۰۰۰	۰/۵۰	۰/۱۸	۰/۷۰
θ=-۲	۰/۰۷	۰/۰۰۰	۰/۵۰	۰/۳۷	۰/۹۴
θ=-۱	۰/۱۸	۰/۰۰۲	۰/۵۰	۰/۶۲	۱/۳۰
θ=۰	۰/۳۷	۰/۰۱۸	۰/۵۰	۰/۸۱	۱/۶۹
θ=+۱	۰/۶۲	۰/۱۱	۰/۵۰	۰/۹۲	۲/۱۵
θ=+۲	۰/۸۱	۰/۵۰	۰/۵۰	۰/۹۷	۲/۷۸
θ=+۳	۰/۹۲	۰/۸۸	۰/۵۰	۰/۹۹	۳/۲۹

با توجه به اطلاعات فوق، خم ویژه هر یک از سؤالات در شکل ۱۰-۳۵ رسم شده است:

برای رسم خم ویژه آزمون باید نموداری رسم کنیم که محور عمودی آن نمره در کل آزمون و محور افقی آن سطوح توانایی است (شکل ۱۱-۳۵.):



شکل ۱۰-۳۵: خم ویژه چهار سؤال



شکل ۱۱-۳۵: خم ویژه آزمون

برآورد توانایی هر یک از دانشجویان

در عمل لازم است که پس از آزمون نمره هر فرد یعنی توانایی تک‌تک دانشجویان شرکت‌کننده در آزمون را به دست آوریم. برای این کار لازم است که مقدار عددی پارامترهای مربوط به سؤال مشخص باشند. سپس به کمک آنها و پاسخ‌هایی که دانشجو به هر سؤال داده است، برآورد توانایی هر دانشجو صورت می‌گیرد. نکته مهم اینکه در نظریه سؤال پاسخ تنها می‌توان پارامتر توانایی دانشجو را «برآورد» کرد و نمی‌توان مقدار آن را به صورت قطعی اعلام کرد. این کار با یک مقدار پیش‌تجربی آغاز می‌شود و یک فرایند تکرارشونده[1] است. همانطور که پارامتر توانایی با θ نشان داده می‌شد، براورد توانایی با θ' نشان داده می‌شود.

به عنوان مثال، دانشجویی در آزمون مذکور با همان چهار سؤال شرکت کرده است که تنها پاسخ سؤال اول را درست انتخاب کرده است. می‌خواهیم سطح توانایی او را برآورد کنیم. پارامترهای سؤالات آزمون به صورت جدول ۸-۳۵ است.

جدول ۸-۳۵: پارامتر دشواری و تمیز چهار سؤال

سؤال	پارامتر دشواری	پارامتر تمیز	پاسخ دانشجو
۱	۰/۵	۱	۱
۲	۲	۲	۰
۳	-۱	۰	۰
۴	-۱/۵	۱	۰

برای این کار مقدار توانایی دانشجو را به صورت پیش‌تجربی صفر فرض می‌کنیم و مقادیر جدول زیر را بر اساس مدل دوپارامتری محاسبه می‌کنیم. P احتمال پاسخگویی دانشجو با $\theta=0$ است که در مثال قبلی محاسبه کرده بودیم و Q، یک منهای این احتمال است. U مشخص کننده جوابی است که دانشجو به سؤال داده است. اگر درست پاسخ داده باشد، یک و اگر غلط جواب داده باشد، مقدار صفر برای آن منظور می‌شود. در نهایت قرار است مجموع (u-P) a و همچنین مجموع a^2 (PQ) را استخراج کنیم (جدول ۹-۳۵).

1. Iterative

جدول ۹-۳۵ : برآورد اجزای لازم برای محاسبه توانمندی با مقدار پیش‌تجربی برابر صفر

سؤال	۱	۲	۳	۴	جمع
U	۱	۰	۰	۰	
P	۰/۳۷	۰/۰۱	۰/۵۰	۰/۸۱	
Q	۰/۶۳	۰/۹۹	۰/۵۰	۰/۱۹	
a (u-P)	۰/۶۳۰	-۰/۰۲۰	۰/۰۰۰	-۰/۸۱۰	-۰/۲۰
a^2 (PQ)	۰/۲۳۳	۰/۰۴۰	۰/۰۰۰	۰/۱۵۴	۰/۴۲۷

برای برآورد توانایی، اختلاف بین دو این مقدار را از طریق تقسیم به دست می‌آوریم. در واقع:

$$\Delta_\theta = \frac{\sum a\,(u-p)}{\sum a^2\,(PQ)} = \frac{-۰/۲۰۰}{۰/۴۲۷} = -۰/۴۶۹$$

مقدار به دست آمده اختلافی است که توانایی دانشجو از سطح پیش‌تجربی یعنی صفر دارد. پس توانمندی دانشجو حدوداً به این مقدار است:

$$\theta' = ۰ - \Delta_{\theta'} = ۰ - (-۰/۴۹۶) = ۰/۴۶۹$$

اما گفتیم که فرایند برآورد یک فرایند تکرارشونده است. چون میزان تفاوت به دست آمده زیاد است، مقدار P سؤالات را دستکاری می‌کنیم تا اختلاف به حداقل برسد (جدول ۱۰-۳۵).

جدول ۱۰-۳۵ : برآورد اجزای لازم برای محاسبه توانمندی با تغییر مقدار پیش‌تجربی

سؤال	۱	۲	۳	۴	جمع
U	۱	۰	۰	۰	
P	۰/۲۷	۰/۰۰	۰/۴۰	۰/۷۱	
Q	۰/۷۳	۰/۱۰	۰/۶۰	۰/۲۹	
a (u-P)	۰/۷۳۰	۰/۰۰۰	۰/۰۰۰	-۰/۷۱۰	۰/۰۲۰
a^2 (PQ)	۰/۱۹۷	۰/۰۰۰	۰/۰۰۰	۰/۲۰۶	۰/۴۰۳

$$\Delta_\theta = \frac{\sum a\,(u-p)}{\sum a^2\,(PQ)} = \frac{-۰/۰۲۰}{۰/۴۰۳} = -۰/۰۵۰$$

$$\theta' = ۰/۴۶۹ \quad -\Delta_\theta = ۰/۴۶۹ - ۰/۰۵۰ = ۰/۴۱۹$$

در تکرار سوم باز هم مقدار P سؤالات را تغییر می‌دهیم (جدول ۱۱-۳۵):

جدول ۱۱-۳۵: برآورد اجزای لازم برای محاسبه توانمندی با تغییر مقدار پیش‌تجربی

سؤال	۱	۲	۳	۴	جمع
U	۱	۰	۰	۰	
P	۰/۲۸	۰/۰۰	۰/۴۱	۰/۷۲	
Q	۰/۷۲	۰/۱۰	۰/۵۹	۰/۲۸	
a (u-P)	۰/۷۲۰	۰/۰۰۰	۰/۰۰۰	-۰/۷۲۰	۰/۰۲۰
a^2 (PQ)	۰/۲۰۲	۰/۰۰۰	۰/۰۰	۰/۲۰۲	۰/۴۰۴

$$\Delta_\theta = \frac{\sum a\,(u-p)}{\sum a^2\,(PQ)} = \frac{-۰/۰۰۰}{۰/۴۰۴} = ۰/۰۰۰$$

$$\theta' = ۰/۴۱۹ \quad -\Delta_\theta = ۰/۴۱۹ - ۰/۰۰۰ = ۰/۴۱۹$$

به این ترتیب توانمندی دانشجوی فوق حدود ۰/۴۲ برآورد می‌شود.

روشی که برای برآورد توانایی دانشجو در اینجا به کار بردیم، در دو حالت قابل استفاده نیست:

- دانشجو به هیچ سؤالی جواب درست ندهد؛ که در این حالت توانایی وی منفی بی‌نهایت برآورد می‌شود.
- دانشجو به همه سؤالات درست پاسخ دهد؛ که توانایی وی مثبت بی‌نهایت برآورد می‌شود.

تابع آگاهی آیتم و تابع آگاهی آزمون

در فصل مربوط به پایایی دیدیم که در نظریه کلاسیک آزمون برای به دست آوردن نمره واقعی دانشجو (که در واقع میانگین نمرات او در بی‌نهایت بار تکرار آزمون بود)، خطای معیار اندازه‌گیری (SEM) را محاسبه می‌کردیم که با داشتن پایایی آزمون و انحراف معیار نمرات قابل محاسبه بود. در نظریه سؤال پاسخ نیز معادل این خطا وجود دارد که به آن خطای معیار برآورد[1] (SEE) گفته می‌شود. به همین ترتیب معادل پایایی که در نظریه کلاسیک آزمون میزان دقت آزمون را نشان می‌داد، در نظریه سؤال پاسخ مفهومی به اسم آگاهی[2] وجود دارد. مفهوم آگاهی در اینجا به میزان دقت در تخمین پارامتر توانایی اشاره دارد. هرچه دقت بیشتر باشد، با اطمینان بیشتری می‌توان در مورد درستی توانایی برآورد شده صحبت کرد.

میزان آگاهی حاصل از یک سؤال برای هر یک از سطوح توانایی جداگانه محاسبه می‌شود. هر سؤال در یک نقطه از سطح توانمندی قادر است که بیشترین اطلاعات را ارائه دهد. در دو سر طیف توانایی، یعنی سطوح منفی سه و مثبت سه، اغلب تعداد دانشجویان کم است و بنابراین مقدار داده‌ای که در اختیار است، اندک است. به همین دلیل هرچه از وسط خم ویژه یک سؤال به دو طرف آن حرکت کنیم، سؤال اطلاعات کمتری در مورد سطح توانایی دانشجویان در اختیارمان می‌گذارد.

آنچه میزان دقت در برآورد را نشان می‌دهد، پراکندگی برآوردها حول مقدار حقیقی توانایی است که نامعلوم است. بنابراین آگاهی سؤال برای هر سطح توانایی از طریق واریانس در فرمول زیر به دست می‌آید:

1. Standard Error of Estimate (SEE)
2. Information

$$\text{information} = \frac{۱}{\sigma^2}$$

برای به دست آوردن خود واریانس که پراکندگی θ' را حول مقدار نامعلوم θ نشان می‌دهد، در یک مدل دو پارامتری از فرمول زیر استفاده می‌شود:

$$\sigma^2 = \frac{۱}{a^2(PQ)}$$

در واقع مخرج کسر فوق همان مخرج معادله‌ای است که در قسمت قبلی برای برآورد توانایی دانشجو ذکر کردیم. بنابراین می‌توان گفت که آگاهی سؤال به این صورت به دست می‌آید:

$$I = a^2(PQ)$$

در مدل تک پارامتری که پارامتر تمیز دخالتی ندارد، فرمول فوق به این صورت قابل خلاصه شدن است:

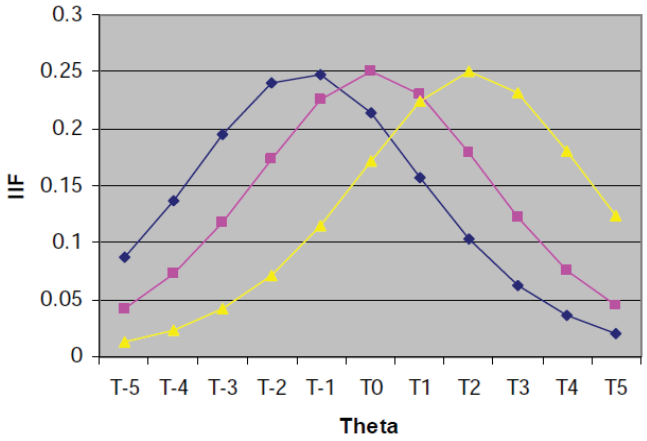
$$I = PQ$$

که در واقع:

$$I = P\,(۱-P)$$

از روی مقدار آگاهی به دست آمده برای هر سطح توانمندی، تابع آگاهی آیتم[1] را می‌توان محاسبه کرد که مشخص می‌کند یک سؤال چقدر قادر به دادن اطلاعات است. برای رسم تابع آگاهی ایتم (IIF)، توانایی را در محور افقی (منفی پنج تا مثبت پنج) و آگاهی را در محور عمودی (۰ تا ۰/۳) قرار می‌دهیم. تابع آگاهی در شکل ۱۲–۳۵ برای سه سؤال رسم شده است:
در مورد سؤال یک، حداکثر اطلاعات هنگامی حاصل می‌شود که سطح توانایی دانشجو منفی یک باشد. وقتی توانایی منفی پنج است، سؤال اطلاعات اندکی در اختیار می‌گذارد (۰/۰۸) اما در توانایی مثبت پنج اطلاعات در حد صفر است. به همین ترتیب سؤال شماره سه، بیشترین اطلاعات را برای دانشجویان با سطح توانایی مثبت دو به دست می‌دهد. اما برعکس سؤال قبلی برای دانشجویان با سطوح بالاتر از متوسط اطلاعات بیشتری در اختیار می‌گذارد تا دانشجویان با سطح پایین‌تر از متوسط.

1. Item Information Function (IIF)



شکل ۱۲-۳۵: تابع آگاهی سه سؤال

به همین ترتیب می‌توان آگاهی را برای کل یک آزمون نیز محاسبه و تابع آگاهی آزمون[1] را رسم کرد که مقدار آگاهی کل آزمون را نشان می‌دهد. همان طور که از IFF اطلاعاتی در مورد دقت تخمین پارامتر توانایی توسط یک سؤال استخراج می‌شود، می‌توان از TIF اطلاعات مربوطه را در مورد کل آزمون به دست آورد. یکی از کاربردهای این تابع در طراحی آزمون‌هایی است که قرار است به صورت موازی یا جایگزین برگزار گردند. در واقع TIF برای یکسان‌سازی و معادل‌سازی[2] آزمون‌ها مورد استفاده قرار می‌گیرد. هرچه تعداد سؤالات آزمون بیشتر باشد، میزان اطلاعاتی که می‌دهد بیشتر است یعنی آزمون برآوردها را با دقت بیشتر انجام می‌دهد. در اینجا هم آگاهی برای هر یک از سطوح توانمندی به صورت جداگانه محاسبه می‌شود وآگاهی آزمون، جمع آگاهی سؤالات آن آزمون در آن سطح توانمندی است.

نقشه سؤال-فرد

می‌توان عملکرد دانشجو و سؤال را در تعامل با یکدیگر نشان داد. برای توضیح این نمودار که به آن نقشه سؤال-فرد[3] می‌گویند، ابتدا مفهوم نسبت شانس[4] و سپس لاجیت[5] را مرور می‌کنیم. نسبت شانس برای سؤال، نسبت تعداد موارد نامطلوب (Q) به تعداد موارد مطلوب (P) است . به عنوان مثال اگر از ۵ دانشجو ۴ نفر به سؤال پاسخ صحیح داده باشند، نسبت شانس پاسخگویی به این سؤال به صورت زیر است:

$$Odds = \frac{Q}{P} = \frac{۱}{۴} = ۰/۲۵$$

مفهوم نسبت شانس را می‌توان به صورت احتمال وقوع رخداد نامطلوب به رخداد مطلوب بیان کرد. بنابراین فرمول آن را به این صورت نیز می‌توان نوشت:

$$Odds = \frac{Q}{P} = \frac{۱-P}{P}$$

1. Test Information Function (TIF)
2. Equating
3. Item-Person Map
4. odd ratio
5. Logit

به این ترتیب میزان احتمال پاسخگویی را می‌توان این گونه محاسبه کرد:

$$P = \frac{Odds}{1+Odds}$$

لاجیت به صورت مقدار لگاریتمی نسبت شانس تعریف می‌شود:

$$Logit = Log\ (Odds)$$

علت استفاده از لاجیت این است که در مقیاس‌بندی معمول فواصل نمرات در مکان‌های مختلف قابل مقایسه با یکدیگر نیستند. به عنوان مثال، تفاوت بین دو سؤال از نظر پارامتر دشواری در نقاط میانی (مثلاً ۵۰ درصد و ۵۵ درصد) مساوی همین میزان تفاوت در قسمت های بالاتر (مثلا ۹۵ درصد و ۱۰۰ درصد) یا پایین‌تر (مثلاً ۵ درصد یا ۱۰ درصد) نیست. به همین ترتیب تفاوت دو دانشجو با سطح توانایی مثبت پنج و مثبت چهار لزوماً با تفاوت بین دو دانشجو با سطوح عملکردی مثبت ۱ و صفر برابر نیست. این مشکل با در نظر گرفتن لاجیت حل می‌شود. برای تقریب ذهن فردی را در نظر بگیرید که قصد کاهش وزن دارد. شاید وی بتواند به راحتی وزن خود را از ۱۲۰ به ۱۰۰ برساند، اما در وزن ۸۰ کیلوگرم برای کم کردن هر کیلو باید تلاش مضاعفی داشته باشد و به راحتی نمی‌تواند بیست کیلوگرم کم کند.

اکنون در مورد مثال جدول ۱۱–۳۵ که شش دانشجو به پنج سؤال جواب داده‌اند، به محاسبه لاجیت می‌پردازیم. البته باید در نظر داشت که فرمول نسبت شانس برای عملکرد دانشجو $\frac{P}{1-P}$ و برای سؤال $\frac{1-P}{P}$ است. یعنی فرمول آنها با یکدیگر یکسان نیست.

جدول ۱۱–۳۵ : مقدار لاجیت محاسبه شده برای نقش دانشجو در پاسخ پنج سوال

شماره دانشجو	سوال ۱	سوال ۲	سوال ۳	سوال ۴	سوال ۵	odd	logit
۱	۱	۱	۱	۱	۰	۴	۰/۶
۲	۱	۱	۱	۰	۰	۱/۵	۰/۱۸
۳	۱	۱	۱	۰	۰	۱/۵	۰/۱۸
۴	۱	۱	۰	۰	۰	۰/۶۷	-۰/۱۷
۵	۱	۱	۰	۰	۰	۰/۶۷	-۰/۱۷
۶	۱	۰	۰	۰	۰	۰/۲۵	-۰/۶
odd	۰	۰/۲	۱	۵	-		
logit	-	-۰/۷	۰	۰/۷	-		

هنگامی که لاجیت یک سؤال منفی است، مانند سؤال شماره دو یعنی سؤال آسانی است. سؤال سوم که لاجیت آن صفر شده است، سؤال متوسطی است. سؤالی که لاجیت آن مثبت است، مانند سؤال چهار دشوار است. به همین ترتیب اگر لاجیت برای دانشجو منفی شود، به این معناست که سطح توانایی وی پایین‌تر از متوسط است و اگر مثبت شود، دانشجوی قوی محسوب می‌شود. همان‌طور که مشخص است اگر سؤالی توسط همه پاسخ داده شود (سؤال یک)، لاجیت آن غیرقابل محاسبه است و اگر

سؤالی توسط هیچ دانشجویی پاسخ داده نشود (سؤال پنج)، نسبت شانس و لاجیت آن غیرقابل محاسبه است.

در شکل ۱۳–۳۵ هیستوگرام پارامتر دشواری سؤالات (سمت چپ) و هیستوگرام توانایی دانشجویان (سمت راست) در مقیاس لاجیت نشان داده شده‌اند.

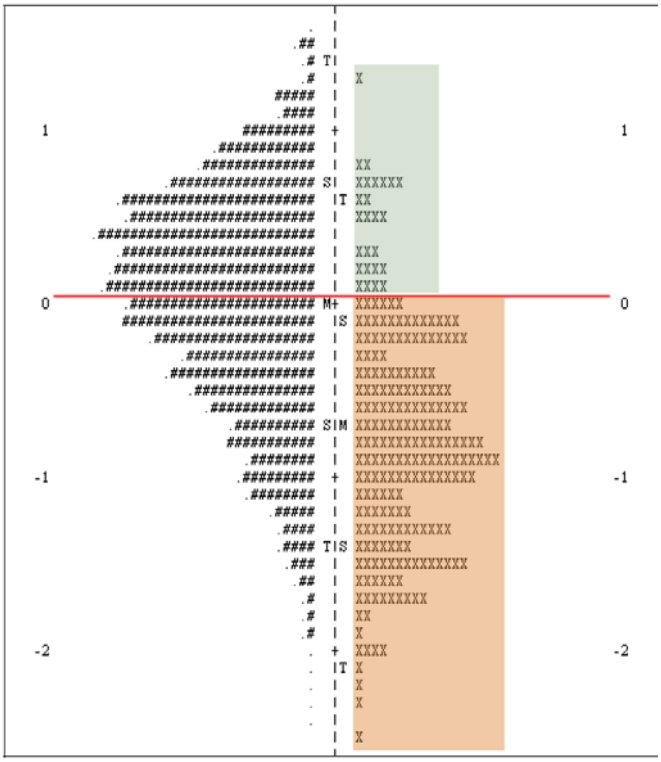
شکل ۱۳–۳۵: هیستوگرام دشواری سؤالات و هیستوگرام توانایی دانشجویان

اگر نحوه قرارگیری دو نمودار را به صورت زیر تغییر دهیم، نقشه سؤال–فرد به دست می‌آید. در این نقشه، سمت راست نمودار، نشان دهنده توانایی دانشجویان و سمت چپ نمودار، نشان دهنده پارامتر دشواری سؤالات است (شکل ۱۴–۳۵).

دانشجویانی که در بالای نمودار هستند، سطح توانایی بهتری نسبت به دانشجویان پایین نمودار دارند. اگر خط صفر را رسم کنیم که نشان دهنده پارامتر دشواری متوسط و توانایی متوسط است، خواهیم دید که دانشجویان پایین‌تر از متوسط، نیمی از سؤالات را به صورت صحیح جواب داده‌اند و دانشجویان قوی تعداد کمی از سؤالات را جواب داده‌اند. بنابراین آزمون به صورت کلی نسبت به سطح شرکت‌کنندگان سخت بوده است.

شکل ۱۴–۳۵: نقشه سؤال-فرد

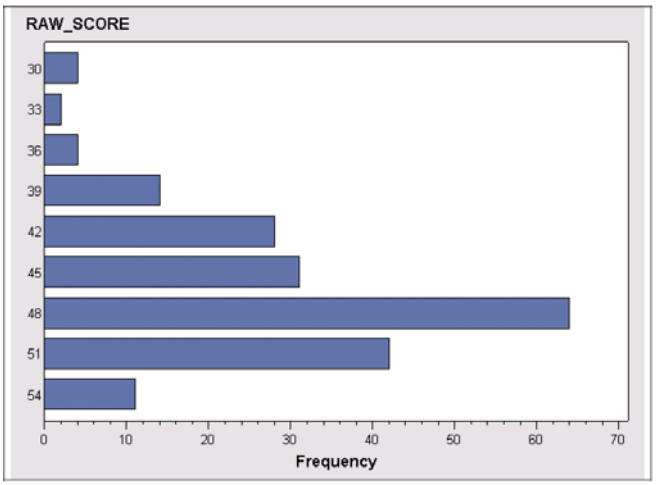
در شکل ۱۵–۳۵، نقشه سؤال–فرد آزمون دیگری قابل مشاهده است که در آن سؤالات در سمت راست و توانایی در سمت چپ نشان داده شده است. سؤالاتی که در بالا قرار گرفته‌اند، برای دانشجویانی که در پایین نمودار قرار دارند، دشوار هستند. در این مثال، به صورت کلی دانشجویان نسبت به سؤالات آزمون سطح بالاتری دارند. اگر خط صفر را که نشان‌دهنده وضعیت متوسط است، در نظر داشته باشیم، اکثر دانشجویان بالای متوسط و اکثر سؤالات پایین متوسط هستند. می‌توان مشاهده کرد که چگونه دانشجویان پایین‌تر از متوسط تعداد کمی از سؤالات را جواب نداده‌اند اما بیشتر سؤالات را جواب داده‌اند.



شکل ۱۵–۳۵: نقشه سؤال–فرد

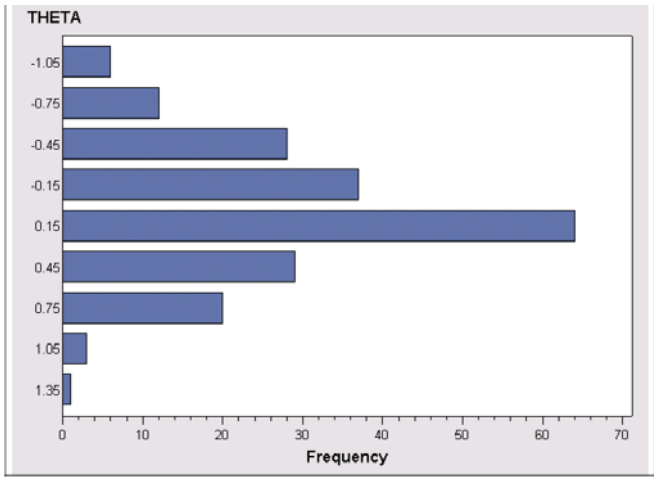
در نقشه سؤال–فرد فوق مشخص است که برخی از سؤالات در دو منتهاالیه نمودار قرار گرفته‌اند. به این موارد به صورت معمول outlier گفته می‌شود و در اینجا لفظ misfit به آنها اطلاق می‌شود (هرچند این دو دقیقاً یکی نیستند و با یکدیگر متفاوت هستند). بروز misfit به معنای دشوار بودن سؤال نیست بلکه به این معنا است که سؤال با کل آزمون در یک راستا نیست. به عنوان مثال، اگر یکی از سؤالات امتحان آناتومی از مباحث بافت‌شناسی طراحی شود، به عبارت دیگر سؤال نامرتبط باشد، احتمال دارد که الگوی پاسخ‌دهی به سؤال بافت‌شناسی از الگوی پاسخ‌دهی به سؤالات آناتومی تبعیت نکند و misfit دیده شود. در نظریه کلاسیک آزمون، چنین سؤالی با استفاده از ضریب همبستگی نقطه‌ای دورشته‌ای یا تحلیل عاملی مشخص می‌شود. در نظریه سؤال پاسخ می‌توان از دو راه میزان fit بودن را مشخص کرد. یکی با توجه به ICC سؤال که تا چه میزان منطبق بر ICC حاصل از داده‌های واقعی است و دیگری از طریق آزمون آماری Chi-Square test of goodness of fit.

همان طور که قبلاً گفته شد در نظریه سؤال پاسخ، پارامتر دشواری، مستقل از توانایی افرادی است که به سؤال جواب می‌دهند و همچنین توانایی افراد، مستقل از دشواری سؤالاتی که پاسخ می‌دهند، تخمین زده می‌شود. این مسأله را به صورت دیگری می‌توان دید. در شکل ۱۶-۳۵ توزیع نمرات دانشجویان در یک آزمون بر اساس نظریه کلاسیک آزمون نشان داده شده است. مشخص است که نمرات به سمت راست نمودار چولگی دارند. در این چارچوب تفسیری که می‌توان از آزمون داشت به این صورت است که اکثر دانشجویان نمره خوبی گرفته‌اند چون یا اکثر دانشجویان قوی بوده‌اند یا سؤالات آسان بوده‌اند. در هر حال، اتفاقی است که در آزمون‌های معیار محور پایان ترم معمولاً می‌افتد.



شکل ۱۶-۳۵: توزیع نمرات دانشجویان در نظریه کلاسیک آزمون

اما اگر از داده‌های همین آزمون برای برآورد توانایی دانشجویان بر اساس نظریه سؤال پاسخ استفاده شود و هیستوگرام آن رسم شود، به صورت شکل ۱۷-۳۵ خواهد بود.



شکل ۱۷-۳۵: توزیع نمرات دانشجویان در نظریه سؤال پاسخ

همان‌طور که دیده می‌شود، توزیع تقریباً زنگوله‌ای شکل است. در واقع توزیع لاجیت توانایی چندان غیرنرمال نیست زیرا همزمان دشواری سؤال را نیز در نظر می‌گیرد. حتی اگر سؤالات خیلی آسان باشند، توانایی دانشجویان بالاتر از مقداری که باید باشد، تخمین زده نمی‌شود. به عبارت دیگر این مسأله که برآورد توانایی مستقل از سؤال[1] است، از مزیت‌های نظریه سؤال پاسخ است. همین مسأله برای پارامتر دشواری نیز صادق است. سؤال دشواری که اتفاقاً بسیاری از دانشجویان به آن جواب داده‌اند، برخلاف آنچه در نظریه کلاسیک آزمون وانمود می‌شود، آسان برآورد نمی‌شود و به اصطلاح سؤال مستقل از نمونه[2] است.

کاربرد نظریه سؤال پاسخ در ارزیابی دانشجویان

همان‌طور که قبلاً گفته شد، یکی از کاربردهای نظریه سؤال پاسخ تولید تابع آگاهی آزمون است که برای تخمین دقت آزمون استفاده می‌شود و به نوعی می‌توان آن را معادل پایایی آزمون قلمداد کرد.

یکی دیگر از کاربردهای نظریه سؤال پاسخ تولید بانک سؤال است. در واقع پارامترهای یک سؤال بر اساس عملکرد دانشجویان در گروه‌ها و سال‌های مختلف محاسبه و نگهداری می‌شوند. سپس این سؤالات برای تدوین امتحانات مخصوصاً امتحانات گواهینامه یا امتحانات سطح بالا و تراکمی و پایانی در دانشکده‌ها به کار می‌روند.

از نظریه سؤال پاسخ در هنگام تدوین دفترچه سؤالات می‌توان استفاده کرد. به این معنا که برای آزمون یک تابع آگاهی خاص در نظر گرفته می‌شود و سپس سؤالاتی انتخاب می‌شوند که بتوانند در کنار هم منحنی مورد نظر را بسازند. هر سؤالی که به مجموعه اضافه می‌شود، تابع آگاهی آزمون رسم می‌شود و این روند آن‌قدر ادامه پیدا می‌کند تا نهایتاً نمودار به شکل مطلوب خود نزدیک شود. این رویکرد در آزمون‌های گواهینامه پزشکی بسیار کاربرد دارد. مخصوصاً در مواردی که به علل امنیتی فرم‌های موازی به کار گرفته می‌شوند. زیرا در این حالت ضروری است که نه تنها آزمون‌ها از نظر محتوا مشابه باشند بلکه باید دقت اندازه‌گیری آنها نیز یکسان باشد (دی چامپلین ۲۰۱۰).

یکی دیگر از کاربردهای نظریه سؤال پاسخ در آزمون‌های مبتنی بر کامپیوتر[3] است که در حال حاضر به خاطر مزایایی که هم برای دانشجو و هم برای موسسات دارند، در بسیاری از مراکز جایگزین آزمون‌های کاغذی شده‌اند. استفاده از این آزمون‌ها موجب انعطاف‌پذیری زمان و مکان آزمون می‌شوند بنابراین برای دانشجو این مزیت وجود دارد که طبق برنامه و شرایط و ترجیحات خودش در آزمون شرکت کند. از نظر اجرایی، کنترلی که روی شرایط و امنیت آزمون وجود دارد بیشتر است و همچنین پایایی و روایی آزمون تضمین می‌شود. نظریه سؤال پاسخ هم در مرحله تدوین دفترچه و هم در مرحله نمره‌دهی و گزارش نتایج آزمون‌های کامپیوتری کاربرد دارد.

در آزمون‌های تطبیقی کامپیوتری[4] استفاده از نظریه سؤال پاسخ از این هم فراتر می‌رود. در واقع برای هر دانشجو یک امتحان منحصر به فرد برگزار می‌شود. به این ترتیب که با ارائه یک سؤال متوسط از نظر دشواری و در نظر گرفتن پاسخ صحیح یا غلط دانشجو به آن برآورد اولیه‌ای از سطح توانایی دانشجو انجام می‌شود و بر این اساس سؤال دوم آسان‌تر یا سخت‌تر در نظر گرفته می‌شود. به همین دلیل به این نوع آزمون تطبیقی گفته می‌شود. به عبارت دیگر، بر اساس سطح توانایی هر دانشجو الگوریتم خاصی طراحی می‌شود و سؤالات با ویژگی‌های خاصی از میان بانک سؤالات برای او انتخاب می‌شوند. روند تا جایی ادامه پیدا می‌کند که دیگر دقت آزمون قابل بهبود نباشد. هدف این است که با حداقل تعداد سؤال ممکن، توانایی دانشجو با حداکثر دقت برآورد شود. شایان ذکر است که برای برگزاری آزمون تطبیقی کامپیوتری، بانک عظیمی از سؤالات لازم است که ویژگی‌های سایکومتریک آنها بر اساس نظریه سؤال پاسخ سنجیده شده باشد.

---

1. Item-independent
2. Sample-independent
3. Computer-based tests
4. Computer Adaptive Tests (CAT)

آنالیزهای مبتنی بر نظریه سؤال پاسخ ویژگی‌های خاصی دارد که لازم است توسط نرم‌افزارهای مخصوصی انجام شوند. بسته‌های متعددی برای انجام این آنالیزها تولید شده‌اند و در دسترس هستند. از جمله آنها می‌توان BILOG-MG را نام برد که برای تست‌های چندگزینه‌ای طراحی شده است. در خصوص نرم‌افزارهایی که برای داده‌های پلی‌توموس مانند پرسشنامه‌ها و آزمون OSCE قابل استفاده هستند، می‌توان به FACET و PARSCALE اشاره کرد.

منابع

1. Bock RD. A Brief History of Item Theory Response. Educational Measurement: Issues and Practice 1997, 16: 21–33
2. Chong Ho Yu. A Simple Guide to the Item Response Theory (IRT) and Rasch Modeling 2010. Available at: http://www.creative-wisdom.com
3. Chong Ho Yu, Jannasch-Pennell A, DiGangi S. A Non-Technical Approach for Illustrating Item Response Theory. Journal of Applied Testing Technology 2008
4. De Champlain AF. A primer on classical test theory and item response theory for assessments in medical education. Medical Education 2010, 44: 109–117
5. Fan X. Item response theory and classical test theory: an empirical comparison of their item/person statistics. Educational and Psychological Measurement 1998; 58(3):357-77
6. Comparison of classical test theory and item response theory and their applications to test development.
7. Hambleton RK, Jones RW. Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development. Educational Measurement: Issues and Practice 1993;12(3):38-47
8. Lawson DM. Applying the Item Response Theory to Classroom Examinations. Journal of Manipulative & Physiological Therapeutics 2006;29(5):393-397
9. Pell G, Fuller R, Homer M, Roberts T. How to measure the quality of the OSCE: A review of metrics – AMEE guide no.49. Med Teach 2010; 32(10): 802-11
10. Partchev I. A visual guide to item response theory. Friedrich-Schiller-Universit¨at Jena 2004
11. Tavakol M, Dennick R. Post-examination interpretation of objective test data: monitoring and improving the quality of high-stakes examinations: AMEE Guide No. 66. Med Teach 2012;34(3):e161-75.
12. Tavakol M, Dennick R. Psychometric evaluation of a knowledge based examination using Rasch analysis: an illustrative guide: AMEE guide no. 72. Med Teach 2013;35(1):e838-48.
13. Tor E, Steketee C. Rasch analysis on OSCE data : An illustrative example Australasian Medical Journal 2011;4(6):339-345

۱۴. پایه های اساسی نظریه سؤال پاسخ (نظریه جدید روانسنجی). نوشته بیکر ف. ۲۰۰۱. ترجمه هومن ح، عسگری ع. تهران، پارسا ۱۳۸۱

۹

نظام ارزیابی

فصل ۳۶

ضرورت نظام ارزیابی

مقدمه

طی سال‌های متمادی، ابزارهای ارزیابی در آموزش پزشکی به صورت مجزا به کار گرفته شده‌اند. پژوهش‌های انجام شده در این حوزه نیز بر ابزارهای ارزیابی جداگانه و بررسی کیفیت روانسنجی هر یک از آن‌ها متمرکز بوده است. این مسأله با در نظر گرفتن دیدگاه غالب حاکم، مبنی بر تعریف توانمندی‌های یک پزشک به صورت اجزای جداگانه –دانش، مهارت، نگرش و حل مسأله– و انتخاب مناسب‌ترین ابزار برای ارزیابی هر یک منطقی به نظر می‌رسد. بهترین مثال تأیید کننده این دیدگاه در نظر گرفتن OSCE[1] به عنوان ابزار ارجح برای اندازه‌گیری مهارت‌های بالینی و آزمون «ویژگی‌های کلیدی»[2] به منظور ارزیابی مهارت‌های حل مسأله است. در واقع، این رویکرد با تمرکز بر یک روش ارزیابی واحد و همچنین معیارهای روان‌سنجی، بینش ارزشمندی را در مورد مزایا و محدودیت‌های ابزارهای ارزیابی و تعادلی که باید در انتخاب ابزارها رعایت کرد فراهم کرده است اما بدون این‌که بخواهیم این ارزش را از نظر دور بداریم، خاطر نشان می‌کنیم که این رویکرد به منظور ارزیابی مطلوب توانمندی به عنوان یک کل کافی نیست (دایکسترا و همکاران[3] ۲۰۱۰).

در طول دهه‌های اخیر، پیشرفت‌های چشمگیری در حوزه ارزیابی دانشجو رخ داده است. یکی از این پیشرفت‌ها، حرکت از ارزیابی ابزار محور به سمت ارزیابی مبتنی بر سیستم، تحت عنوان «نظام ارزیابی[4]» است. یکی از دلایل این تغییر، برنامه‌های درسی مبتنی بر توانمندی است که به روش‌های ارزیابی متنوعی برای تعیین میزان دستیابی به توانمندی نیاز دارند. توانمندی جمع اجزای جداگانه نیست بلکه یک کل تلفیق یافته است. بنابراین هیچ ابزار ارزیابی به تنهایی، حتی اگر از نظر روانسنجی مناسب باشد، نمی‌تواند همه اطلاعات مربوط به ارزیابی جامع توانمندی در یک حوزه گسترده مانند پزشکی را فراهم کند. توانمندی به عنوان ظرفیت ترکیب و ادغام دانش، مهارت و نگرش در محیط کاری خاص در نظر گرفته می‌شود و بنابراین ارزیابی آن نیز باید بر تلفیق این سه عنصر تمرکز داشته باشد (ون‌درولوتن و شوورث[5] ۲۰۰۵).

هرم میلر[6] که اخیراً شهرت بسیاری یافته است، چارچوبی را به منظور انتخاب ابزارهای ارزیابی مناسب برای عناصر مجزای توانمندی در چهار سطح فراهم می‌آورد. ولی هرم میلر ارتباط بین سطوح یا تلفیق ابزارها را توصیف نمی‌کند. متأسفانه در مورد ارتباطات و مبادلات بین سطوح هرم میلر و تلفیق ابزارهای ارزیابی اطلاعات کمی در دسترس است. به

1. Objective Structured Clinical Exam
2. Key Feature (KF)
3. Dijkstra et al
4. Programmatic assessment
5. Van der Vleuten & Schuwirth
6. Miller

طور یقین تنها مخلوط کردن ابزارهای ارزیابی و استفاده تصادفی از آنها کفایت نمی‌کند؛ بلکه چیدمان هدفمندی از روش‌ها به منظور ارزیابی جامع توانمندی مورد نیاز است. در قیاس با این دیدگاه پذیرفته شده که یک آزمون خوب چیزی بیش از تنظیم تصادفی سؤالات با کیفیت مناسب است، یک برنامه ارزیابی خوب هم بیش از مجموعه‌ای تصادفی از ابزارهای مناسب است. برنامه ارزیابی حتی فراتر از تشبیه فوق است. به این دلیل که، هر چند سؤالات با کیفیت خود قابل دستیابی هستند ولی چیزی به عنوان یک ابزار ایده‌آل وجود ندارد (دایکسترا ۲۰۱۰).

تعریف نظام ارزیابی

ارائه تعریف واحد برای نظام ارزیابی کار آسانی نیست اما می‌توان گفت که نظام ارزیابی یک فرایند طراحی است که طی آن تصمیمات آگاهانه، مبتنی بر شواهد و منطقی درباره حوزه‌های مورد ارزیابی، روش‌های ارزیابی مربوط، نحوه ترکیب نتایج به دست آمده از منابع مختلف و مصالحه‌ای که باید بین مزایا و محدودیت‌های اجزای برنامه در نظر گرفته شود اتخاذ می‌شود (دایکسترا ۲۰۱۰).

به عبارت دیگر، برنامه ارزیابی، چیدمان آگاهانه و سنجیده‌ای از فعالیت‌های ارزیابی دانشجو است که در آن یک ابزار ارزیابی در ارتباط با دیگر عناصر و کل برنامه معنای خود را پیدا می‌کند. در یک برنامه ارزیابی با چیدمان مناسب، تلاش می‌شود سنت‌های حاکم در محیط آموزشی و ارزیابی دانشجو، قوانین و مقرارت، فلسفه آموزشی و منابع در نظر گرفته شوند. برنامه ارزیابی آرایشی از روش‌های ارزیابی است که به منظور به حداکثر رساندن کیفیت یا تناسب با هدف برنامه‌ریزی شده است. با در نظر داشتن کیفیت، توصیه می‌شود که برنامه ارزیابی هدفمند طراحی شود، عناصر آن در راستای هدف آن باشد، استقرار و اجرای آن به طور مرکزی اداره شود و به طور مرتب ارزشیابی و اصلاح شود. در واقع، انتخاب و چیدمان هدفمند ابزارهای ارزیابی در برنامه ارزیابی موجب می‌شود که کل چیزی بیش از جمع اجزای آن باشد. بنابراین، لازم نیست هر روش ارزیابی به تنهایی کامل باشد بلکه در برنامه ارزیابی، ترکیبی از روش‌های سنتی و نوین ارزیابی می‌توانند با توجه به هدف برنامه ارزیابی مورد استفاده قرار گیرند.

در یک برنامه ارزیابی خوب بین عناصر مختلف توانمندی و منابع متعدد جمع‌آوری اطلاعات ارتباط برقرار می‌شود تا توانمندی‌ها در موقعیت‌های مختلف با استفاده از استانداردهای معتبر مورد ارزیابی قرار گیرد. سپس اطلاعات به دست آمده با هم تجمیع می‌شوند تا مبنای تصمیم‌گیری نهایی قرار گیرند. به بیان دیگر جمع‌آوری اطلاعات در سطح روش‌های ارزیابی شروع می‌شود ولی قضاوت نهایی بر اساس مجموعه‌ای مکفی از اطلاعات فراهم شده توسط روش‌های واحد ارزیابی صورت می‌گیرد. وقتی همه منابع جمع‌آوری اطلاعات در یک جهت باشند، اطلاعات منسجم بوده و تصمیم‌گیری نسبتاً ساده خواهد بود. برنامه ارزیابی همچنین شامل یک دیدگاه طولی یادگیری و ارزیابی در ارتباط با یک پیامد یادگیری مشخص است. بنابراین، رشد و ارتقای فراگیران تسهیل و پایش می‌شود.

علاوه بر این، هر برنامه ارزیابی خوب باید مجهز به مکانیسم‌های ارائه بازخورد به فراگیران در زمان مناسب خود باشد تا تصمیم نهایی برای فراگیران دور از انتظار و تعجب برانگیز نباشد. اگر چنین اتفاقی رخ دهد و تصمیم نهایی دور از انتظار فراگیران باشد یعنی در بخشی از برنامه ارزیابی مکانیسم بازخورد حذف یا دچار مشکل شده است (ون‌درولوتن و شوورث ۲۰۰۵).

ضرورت و مزایای نظام ارزیابی

ون‌درولوتن و همکارن به خوبی به این سول پاسخ داده‌اند که چرا باید به برنامه ارزیابی فکر کنیم و برنامه ارزیابی در مقابل روش‌های ارزیابی جداگانه چه محاسنی دارد (ون‌درولوتن و همکارن ۲۰۱۲):

- توانمندی کاربرد ادغام‌یافته دانش، مهارت و نگرش است، بنابراین یک روش ارزیابی به تنهایی نمی‌تواند توانمندی را ارزیابی کند و تلفیقی از روش‌های ارزیابی مورد نیاز است.
- برنامه ارزیابی، تصویری کلی از آنچه اندازه‌گیری می‌شود و آنچه اندازه‌گیری نمی‌شود، فراهم می‌آورد. بنابراین بین محتوا و جنبه‌های دیگر توانمندی تعادل برقرار می‌کند و از تأکید بیش از حد بر ارزیابی عناصری مانند دانش مربوط به حقایق غیرمرتبط که سنجش آن آسان است پیشگیری می‌کند. علاوه بر این، کارآیی ارزیابی به دلیل کاهش در جمع‌آوری اطلاعات تکراری افزایش می‌یابد. وقتی اطلاعات در مورد یک موضوع از طریق یک ابزار جمع آوری شده است، زمان و فضا می‌تواند به ارزیابی موضوعات دیگر اختصاص یابد.
- یک برنامه ارزیابی، یعنی تلفیق ارزیابی‌های مختلف، موجب می‌شود ضعف‌های برخی از ابزارها با مزایای ابزارهای دیگر جبران شود. در نتیجه طیف متنوعی از ابزارهای اندازه‌گیری مکملِ هم را ایجاد می‌کند که می‌توانند توانمندی را به عنوان یک کل ارزیابی کنند
- در آزمون‌های مهم[1]، اطلاعات از منابع مختلف منجر به دستیابی به تصمیمات آگاهانه و کاملاً قابل دفاع می‌شود.
- در برنامه ارزیابی، توجه به چیدمان همه ارزیابی‌ها در کل برنامه آموزشی است. حسن این رویکرد این است که تأکید بر پایایی آزمون‌های با حساسیت و اهمیت کم کاهش یافته و در نتیجه منابع می‌تواند صرف طراحی آزمون‌های پایا و پرهزینه در جای مناسب آن یعنی آزمون‌های حساس و مهم شود.
- پایایی بیش از آن‌که به استاندارد، ساختارمند و عینی نمودن روش‌های ارزیابی بستگی داشته باشد به نمونه‌گیری بستگی دارد. برنامه ارزیابی با استفاده از ابزارها و منابع مختلف در ارزیابی، که برخی از آن‌ها کیفی و مبتنی بر قضاوت خبرگان است پایایی ارزیابی دانشجو را افزایش می‌دهد.

مبانی پایه نظام ارزیابی

در این بخش مفاهیم و مبانی پایه‌ای که برنامه ارزیابی بر آن بنا نهاده شده است را بازگو می‌کنیم (ون‌درولوتن و همکاران، ۲۰۱۲):

- **هر نوع اطلاعات ارزیابی مربوط به یک ابزار ارزیابی واحد، ناقص است.** اجرای یک روش ارزیابی به تنهایی در هر سطحی از هرم میلر دارای محدودیت ذاتی است. به دلیل ویژگی محتوا دستیابی به پایایی مطلوب با یک روش ارزیابی مشکل است. علاوه بر این، یک روش ارزیابی به تنهایی نمی‌تواند همه بخش‌های هرم میلر را ارزیابی کند. همچنین یک روش ارزیابی به تنهایی نمی‌تواند موجب تغییر یا بهبود رفتار فراگیر شود. محدودیت‌های استفاده از اطلاعات یک روش ارزیابی واحد تفکر ما را به سوی برنامه ارزیابی سوق می‌دهد، مشروعیت می‌بخشد و جهت می‌دهد.
- **به منظور تضمین روایی ابزارهای سطح بالای هرم میلر آموزش ارزیابان اهمیت حیاتی دارد.** در همه ابزارهای ارزیابی در سه سطح اول هرم میلر که قابلیت استاندارد شدن را دارند، با دقت در تدوین محتوا و پروسیجرهای نمره‌دهی و اجرا می‌توان روایی ابزار را تأمین کرد. رعایت پروسیجرهای تضمین کیفیت در زمان ساخت آزمون مانند

1. High-stake

آموزش ارزیابان، عینی نمودن نمره‌دهی، استانداردسازی بیماران شبیه سازی‌شده و غیره تأثیر بارزی بر کیفیت آزمون دارد اما روایی ابزارهای غیراستاندارد بیشتر به استفاده کنندگان آن بستگی دارد تا خود ابزارها. در یک برنامه ارزیابی کامل به طور اجتناب‌ناپذیری لازم است از روش‌های ارزیابی غیر استاندارد استفاده شود. به ویژه زمانی که ارزیابی در سطح بالای هرم میلر، در محیط واقعی، مطرح است. کیفیت ارزیابی‌های مذکور به این بستگی دارد که استفاده‌کنندگان ارزیابی، چقدر در زمینه انجام ارزیابی تخصص دارند، به چه میزان ارزیابی را جدی می‌گیرند و میزان زمانی که صرف آن می‌کنند چقدر است. هر چند آموزش وسیعی برای فردی که آزمون چند گزینه‌ای طراحی می‌کند لازم نیست، ولی تربیت افرادی که درگیر ارزیابی غیر استاندارد به صورت مشاهده مستقیم عملکرد در محیط واقعی هستند اهمیت حیاتی دارد. این که افراد چقدر ارزیابی را جدی می گیرند در زمانی که برای بازخورد اختصاص می‌دهند یا در جملاتی که در فرم ثبت می‌کنند نمایان می‌شود. به این دلیل که یک برنامه ارزیابی بدون روش‌های غیر استاندارد غیر قابل تصور است ما نیازمند طراحی یک سیستم به منظور کمک به استفاده‌کنندگان به منظور ایفای عملکرد صحیح در نقش ارزیابی ایشان هستیم. این سیستم شامل آموزش، تمرین و بازخورد به منظور ارتقای درک و فهم استفاده کنندگان از مفهوم و هدف ارزیابی است.

- **از لحاظ چارچوب مفهومی در نظام ارزیابی تمایز قائل شدن بین ارزیابی تکوینی و تراکمی خیلی مفید نیست.** هر روش ارزیابی می‌تواند هم تکوینی و هم تراکمی باشد تنها درجه آن متفاوت است. بنابراین مفهوم سازی اهمیت[1] آزمون به عنوان یک طیف از اهمیت کم تا زیاد به نظر مفید می‌رسد. در ارزیابی‌های با اهمیت کمتر، نتایج ارزیابی عواقب کمتری برای فراگیران از نظر ارتقا به سطح بالاتر، انتخاب یا اعطای مدرک پیدا می‌کند، در حالی‌که می‌تواند پیامدهای گسترده و چشمگیری بر یادگیری و عملکرد فراگیران داشته باشد. در برنامه ارزیابی، تنها ارزیابی‌های با حساسیت کم می‌توانند بر داده‌های واحد (استفاده از یک روش ارزیابی) بنا شوند، در حالی که تصمیم‌گیری‌های مهم و حساس نیاز به داده‌های بیشتری دارند. هر چه حساسیت و اهمیت آزمون بیشتر می‌شود، نقش قضاوت در آن پررنگ‌تر می‌شود. در نتیجه نقش استاد به عنوان حمایت کننده به قضاوت کننده تغییر می‌یابد. نقش قضاوتی استاد ممکن است نقش حمایتی وی را تحت تاثیر قرار دهد و بالعکس. هر چقدر تصمیمات مهم با داده‌های متنوع حمایت شوند این تضاد نقش را کمتر می‌کند. در مجموع در برنامه ارزیابی اهمیت و حساسیت آزمون به عنوان یک طیف در نظر گرفته می‌شود که در آن ارتباط نسبی بین افزایش اهمیت آزمون و تعداد داده‌های در نظر گرفته شده وجود دارد.
- **ارزیابی، یادگیری را جهت می‌دهد.** این مفهوم پذیرفته شده‌ای در ادبیات ارزیابی و از مبانی اصلی برنامه ارزیابی است ولی با این وجود به خوبی درک نشده است. به احتمال زیاد، بسیاری از روش‌های ارزیابی راهبردهای یادگیری نامطلوب را تقویت می‌کند به این دلیل که ارزیابی با اهداف برنامه درسی هم‌راستا نیست. این وضعیت به ویژه در سیستم‌های با داده‌های ضعیف و به صورت کاملا تراکمی رایج است.

ما نیازمند شفاف‌سازی نظری بیشتری در مورد چرایی و چگونگی جهت‌دهی یادگیری از طریق ارزیابی هستیم و پژوهش‌های بیشتری در این زمینه در حال انجام است. هدف این است که یادگیری در جهت مناسبی هدایت شود و رویکردهای یادگیری عمیق تقویت شود (همچنین در صورت نیاز، یادگیری در حد تسلط). شواهد محکمی وجود دارد که بازخورد تکوینی یادگیری را افزایش می‌دهد.

یک روش ارزیابی در صورتی هدایت کننده یادگیری است که اطلاعات معنی‌داری برای فراگیران فراهم کند. به منظور ارائه اطلاعات معنی‌دار به فراگیران و ارتقای یادگیری آن، باید اطلاعات ارزیابی تا حد امکان غنی باشد که این اطلاعات می‌تواند به روش‌های مختلفی، هم کیفی و هم کمی جمع‌آوری شود. در حال حاضر، ارزیابی

1. Stake

اغلب تنها با نمرات ارتباط دارد و فراگیران از طریق نمره بازخورد دریافت می‌کنند. نمره یکی از ضعیف‌ترین اشکال بازخورد است. در حالی‌که انواع متفاوتی از اطلاعات کمی مانند نیمرخ نمرات و اطلاعات عملکرد مرجع[1] مورد نیاز است، اطلاعات کیفی نیز اهمیت دارند. اطلاعات نقلی[2] ابزار قوی برای بازخورد کیفی هستند و می‌توانند به طور قابل ملاحظه‌ای در معنادار نمودن اطلاعات کمی سهیم باشند. به منظور ارائه نتایج معنی‌دار به فراگیر لازم است مهارت‌های ارائه بازخورد ارزیابان توسعه یابند. خود فرایند انجام آزمون نیز باید به گونه‌ای باشد که اطلاعات مهم و معنی‌داری جمع‌آوری شود. اگر فراگیران با به خاطر سپردن چک لیست‌ها بدون دیدن بیمار می‌توانند از سد آزمون OSCE عبور کنند عملکردشان بی اهمیت است؛ اگر یک ارزیاب همه آیتم‌های فرم ارزیابی رفتار حرفه‌ای را با یک حرکت قلم تیک می‌زند، ارزیابی همه معنای خود را از دست می‌دهد و بی‌مفهوم می‌شود.

بنابراین، یک روش ارزیابی واحد باید تا حد امکان با معنی باشد تا یادگیری را سرعت بخشد و تصمیمات مهم نیز باید بر اساس تعداد زیادی از روش‌های ارزیابی واحد باشند. تجمیع هدفمند روش‌های ارزیابی واحد منجر به تصمیمات حساس و مهم معنی‌دار می‌شود. تنها یک استثنا وجود دارد که یک ارزیابی واحد می‌تواند مبنای تصمیم‌گیری‌های حساس قرار گیرد و آن وقتی است که یادگیری یک تکلیف در حد تسلط و مهارت مدنظر است (به عنوان مثال، یادگیری جدول ضرب برای کودکان یا عملیات احیای قلبی و ریوی برای دانشجویان پزشکی) (ون‌درولوتن و همکاران ۲۰۱۲). در نتیجه در برنامه ارزیابی باید این استثنا را لحاظ کنیم و این کسب مهارت و تسلط در این وظایف نیاز به اعطای گواهی دارد. هر چند این به این معنی نیست که وظایف یادگیری در حد تسلط نیاز به بازخورد ندارند.

- **در برنامه ارزیابی، قضاوت متخصصان امری ضروری است.** در اغلب موارد سعی می‌شود با طراحی روبریک‌های نمره‌دهی، آموزش ارزیابان و تدوین استانداردهای عملکرد ذهنی بودن قضاوت را کاهش داد، ولی لازم است به این نکته توجه کنیم که اصرار بر عینی سازی کامل، فرایند ارزیابی را خالی از مفهوم می‌کند. توانمندی امری پیچیده است. بنابراین ارزیابی مبتنی بر توانمندی نیازمند قضاوت افراد متخصص و مطلع در بخش‌های مختلف فرایند ارزیابی است.

علاوه بر این، در ترکیب اطلاعات روش‌های ارزیابی واحد نیز به قضاوت متخصصان نیاز داریم. اغلب از راهبردهای کمی مانند معدل‌گیری نمرات و شمارش تعداد قبول شدگان برای تجمیع منابع اطلاعات استفاده می‌شود، ولی زمانی که روش‌های ارزیابی واحد نیز غنی بوده و به ویژه زمانی که شامل اطلاعات کیفی است، تجمیع کمی اطلاعات خالی از اشکال نیست و ما نیازمند توسل به قضاوت متخصصان هستیم. بر اساس شواهد مربوط به تصمیم‌گیری، ذهن انسان در مقایسه با تصمیم‌گیری‌های مبتنی بر آمار جایز الخطا است. در عین حال سوگیری‌های تصادفی مربوط به قضاوت، با راهبردهای نمونه‌گیری و سوگیری‌های سیستماتیک با تدوین و اجرای پروسیجرهای مناسب برای تصمیم گیری کاهش می‌یابد. نمونه‌گیری در بسیاری از موقعیت‌های ارزیابی مؤثر بوده است؛ می‌توان با استفاده از قضاوت‌های بیشتر اطلاعات پایایی تولید کرد. البته، روش‌های ارزیابی که به میزان نسبتاً زیادی بر قضاوت متکی است، به طور قابل ملاحظه‌ای نسبت به روش‌های استاندارد و عینی شده به حجم نمونه کمتری نیاز دارد. پیشگیری از سوگیری هر چند مشکل است، ولی می‌توان با استفاده از پروسیجرهای مناسب در تصمیم‌گیری میزان آن را کاهش داد. به عنوان مثال، تصمیم‌گیری در مورد داوطلبان مرزی، به فرایند جمع‌آوری اطلاعات به صورت موشکافانه‌تر و حتی شاید اطلاعات بیشتر و بررسی بیشتر اطلاعات جمع‌آوری شده نیاز دارد. در نتیجه باید پروسیجرهای مناسبی برای قضاوت در مورد عملکرد این گروه از فراگیران در نظر گرفته

1. Reference
2. Narrative

شود تا سوگیری کاهش یابد. علاوه بر این، تفسیر و قضاوت اطلاعات کیفی نیز نیاز به پروسیجرهای دقیق دارد. روش‌های مورد استفاده برای تأمین روایی و پایایی پژوهش‌های کیفی می‌تواند به عنوان الهام بخش و مبنایی برای تدوین پروسیجرهای مربوط مورد استفاده قرار گیرد. بسته به مراقبتی که در تدوین و اجرای این پروسیجرها می‌شود سوگیری می‌تواند کاهش پیدا کند و منجر به تصمیمات معتبرتر و قابل دفاع‌تری شود.

به نظر می‌رسد راهبردهای ذکر شده در بالا بتوانند اطلاعات ذهنی (تلفیق شده با اطلاعات عینی) را مدیریت و تصمیم‌گیری‌های منتج شده را تقویت کنند تا نیاز به عینی کردن همه بخش‌های برنامه ارزیابی برطرف گردد و تقلیل‌گرایی[1] بیهوده در ارزیابی رخ ندهد.

- **کیفیت برنامه ارزیابی توانمندی باید به صورت ادغام یافته ارزشیابی شود.** یک برنامه ارزیابی توانمندی شامل هر دو نوع اهداف تکوینی و تراکمی است. برخی عقیده دارند که پایایی برای ارزیابی تراکمی اهمیت بسیار دارد ولی برای ارزیابی تکوینی از اهمیت کمی برخوردار است. چنین تمایزی بین ارزیابی تکوینی و تراکمی با این خطر همراه است که ارزیابی تکوینی بر اساس معیارهای مربوط به «یادگیری» ارزشیابی شود و ارزیابی تراکمی بر اساس معیارهای تکنیکی سنتی مورد نقد و ارزشیابی قرار گیرد. در حالی که ارزیابی تراکمی نیز پتانسیل سازنده بودن را به دلیل هدایت فرایند یادگیری دارد. بنابراین معیارهای کیفیت مربوط به یادگیری برای ارزیابی تراکمی نیز مهم هستند و کیفیت برنامه ارزیابی توانمندی باید به صورت ادغام‌یافته ارزشیابی شود.

1. Reductionism

منابع

1. Amin Z. Purposeful assessment. Med educ. 2012; 46(1): 4–7.
2. Dijkstra J, Van der Vleuten C, Schuwirth L. A new framework for designing programmes of assessment. Adv Heal Sci Educ. 2010;15:379–93.
3. Schuwirth LW, Van der Vleuten CP. Programmatic assessment: From assessment of learning to assessment for learning. Med Teach. 2011; 33(6):478–85.
4. Van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: from methods to programmes. Med Educ. 2005; 39(3):309-17.
5. Van der Vleuten CPM, Schuwirth LWT, Driessen EW, Dijkstra J, Tigelaar D, Baartman LKJ, et al. A model for programmatic assessment fit for purpose. Med Teach. 2012; 34(3):205-14.
6. van der Vleuten CPM, Dannefer EF. Towards a systems approach to assessment. Med Teach. 2012; 34(3); 185-6.
7. van der Vleuten CP, Schuwirth LWT, Driessen EW, Govaerts MJB, Heeneman S.. 12 Tips for programmatic assessment. Med Teach. 2015;37(7):641-6.

فصل ۳۷

طراحی و اجرای نظام ارزیابی

مقدمه

ارزیابی دانشجو یک مسأله مربوط به اندازه‌گیری و روانسنجی یک روش ارزیابی نیست، بلکه یک نوع طراحی آموزشی است که مشتمل بر جنبه‌های آموزشی، اجرایی و صرف منابع است و کل برنامه درسی را تحت تاثیر قرار می‌دهد.

برخلاف تصور رایج، در برنامه ارزیابی، موضوع سنتی یا نوین بودن روش ارزیابی اهمیت ندارد، بلکه باید دلیل و نحوه انتخاب یک ابزار با توجه به شرایط، توجیه منطقی داشته باشد (ون‌درولوتن و شوورث ۲۰۰۵). بنابراین، برنامه ارزیابی نیاز به مدیریت مرکزی و یک برنامه تدوین شده مشخص دارد تا اجزای آن را در کل برنامه درسی هماهنگ کند و به طور مستمر مورد پایش و اصلاح واقع شود.

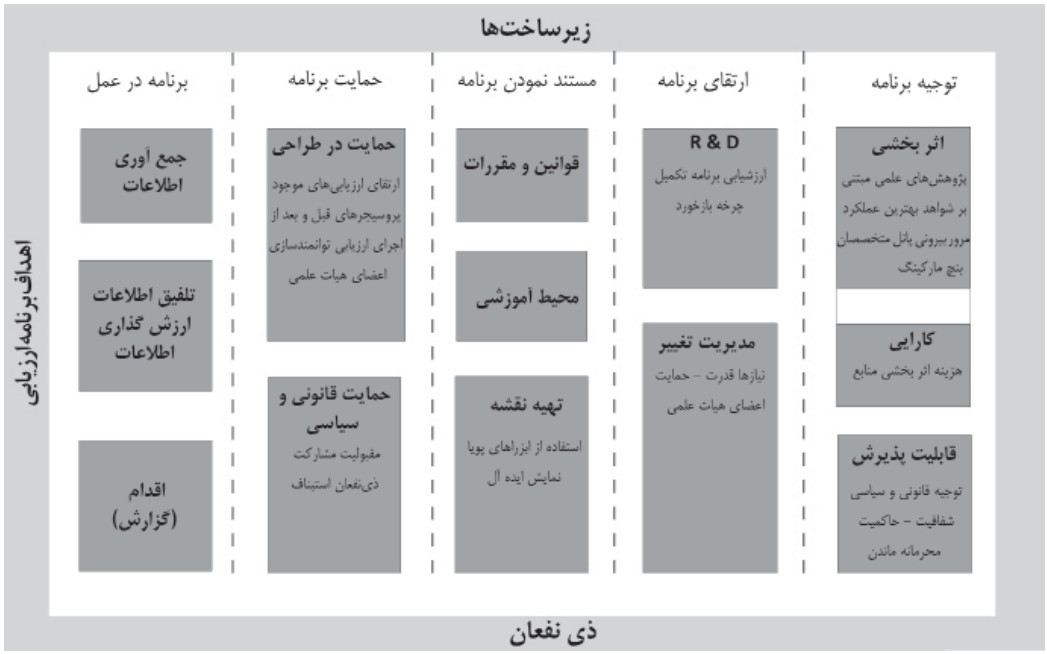
گام‌های طراحی و اجرای نظام ارزیابی

طراحی یک برنامه ارزیابی مؤثر، کاری مشکل و چالش برانگیز است. برنامه ارزیابی ساختار مشخصی (مشتمل بر اعضای هیات علمی، روش‌های ارزیابی و منابع) برای اجرای فرایندهای آن (شامل تجربیات و فعالیت‌های یادگیری و ارزیابی) به منظور دستیابی به پیامدها (توانمندی‌های انتهای برنامه) دارد. در شکل ۱ عناصر تشکیل دهنده و وسعت برنامه ارزیابی نشان داده شده است (دایکسترا و همکاران ۲۰۱۰ و ۲۰۱۲).

بر اساس مدل مذکور، مبنای طراحی برنامه ارزیابی، اهداف[1] برنامه هستند. همچنین زیرساخت‌ها[2] و ذی‌نفعان[3] بر همه ابعاد برنامه ارزیابی و اهداف آن تاثیرگذار هستند. علاوه بر سه عنصر پایه و ضروری مذکور، برنامه ارزیابی از پنج بُعد دیگر تشکیل شده است. بعد «برنامه در عمل[4]» شامل چهار عنصر درهم‌تنیده جمع‌آوری[5]، ترکیب[6]، ارائه گزارش[7] و تصمیم‌گیری[8] است. برنامه در عمل به معنی کلیه فعالیت‌هایی است که به عنوان حداقل‌های اجرای برنامه ارزیابی مورد نیاز هستند و همانطور که ذکر شد در یک طیف از جمع‌آوری اطلاعات تا اقدام بر اساس این اطلاعات متنوع هستند. بخش اعظمی از برنامه ارزیابی به فعالیت‌های تضمین کیفیت ارزیابی تعلق دارد که می‌توان آن را در چهار بعد قرار داد و به جای آن‌که

1. Purpose
2. Infrastructure
3. Stakeholders
4. Progrmme in action
5. Collecting
6. Combining
7. Reporting
8. Decision taking

به صورت یک عنصر مجزا در انتهای فرایند باشند، با اهداف و برنامه در عمل در هم‌تنیده هستند. ابعاد مربوط به تضمین کیفیت عبارتند از: «حمایت[1]»، «مستندسازی[2]»، «بهبود[3]» و «توجیه[4]» برنامه. همان‌طور که در شکل ۱-۳۷ ملاحظه می‌شود، هر یک از این ابعاد به نوبه خود از عناصری تشکیل شده است. در زیر گام‌های طراحی برنامه ارزیابی با توجه به ابعاد و عناصر چارچوب فوق ارائه می‌شود.



شکل ۱-۳۷ چارچوب برنامه ارزیابی (دایکسترا و همکاران ۲۰۱۲)

۱. تعریف شفاف هدف برنامه ارزیابی

کیفیت یک برنامه ارزیابی یعنی تناسب با اهداف آن. بنابراین تعریف اهداف ارزیابی، بدون توجه به محیط آموزشی (به عنوان مثال، آموزش سنتی یا یادگیری مبتنی بر حل مسأله) یا عملکرد خاص ارزیابی (به عنوان مثال، ابزاری برای یادگیری یا تصمیم‌گیری برای ارائه مدرک)، یکی از مهمترین اصول برنامه ارزیابی است و باید در ابتدای طراحی برنامه مشخص شود.

تعداد و جزییات اهداف می‌تواند متنوع باشد ولی بهتر است یک هدف به عنوان هدف اصلی برنامه در نظر گرفته شود. بهتر است اهداف کوتاه مدت و طولانی مدت برنامه ارزیابی هر چند محدود تعریف شود. به عنوان مثال، برنامه ارزیابی که در دانشگاه ماستریخ[5] طراحی و اجرا شد و در انتهای فصل حاضر به آن اشاره می‌شود، با هدف ارتقای یادگیری تدوین شده است.

1. Supporting
2. Documenting
3. Improving
4. Justifing
5. Maastricht

۲. تعیین زیرساخت‌های برنامه ارزیابی

فرصت‌ها و محدودیت‌های در اختیار برای برنامه ارزیابی باید در مراحل اول طراحی برنامه تعیین شوند و تصمیمات بعدی متناسب با فرصت‌ها و محدودیت‌ها اتخاذ شوند. علاوه بر این، زیر ساخت‌های مورد نیاز و متناسب با اهداف برنامه تامین شوند.

۳. تعیین ذی‌نفعان برنامه ارزیابی

ذی‌نفعان برنامه ارزیابی باید شناسایی شوند، تخصص‌های ایشان مورد استفاده واقع شود و نقش‌هایی که باید ایفا کنند مشخص شود. همچنین سطح مشارکت ذی‌نفعان در برنامه ارزیابی متناسب با اهداف برنامه ارزیابی و نیاز ایشان تعیین می‌شود.

۴. تهیه یک طرح جامع ارزیابی

منظور از طرح جامع برای ارزیابی یک ساختار جامع بر اساس چارچوب توانمندی است. این طرح جامع اهمیت زیادی دارد زیرا در برنامه ارزیابی، قضاوت بر اساس یک ارزیابی واحد صورت نمی‌گیرد و اطلاعات ارزیابی واحد با هم ترکیب می‌شود تا در مورد پیشرفت در یک توانمندی یا یک نقش اطلاعات در اختیار قرار دهد. بنابراین، طرح جامع باید نقشه‌ای از روش‌های واحد ارزیابی برای چارچوب توانمندی و نیز برنامه درسی فراهم کند. به این صورت که دلیل انتخاب یک روش ارزیابی خاص و هدفمند بودن انتخاب محتوای آن در یک بخش از برنامه درسی و در یک برهه زمانی مشخص باشد. بنابراین، بسته به نوع برنامه درسی و هر مرحله آن طرح جامع ارزیابی مشتمل بر تنوعی از محتوای ارزیابی، روش‌های ارزیابی استاندارد و غیر استاندارد و عناصر ارزیابی مقطعی و طولی است. برای هر انتخاب باید نقش آن در برنامه جامع و هم‌راستایی آن با برنامه درسی مشخص شود (ون‌درولوتن و همکاران ۲۰۱۴).

۵. تعیین نحوه و روش جمع‌آوری اطلاعات

جمع‌آوری اطلاعات به معنی فعالیت‌های مربوط به جمع‌آوری انواع مختلف اطلاعات کمی و کیفی درباره توانایی آزمون‌شونده است. موضوعات مربوط به این حوزه شامل محتوای ارزیابی، انتخاب ساختار آزمون، استفاده از ابزارها، سیستم نمره‌دهی و برنامه زمان‌بندی ارزیابی است. عامل اصلی در انتخاب ابزارهای ارزیابی اهداف برنامه، قابلیت اجرای آن (زیر ساخت‌ها) و قابلیت پذیرش آن (ذی‌نفعان) است.

در انتخاب ابزار، محتوایی که در ابزار گنجانده می‌شود، بسیار مهمتر از نوع ابزار است. همچنین باید نمونه‌گیری مناسبی برای پوشش ارزیابی فرایندهای شناختی، رفتاری و نگرشی در سطوح مورد نظر انجام شود. علاوه بر این، در انتخاب ابزار باید تاثیری که ابزار بر رفتار آزمون‌شونده خواهد داشت، مدنظر قرار گیرد، هزینه آن در نظر گرفته شود و در صورت وجود جایگزین مناسب‌تر از آن استفاده شود، و قبل از اجرا در محیط به طور آزمایشی مورد استفاده واقع شود. در برنامه ارزیابی باید ارتباط بین ابزارهای ارزیابی مختلف مشخص باشد. در مجموع، در برنامه ارزیابی تاکید بر یک ابزار به صورت جداگانه نیست، بلکه بر نقشی که هر ابزار در کل برنامه ایفا می‌کند، تاکید می‌شود.

شرایط اجرای روش‌های ارزیابی مانند زمان و مکان اجرای آن، و محتوای آن مانند دشواری، پیچیدگی و واقعی بودن آن باید با هدف روش ارزیابی انتخاب شده متناسب باشد. برنامه‌ریزی زمانی ارزیابی باید به گونه‌ای باشد که زمان کافی برای آموزش و یادگیری مطالب وجود داشته باشد.

۶. تلفیق اطلاعات جمع‌آوری شده

تلفیق اطلاعات به دست آمده از روش‌های ارزیابی مختلف باید معنادار و بر اساس اهداف و محتوای روش‌های ارزیابی باشد. تلفیق اطلاعات شامل دو جنبه فنی و مفهومی است. جنبه فنی به معنی ترکیب اطلاعات از منابع متعدد و ترکیب انواع مختلف داده‌ها است، ولی نباید منجر به از دست رفتن سطح اندازه‌گیری یک ابزار شود و باید نتایج و اثرات تلفیق اطلاعات بر همه ذی‌نفعان مشخص شود.

۷. ارزش‌گذاری اطلاعات

استفاده از ابزارهای مختلف منجر به جمع‌آوری داده‌های غنی از منابع مختلف می‌شود. به منظور اقدام در مورد مجموعه اطلاعات غنی و متنوع لازم است این اطلاعات تفسیر شوند تا ارزش افزوده‌ای نسبت به اطلاعات مربوط به یک ابزار داشته باشند. جنبه مفهومی جمع‌آوری اطلاعات به استفاده از اطلاعات ترکیب شده شامل تفسیر، ارزش‌گذاری و انتخاب داده‌ها برمی‌گردد. ارزش گذاری شامل تعیین حد نصاب قبولی و همچنین تعیین نقاط قوت و ضعف آزمون‌شوندگان و اولویت بندی اهداف یادگیری بر اساس اطلاعات به دست آمده از ارزیابی است.

۸. اقدام مناسب بر اساس نتایج ارزیابی

اقدامات نیز باید متناسب با هدف ارزیابی و عواقب آن توجیه پذیر باشد. نحوه دسترسی ذی‌نفعان به اطلاعات و ارائه بازخورد به آن‌ها مشخص باشد. این مرحله به معنی تکمیل و بستن حلقه است که می‌تواند از اقداماتی مانند ادامه/عدم ادامه تا ارائه بازخورد یا حتی دوره‌های جبرانی متفاوت باشد. این مرحله نتایج را به ارزیابی متصل می‌کند.

۹. حمایت برنامه ارزیابی

هر چند عناصر برنامه در عمل برای استقرار یک سیستم ارزیابی کفایت می‌کند ولی کیفیت آن را تضمین نمی‌کند. بخشی از فعالیت‌های تضمین کیفیت به حمایت برنامه در حال اجرا بر می‌گردد. لازم است یک مدیریت مرکزی برای هماهنگ نمودن اجزای برنامه ارزیابی در نظر گرفته شود. علاوه بر مدیریت مرکزی برنامه ارزیابی نیاز به رهبری آموزشی دارد. ضروری است لیدر برنامه ارزیابی از ویژگی‌های زیر برخوردار باشد:

- هر آنچه از دیگران انتظار انجام آن را دارد (مانند اختصاص زمان برای مشاهده مستقیم و بازخورد)، خود آن را انجام دهد. این موضوع موجب اعتبار فرایند ارزیابی و تسهیل تغییر می‌شود.
- به روش‌شناسی ارزیابی و بازخورد آگاه باشد.
- همکاری نزدیکی با اعضای هیات علمی، دانشجویان، و دیگر ذی‌نفعان سیستم ارزیابی مانند مدیران، کارکنان، پرستاران، بیماران و افرادی که به نوعی در برنامه ارزیابی نقشی دارند داشته باشد.
- نتایج منفی ارزیابی را جدی بگیرد.

علاوه بر این، دو مضمون دیگر مربوط به حمایت، می‌تواند کیفیت برنامه ارزیابی یا تناسب آن با اهداف را تأمین کند.

- حمایت فنی که مرتبط با کیفیت مواد ارزیابی است. این حمایت فنی می‌تواند شامل فعالیت‌هایی قبل از اجرای آزمون (مثل پانل مرور سؤالات، توانمند سازی اعضای هیأت علمی) و پایش بعد از ارزیابی (تحلیل روانسنجی و تحلیل‌های دیگر) باشد. به این دلیل که موفقیت یک ارزیابی به میزان زیادی به استفاده‌کنندگان آن بستگی دارد توانمندسازی ارزیابان به منظور ارتقای کیفیت برنامه ارزیابی از اهمیت زیادی برخوردار است. گروه‌های مختلفی از اعضای هیات علمی، مسؤول دوره یا برنامه آموزشی، مدیر گروه آموزشی و کمیته ارزیابی گرفته تا دستیاران، دانشجویان پزشکی و پرستاران ممکن است در برنامه ارزیابی به عنوان ارزیاب همکاری داشته باشند. ضروری است ارزیابان به نقش ارزیابی

خود علاقمند باشند و دانش، مهارت و نگرش مورد نیاز و زمان و فرصت کافی برای ایفای آن را داشته باشند. اغلب ساعت‌ها وقت صرف آموزش ارزیابی مواجهه با بیمار استاندارد شده بر اساس یک چک لیست روا و پایا می‌شود، در حالی‌که ارزیابان برای ایفای نقش ارزیابی خود برای قضاوت در مورد تعامل پیچیده فراگیر با بیمار در محیط‌های غیر قابل پیش‌بینی طبابت آموزشی نمی‌بینند. حمایت فنی همچنین به دانش، مهارت و نگرش مورد نیاز برای طراحی و اجرای یک سیستم ارزیابی با تأثیر آموزشی اشاره دارد. با وجود این، حمایت فنی به تنهایی کافی نیست و یک سیستم ارزیابی با حمایت فنی می‌تواند به دلیل مقاومت ذی‌نفعان با خطر شکست مواجه شود.

- دومین مضمون مرتبط با حمایت برنامه ارزیابی، حمایت قانونی و سیاسی آن به منظور افزایش قابلیت پذیرش ارزیابی است. این امر با جلب مشارکت ذی‌نفعان به وقوع می‌پیوندد. مشارکت ذی‌نفعان در طراحی برنامه ارزیابی نه تنها احتمال طرح عقاید خلاق را مطرح می‌کند، بلکه تناسب برنامه را با محیط عمل افزایش می‌دهد. مشارکت ذی‌نفعان به آن‌ها حس مالکیت می‌دهد و به این وسیله حمایت آن‌ها را جلب می‌کند. تمامی پروسیجرها و پروتکل‌های ارزیابی باید مشخص باشند و به ذی‌نفعان به ویژه فراگیران اطلاع‌رسانی شود. همچنین در نظر گرفتن رویه استیناف و اعتراض به منظور پرهیز از ارزیابی ناعادلانه یکی از روش‌های حمایت از برنامه است. مسایل و ملاحظات قانونی ملی و بین المللی نیز باید لحاظ شوند و میزان آزادی در طراحی برنامه ارزیابی را متأثر می‌سازند. اقدامات حمایتی تأثیر فوری بر اقدامات ارزیابی در حال اجرا دارند و این دو بعد برنامه ارزیابی یک فرایند چرخه‌ای را با هدف بهینه سازی سیستم ارزیابی درونی شکل می‌دهند.

تجربه دانشگاه اوتاگو [1] (ویلکینسون [2] و همکاران ۲۰۱۱)

در دانشگاه اوتاگو در کشور نیوزلند برنامه ارزیابی در مقطع پزشکی عمومی طراحی و اجرا شد. در این برنامه دستورالعمل‌های شفاف در خصوص اجزای مختلف برنامه ارزیابی تدوین شد. به عنوان مثال، تصمیمات نهایی در خصوص عملکرد دانشجو به صورت قبول، رد و مشروط بود. معیارهای وضعیت مشروط به صورت زیر تعریف شدند:

۱. عدم احزار وضعیت مطلوب مورد نیاز در یک روش ارزیابی (به عنوان مثال، OSCE یا تکالیف)، که در این صورت دانشجو باید ارزیابی را تکرار کند یا در ماژول بعدی نشان دهد که به وضعیت مطلوب دست یافته است.

۲. عدم آمادگی برای جلسات بررسی عملکرد در مهلت تعیین شده یا عدم حضور در جلسات که در این صورت باید در ماژول‌های بعدی در مهلت مقرر موارد مربوط را رعایت کند.

در مجموع، موفقیت در یک حوزه عملکردی نمی‌تواند جبران کننده ضعف عملکرد در حوزه‌های دیگر باشد. انتظار می‌رود دانشجو حد نصاب قبولی در همه اجزای ارزیابی پایانی مربوط به یک بلوک، از جمله تعهد حرفه‌ای را کسب کند. ضعف در یک حوزه عملکردی (به عنوان مثال، نقص دانش) می‌تواند در ماژول بعدی توسط همان حوزه عملکردی (و نه مهارت بالینی) جبران شود.

1. Otago
2. Wilkinson

۱۰. مستند نمودن فعالیت‌های برنامه ارزیابی

مستندسازی به دو منظور تسهیل یادگیری سازمانی و افزایش شفافیت برنامه انجام می‌شود. بنابراین همه عناصر مربوط به اقدامات اجرایی و حمایتی شامل مسؤولیت‌ها، حقوق، تعهدات، قوانین و مقررات باید ثبت شوند تا فرایند ارزیابی بدون ابهام و قابل دفاع باشد. در این ارتباط سه عامل اهمیت ویژه دارند:

- اولین عامل، **محیط و شرایط یادگیری** است. برنامه ارزیابی در خلا اجرا نمی‌شود. محیط و قابلیت کاربرد برنامه ارزیابی باید به صورت شفاف توصیف شود. ذی‌نفعان باید بتوانند وجود و نحوه تأثیر برنامه ارزیابی را بر خود تعیین کنند. ارتباط برنامه ارزیابی با سیستم آموزشی باید مشخص شود.
- دومین عامل، **قوانین و مقررات** است که مبنایی برای ذی‌نفعان فراهم می‌کند تا هدف ارزیابی و وظایف و تعهدات ذی‌نفعان مختلف را در ارتباط با برنامه در حال اجرا و اقدامات حمایتی بدانند. مستندسازی شفافِ مقررات موجب پیشگیری از طفره رفتن افراد از مسؤولیت‌هایشان می‌شود.

- سومین عنصر تهیه **بلوپرینت**[1] یا **جدول مشخصات ارزیابی** است که محتوای برنامه ارزیابی و ابزارهای مورد استفاده را به تصویر می‌کشد. به طور قطع، در طراحی ارزیابی در هر سطحی محتوا بخشی از معادله است. محتوای اختصاصی بر کلیت فرایند طراحی تأثیری ندارد ولی قویاً با اهداف ارزیابی ارتباط دارد، در نتیجه باید به عنوان منبعی برای آینده مستند شود.

۱۱. ارتقای برنامه ارزیابی

تا این‌جا فعالیت‌های ارتقای برنامه در ابعاد حمایتی و مستند سازی ارائه شد. اما نوع دیگری از فعالیت‌های مربوط به ارتقای برنامه در پاسخ به ارزیابی نقادانه از دیدگاه‌هایی دورتر به برنامه است. فعالیت‌ها در این بعد عموماً تأثیر فوری بر برنامه‌های اخیراً در حال اجرا ندارد، بلکه تأثیر آن‌ها در زمان طراحی مجدد تمام یا بخشی از برنامه معمولاً در زمانی دورتر آشکار می‌شود. اکثر فعالیت‌های ارتقا شامل فعالیت‌های پژوهشی و توسعه‌ای با هدف ارزشیابی دقیق برنامه در راستای تعیین جنبه‌های مشکل‌دار آن و تکمیل چرخه است.

۱۲. توجیه‌پذیر بودن فعالیت‌های برنامه ارزیابی

در حالی که ابعاد قبلی چارچوب برنامه ارزیابی به جنبه‌های درونی موسسه یا سازمان مسؤول برنامه ارزیابی مربوط بود، توجیه‌پذیر بودن برنامه به تقاضای در حال افزایش برای پاسخگویی عمومی مربوط است. به منظور توجیه برنامه ارزیابی چهارنوع فعالیت باید در نظر گرفته شود:

- انجام **پژوهش‌های علمی** به منظور تأمین شواهدی برای حمایت از عمل و کاربرد دیدگاه‌های علمی جدید در برنامه ارزیابی یکی از فعالیت‌ها است.
- توجیه‌پذیری همچنین نیازمند **مرور بیرونی** برنامه ارزیابی توسط متخصصان به صورت بازدید از محل و شنیدن دیدگاه‌های ذی‌نفعان برنامه است. مرور بیرونی بیشتر با هدف اعتباربخشی یا مقایسه با برنامه‌های ارزیابی مشابه صورت می‌پذیرد.
- به دلیل محدودیت منابع، **هزینه–اثربخشی** نیز یکی از اهداف مطلوب برنامه ارزیابی است. به بیان دیگر کیفیت برنامه ارزیابی در میزان دستیابی به اهداف به میزان در دسترس بودن منابع بستگی دارد. به منظور توجیه برنامه همه هزینه‌های برنامه باید مشخص شود و هزینه–سودمندی برنامه در سایه اهداف آن تعیین شود و در صورت وجود جایگزین‌های بهتر تغییر مورد نیاز انجام شود.
- علاوه بر این برنامه ارزیابی باید **قابلیت پذیرش سیاسی و قانونی** داشته باشد. بنابراین، باید قوانین بالادستی یا قوانین موجود را نقض نکند و اطلاعات محرمانه بماند.

تجربه دانشگاه علوم پزشکی تهران (گندم‌کار و همکاران ۱۳۹۳):

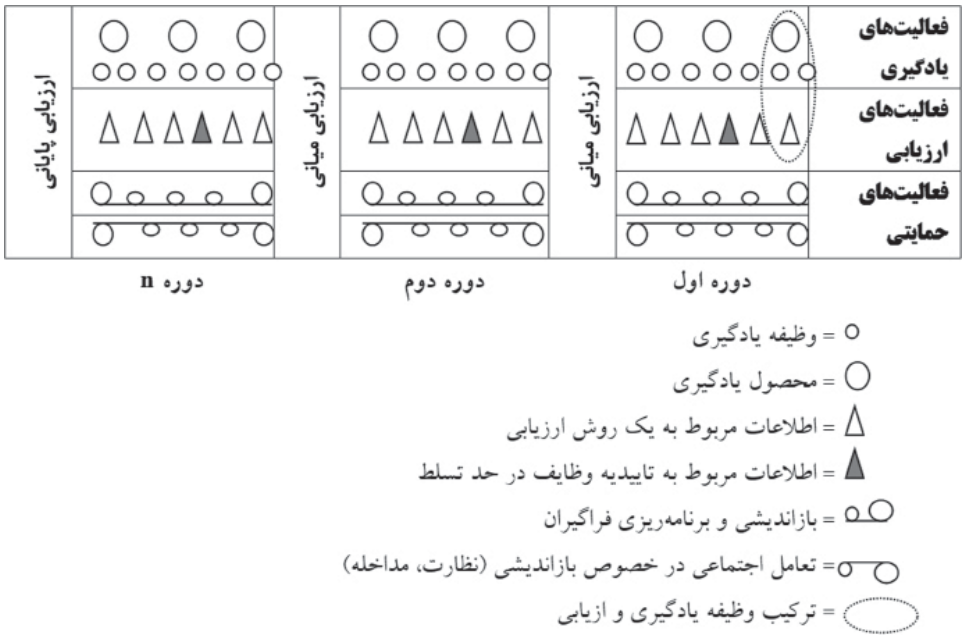
رویکرد اخیر دانشگاه علوم پزشکی تهران به آموزش مبتنی بر توانمندی ایجاب کرد ارزیابی دانشجو به گونه‌ای طراحی شود که در راستای آموزش ارائه شده باشد. بنابراین، به منظور فراهم آوردن چارچوبی برای طراحی نظام ارزیابی در دانشکده‌های دانشگاه علوم پزشکی، مرکز مطالعات و توسعه آموزش دانشگاه اقدام به «تدوین آیین‌نامه نظام جامع ارزیابی» دانشجو نمود. آیین‌نامه مذکور در سه بخش «ساختار و تشکیلات ارزیابی دانشجو»، «طراحی و اجرای نظام ارزیابی دانشجو» و «تضمین کیفیت نظام ارزیابی دانشجو» و در ۲۴ بند تدوین شد.

1. Blueprint

نمونه‌های عملی نظام ارزیابی

در این بخش نمونه‌هایی از برنامه‌های ارزیابی که بر اساس اصول ذکر شده در قسمت پیش طراحی شده‌اند، ارائه می‌شود.

ون‌درولوتن و همکاران (۲۰۱۲) مدلی را برای نظام ارزیابی پیشنهاد دادند که هدف اصلی آن ارزیابی برای یادگیری بود و آزمون‌های اعطای مدرک را شامل نمی‌شد. این مدل برای هر برنامه آموزشی، برنامه‌هایی که با آموزش نظری سر و کار دارند یا برنامه‌هایی که مبتنی بر آموزش در محل کار هستند مانند برنامه‌های دستیاری مناسب است. مدل مذکور شامل سه عنصر اصلی فعالیت‌های یادگیری[1]، فعالیت‌های ارزیابی[2] و فعالیت‌های حمایت فراگیران[3] است (شکل ۲-۳۷).



شکل ۲-۳۷: مدل ارزیابی برنامه‌ای در عمل (ون‌درولوتن و همکاران ۲۰۱۲)

تجربه دانشکده پزشکی، دانشگاه ماستریخ (دریسن و همکاران ۲۰۱۲)

در دانشگاه ماستریخ، از مدل نظام ارزیابی استفاده شد تا برنامه ارزیابی مبتنی بر محل کار را در سال آخر کارآموزی طراحی کنند. با توجه به این‌که یادگیری در محیط بالینی نسبت به سال‌های قبل از آن کمتر ساختارمند است، بنابراین در طراحی سیستم ارزیابی نیز باید بر موانع خاصی غلبه کرد. هر چند اهداف و توانمندی‌های انتهای دوره بالینی از قبل تعیین می‌شود ولی واقعیت این است که فرصت‌های یادگیری به تنوع بیماران و حضور اساتید بالینی بستگی دارد. بنابراین مهمترین چالش در طراحی سیستم ارزیابی در محیط بالینی آن است که علیرغم تفاوت در محیط‌های بالینی بتواند همه دانشجویان را به سمت سطح توانمندی مورد نظر جهت دهد و فرصت‌های کافی را فراهم کند که نشان می‌دهد آن‌ها به توامندی مورد نظر دست یافته‌اند. بنابراین برنامه ارزیابی باید با چارچوب توانمند کلی هم راستا باشد و همچنین تفاوت‌های محیط یادگیری را نیز دربرداشته باشد. برنامه ارزیابی برای کارآموزان سال ۶ اجرا شد که در آن، هر چرخش بالینی ۱۸ هفته طول می‌کشد. هر دانشجو یک منتور دارد که مسؤول نظارت بر یادگیری و ارزیابی وی است. منتور و دانشجو مسؤولیت مراقبت از بیمار را به عهده دارند که مسؤولیت دانشجو در مراقبت از بیمار نسبت به کارآموزی‌های قبلی افزایش پیدا می‌کند.

1. Learning activities
2. Assessment activities
3. Learner support activities

در ادامه اجزای مدل شرح داده می‌شود و نمونه‌هایی از تجربه‌های استفاده از آن ذکر خواهد شد[۱]:

۱. فعالیت‌های یادگیری

واحد تشکیل‌دهنده فعالیت‌های یادگیری، وظیفه یادگیری[۲] است. هر چیزی که می‌تواند منجر به یادگیری شود یک وظیفه یادگیری است: یک جلسه سخنرانی، یک جلسه کار عملی، مواجهه با بیمار، یک جلسه حل مسأله، یک پروژه، تکالیف یادگیری یا مطالعه توسط فراگیر.

وظایف یادگیری به نوبه خود در راستای یک برنامه درسی منسجم که بر اساس اصول طراحی آموزشی تدوین شده است هستند. برخی از وظایف یادگیری می‌توانند منجر به تولید آثار و محصولات یادگیری شوند. برخی از محصولات یادگیری مانند گزارش یک پروژه پیامدی هستند و برخی دیگر مانند فهرستی از پروسیجرهای جراحی انجام شده در اتاق عمل فرایندی هستند.

تجربه دانشکده پزشکی، دانشگاه ماستریخ (دریسن و همکاران ۲۰۱۲)

- فعالیت‌های یادگیری

ساختار برنامه ارزیابی در این دانشکده بر اساس چارچوب توانمندی‌های CanMEDS است. برای هر توانمندی یک روبریک طراحی شده است که در آن سطح دستیابی هر توانمندی در انتهای دوره و عملکردهای پایین‌تر و بالاتر از سطح مورد نظر توصیف شده‌اند. به منظور هم‌راستا نمودن فعالیت‌های یادگیری یا محیط‌های یادگیری مختلف و نیز نیازهای یادگیری فردی، دانشجویان یک برنامه یادگیری با مشاوره منتور خود قبل از این‌که دوره شروع شود آماده می‌کنند. مهم این است که برنامه یادگیری در راستای توانمندی‌های کسب شده توسط دانشجو در دوره کارآموزی قبلی و فرصت‌های یادگیری اختصاصی کارآموزی فعلی باشد.

تجربه دانشکده دامپزشکی، دانشگاه یوتریچ[1] (بوک و همکاران ۲۰۱۳)

- فعالیت‌های یادگیری
  - برنامه بالینی (سال‌های ۴، ۵ و ۶) بر اساس چارچوب توانمندی حرفه‌ای دامپزشکی سازمان‌دهی شده است.
  - برنامه تقریبا شامل ۱۷ چرخش بالینی است.
  - دانشجویان وظایف بالینی را در بیمارستان وابسته به دانشگاه یا در کلینیک‌های خارج از دانشگاه انجام می‌دهند.
  - دانشجویان در تیم‌هایی متشکل از دانشجویان با سطوح مختلف تجربه کار می‌کنند.
  - ۳۰ درصد از زمان کل آموزش به جلسات سخنرانی و سمینارها در خصوص موضوعات و موارد بیماری خاص اختصاص دارد.
  - ۳۵ درصد زمان دانشجویان نیز به مطالعه موارد بالینی به منظور تعمیق بینش ایشان انجام می‌شود.

1. Utrecht

۲. فعالیت‌های ارزیابی

اولین فعالیت‌های ارزیابی در برنامه ارزیابی، روش‌های ارزیابی واحدِ مرتبط با هر یک از سطوح هرم میلر هستند. روش‌های ارزیابی می‌تواند شامل آزمون کتبی، OSCE، مشاهده مواجهه بالینی (مانند Mini-CEX)، ارزشیابی توسط همگنان در جلسات یادگیری مبتنی بر حل مسأله و ... یا ارزیابی محصولات منتج از وظایف یادگیری باشد. نمونه‌هایی از مورد آخر، ارزیابی مستندات مربوط به بررسی بیمار، تهیه شده توسط فراگیران یا ارزیابی نحوه ارائه نتایج کارهای پژوهشی وی است. در زمان طراحی فعالیت‌های ارزیابی باید نکات زیر را مد نظر قرار داد:

- همه فعالیت‌های ارزیابی باید چنان منظم و سازمان‌دهی شوند تا از یادگیری مستمر دانشجو حمایت کنند. بنابراین، باید معنادار باشند و نسبت به عملکرد فراگیرد بازخورد غنی از اطلاعات، چه کمی و چه کیفی، فراهم کنند.
- اطلاعات فعالیت‌های ارزیابی باید مستند شوند و قابل پیگیری باشند.
- اگر چه هر روش ارزیابی در خصوص عملکرد دانشجو بازخورد ارائه می‌دهد، ولی از حساسیت کمی برخوردار است

۱. با توجه به این‌که یکی از روش‌های اصلی مستند سازی برنامه ارزیابی کارپوشه است، بخش عمده‌ای از تجربیات مربوط به اجرای برنامه ارزیابی در فصل کارپوشه ارائه شده است.
2. Learning task

و در مورد تصمیم‌گیری در مورد ردی یا قبولی فراگیران بر اساس نتیجه یک ارزیابی واحد باید با احتیاط برخورد کرد. تنها استثنا آزمون‌های در حد تسلط هستند که بر اساس نتیجه یک آزمون هم می‌توان در مورد دستیابی یا عدم دستیابی فراگیران به مهارت مورد نظر در حد تسلط تصمیم گیری کرد.

- در این مدل، هر روش ارزیابی واحد جزئی از برنامه طولی روش‌های ارزیابی است.
- نمره نباید به عنوان تنها شکل از بازخورد به فراگیران در نظر گرفته شود. به این دلیل که دانشجو متوجه نمی‌شود چه چیزی را و چگونه یاد گرفته است. در اغلب موارد مدرسان به دلیل عینیت نمره و در نتیجه کافی دانستن آن از دادن بازخورد پرهیز می‌کنند.
- در صورت امکان از فناوری اطلاعات مانند کارپوشه الکترونیکی و ... در برنامه ارزیابی استفاده شود.
- روش‌های ارزیابی فی نفسه مزیتی بر یکدیگر ندارند، بلکه سودمندی آن‌ها وابسته به عملکردشان در برنامه ارزیابی است.
- قضاوت خبرگان در برنامه ارزیابی غیر قابل اجتناب است که می‌تواند معلم، مربی، همگنان، بیمار یا حتی خود دانشجو باشد.

تجربه دانشکده پزشکی، دانشگاه ماستریخ (دریسن و همکاران ۲۰۱۲)

- فعالیت‌های ارزیابی

فعالیت‌های ارزیابی به گونه‌ای طراحی شده است که اطمینان حاصل شود که دانشجویان در حوزه‌های مختلف توانمندی بازخورد مناسب با تاکید بر بازخورد کیفی دریافت کنند. همه فعالیت‌های ارزیابی در ارزیابی نهایی سهیم است. به این دلیل که فعالیت‌های ارزیابی طوری طراحی شده است که متناسب با همه رشته‌ها و محل‌هایی که توسط دانشجویان انتخاب می‌شود باشد، ابزارهای ارزیابی کلی هستند و با رشته‌ها و محل‌های مختلف متناسب هستند. در طراحی فعالیتهای ارزیابی امکانات و محدودیت‌های کارآموزی، کاربری آسان و کارایی و قابلیت اجرای آسان در محیط بالینی پر از ازدحام در نظر گرفته شده است. تاکید بر کارا بودن، سادگی و دانشجو محوری است. ابزارهای ارزیابی ساده و کوتاه با بخش‌های عددی محدود و فضایی برای بازخوردهای کیفی طراحی شده‌اند. حداقل ۵ فعالیت ارزیابی که لایه‌های مختلف هرم میلر را پوشش می‌دهد جزء الزامات است:

- پنج آزمون mini-CEX
- دو ارزیابی ۳۶۰ درجه (MSF)
- دو ارزیابی نقادانه یک موضوع[1] (CAT)
- دو آزمون پیشرفت تحصیلی
- یک آزمون OSCE

در ابتدای کارآموزی کارپوشه کارآموزان، تنها شامل برنامه یادگیری است و در طول دوره بخش‌های مختلفی به آن افزوده می‌شود: ارزیابی‌ها، ارزیابی میانی، گزارشات جلسات پیشرفت کار، بازاندیشی دانشجویان بر فرایند یادگیری خود و سایر مواردی که دانشجویان تمایل دارند به آن اضافه کنند. مواردی که در بالا آماده است جزء حداقل موارد است و منتور و دانشجویان می‌توانند مواردی را به آن اضافه کنند. وظیفه دانشجویان است که ارزیابی ها را از منابع مختلف در موقعیت‌های متفاوت به منظور دستیابی به نمونه‌های متنوع از ارزیابان و موارد جمع‌آوری کنند.

1. Clinical Appraisala of a Topic

تجربه دانشکده دامپزشکی، دانشگاه یوتریچ (بوک و همکاران ۲۰۱۳)

- فعالیت‌های ارزیابی

ابزارهای زیر به منظور ارزیابی توانمندی دانشجویان و متناسب با چارچوب توانمندی‌ها مورد استفاده قرار گرفت:

- mini-CEX
- MSF
- گزارش موارد بیماری مبتنی بر شواهد[1] (EBCR)

ابزارهای ارزیابی استاندارد شامل سؤالات با مقیاس لیکرت ۵ تایی و فضایی برای ارائه بازخورد کتبی مورد استفاده قرار می‌گیرند.
از کارپوشه آنلاین ساختارمند بر اساس توانمندی‌های حرفه دامپزشکی به منظور مستند سازی نتایج ارزیابی‌ها استفاده می‌شود.
سالانه دو آزمون پیشرفت تحصیلی با هدف ارزیابی مهارت‌های استدلال بالینی مورد استفاده قرار می‌گیرند.

1. Evidence Based Case Reports

۳. فعالیت‌های حمایتی

فعالیت‌های حمایتی در یک دوره به دو گونه است. اول، فراگیران بر اطلاعات به دست آمده از فعالیت‌های یادگیری و ارزیابی بازاندیشی می‌کنند. دوم، اساتید راهنما یا مشاور با ارائه بازخورد از یادگیری خود راهبر حمایت می‌کنند. ممکن است فعالیت‌های بازاندیشی در ابتدا و انتهای دوره بیشتر باشند ولی فعالیت یادگیری خودراهبر مداوم است. به این صورت که بازخوردها تفسیر می‌شوند و به منظور برنامه‌ریزی برای اهداف یا وظایف یادگیری جدید استفاده می‌شوند. حمایت از یادگیری خود راهبر بر منتورینگ و مربی‌گری استاد راهنما یا مشاور استوار است، ولی می‌تواند توسط فراگیران سال بالاتر یا همگنان نیز فراهم شود.

تجربه دانشکده پزشکی، دانشگاه ماستریخ (دریسن و همکاران ۲۰۱۲)

• فعالیت‌های حمایتی

دانشجویان توسط منتور خود و از طریق کارپوشه حمایت می‌شوند. منتور از میان استادان بخش انتخاب می‌شود. انتظار می‌رود دانشجو و منتور سه جلسه مرور پیشرفت، در هفته ۴، ۸ و ۱۲ به منظور پایش پیشرفت دانشجو ترتیب دهند. بحث بر اساس اطلاعات کارپوشه و خود ارزیابی دانشجو از عملکرد خود در برابر برنامه یادگیری و میزان دستیابی به توانمندی‌های CanMEDS شروع می‌شود. سپس منتور بازخورد ارائه می‌دهد و بازاندیشی دانشجویان را هدایت می‌کند. بر اساس نتایج ملاقات، برنامه یادگیری اصلاح می‌شود.

تجربه دانشکده دامپزشکی، دانشگاه یوتریچ (بوک و همکاران ۲۰۱۳)

• فعالیت‌های حمایتی
- انتظار می‌رود دانشجویان بر اطلاعات به دست آمده از فعالیت‌های یادگیری و ارزیابی بازاندیشی کنند.
- دانشجویان بر اساس بازخوردی که دریافت می‌کنند، نقاط ضعف و قوت خود را مشخص می‌کنند و سپس سؤالات یادگیری را مطرح می‌کنند.
- سؤالات یادگیری در جلسات گروه همگنان با شش دانشجوی دیگر و با یک معلم بالینی (منتور) مورد بحث قرار می‌گیرد.
- بحث انجام شده منجر به شکل‌گیری اهداف یادگیری برای دوره بعدی می‌شود.
- کل فرایند بر اساس برنامه رشد فردی تدوین شده بر اساس چارچوب توانمندی‌ها هدایت می‌شود.

۴. ارزیابی میانی

در پایان هر دوره، همه محصولات مربوط به فعالیت‌های یادگیری، اطلاعات ارزیابی و اطلاعات انتخاب شده مربوط به فعالیت‌های حمایتی تحت عنوان ارزشیابی میانی میزان پیشرفت ارزیابی می‌شوند. تمام اطلاعات جمع شده از تمام ارزیابی‌های به عمل آمده در طول دوره در برابر مجموعه‌ای از استانداردهای عملکرد، توسط گروهی از ارزیابان مستقل و معتبر به عنوان مثال، کمیته ارزیابان بررسی می‌شود. در این مرحله متخصصان بر اساس داده‌های کمی و کیفی قضاوت می‌کنند. تجمیع و مقایسه نمرات آزمون‌های پیشرفت تحصیلی در طول زمان می‌تواند یکی از منابع اطلاعات کمی باشد که قابلیت پیشگویی عملکرد آینده فرد را بر اساس عملکرد گذشته وی نیز دارد. هر چند برخی از اطلاعات ارزیابی مانند اطلاعات ثبت شده در پرونده بیمار، کیفی و نقلی هستند و نیازمند تفسیر انسانی اطلاعات می‌باشند. داده‌های ارزیابی ترجیحا باید به روشی معنی‌دار تجمیع شوند. به طور سنتی، از روش‌های ارزیابی (یا سطوح هرم میلر) برای تجمیع اطلاعات استفاده می‌شود، ولی تقسیم‌بندی‌های معنی‌دار دیگری، مانند موضوعات برنامه درسی یا چارچوب توانمندی نیز قابل تصور هستند.

فرایند ارزیابی میانی باید از استحکام و دقت کافی برخوردار باشد. استفاده از افراد متخصص و مطلع در زمینه قضاوت و ارزشیابی، آموزش ایشان و استفاده از ابزارهای حمایتی مانند روبریک و استانداردهای عملکرد از جمله راهبردهای مؤثر در این زمینه است. افراد متخصصی که عضو کمیته هستند به مرور زمان و با کسب تجربه می‌توانند پروسیجرها و ابزارهای حمایتی را اصلاح کنند. یکی دیگر از عواملی که بر دقت ارزشیابی میانی تأثیر گذار است وسعت بحث‌های انجام شده در کمیته است. در اکثر موارد، در صورتی که اطلاعات روش‌های ارزیابی واحد از ثبات کافی و سطح مناسب برخوردار باشند،

فرایند ارزیابی سریع و کاراً خواهد بود. هر چند برای برخی از فراگیران کمیته نیازمند بحث، مشورت و استدلال بیشتری است.

این تصمیم‌گیری‌ها نه تنها اطلاعاتی را در مورد دستیابی به استانداردهای عملکرد فراهم می‌آورد، بلکه دارای ارزش تشخیصی (با ارائه نقاط قوت و ضعف فراگیر)، درمانی (پیشنهاد اصلاحی به منظور دستیابی به نتایج مورد نظر) و پیش آگهی (پیش‌بینی عملکرد آینده فرد) است. نکته قابل توجه این است که ارزیابی میانی دارای رویکردی اصلاحی است و این نکته مهم نظام ارزیابی را از انواع سنتی ارزیابی که به طور معمول مبتنی بر رسیدن به حد تسلط است تمایز می‌دهد؛ اگر دستیابی به حد تسلط رخ ندهد فراگیران می‌بایست دوره را دوباره بگذرانند و مجدداً مورد ارزیابی قرار گیرند. در ارزیابی برنامه ریزی شده رویکرد در وهله اول توسعه‌ای است؛ به این صورت که توصیه‌هایی برای یادگیری بیشتر بر اساس اطلاعات غنی به دست آمده از عملکرد دانشجو و متناسب با هر دانشجو ارائه می‌شود. ارزیابی کمیته از حساسیت و اهمیت متوسطی برخوردار است. به این صورت که هر چند اطلاعات ارزیابی مذکور برای فراگیر عواقبی از لحاظ ردی یا قبولی ندارد ولی نباید نادیده گرفته شوند. لازم است فراگیران از آن برای برنامه‌ریزی برای یادگیری بیشتر استفاده کنند.

یکی از چالش‌های ارزیابی میانی، نحوه استفاده از اطلاعات سیستم حمایتی است. افرادی مانند مربی و منتور که طولانی مدت با فراگیر در تعامل بوده‌اند غنی‌ترین اطلاعات را در خصوص فراگیر در اختیار دارند. اگر تصمیم‌گیری و قضاوت در مورد اطلاعات حمایتی به افراد دخیل در سیستم حمایتی سپرده شود ممکن است به ارتباط بین حمایت کننده و فراگیر خدشه وارد کند، و در طولانی مدت تاثیر سیستم حمایتی را کاهش دهد. از سوی دیگر، اگر تصمیم گیری تنها به کمیته ارزیابی سپرده شود یک دیوار دفاعی غیر قابل نفوذ بین فعالیت‌های حمایتی و تصمیم‌گیری ایجاد می‌شود که موجب می‌شود کمیته ارزیابی از اطلاعات با ارزشی محروم شود. در این صورت ارزیابان برای به دست آوردن اطلاعات باید تلاش بیشتری کنند، اطلاعات به دست آمده ممکن است همراه با سوگیری باشد و همچنین می‌تواند هزینه بیشتری را به دنبال داشته باشد.

یک راه حل میانی برای رفع معضل مذکور این است که مربی یا منتور از فراگیر اظهارنامه‌ای مبنی بر این که اطلاعات حمایتی تصویر معتبری از فراگیر فراهم می‌کند اخذ کند و سپس در مورد عملکرد وی قضاوت کند. در یک گام جلوتر، ممکن است از منتور خواسته شود که توصیه‌های خود را در مورد تصمیمات گرفته شده در مورد عملکرد در کمیته ارزیابی ارائه دهد و فراگیر نیز فرصت اظهار نظر و احتمالاً اصلاح آن را داشته باشد. به طور خلاصه، یک راهبرد واحد به عنوان راه حل معضل ارزیابی میانی وجود ندارد و راه حل‌ها بسته به منابع در دسترس، منطق، احساسات، فرهنگ و میزان حساسیت تصمیم‌گیری متفاوت است.

در مجموع، یک چرخه در مدل وندرولوتن و همکاران (۲۰۱۲) شامل فعالیت‌های آموزشی، ارزیابی و فعالیت‌های حمایتی است. این چرخه می‌تواند به صورت نامحدود تکرار شود. تعداد چرخه‌ها به ماهیت برنامه آموزشی و در دسترس بودن منابع بستگی دارد. در مدل ارائه شده، در هر سال تحصیلی چرخه سه بار تکرار می‌شود. هر دوره شامل چندین دوره درسی است. مهمترین موضوع، پیشرفت منطقی طولی فراگیران از طریق وظایف یادگیری، بازخورد مناسب و خودراهبری حمایت شده است. این در نقطه مقابل رویکرد مبتنی بر حد تسلط خالص قرار دارد که در آن قبولی در یک آزمون به معنی کسب توانمندی مادام العمر است. همچنین نکته مهم این است که اطلاعات کافی در مورد عملکرد و اصلاحات مورد نیاز باید قبل از تصمیم‌گیری‌های مهم نهایی رخ دهد.

تجربه دانشکده پزشکی، دانشگاه ماستریخ (دریسن و همکاران ۲۰۱۲)

• ارزیابی میانی
با توجه به این‌که امکان جمع‌آوری اطلاعات در مدت زمان کوتاه برای دانشجویان امکان‌پذیر نبود، تنها یک ارزیابی میانی در هفته ۴ به عمل آمد. به این دلیل که ارزیابی همه دانشجویان (۳۴۰نفر) توسط یک کمیته امکان‌پذیر نبود منتورها در مورد عملکرد دانشجویان در دوره و امکان ادامه دوره بر اساس اطلاعات موجود در کارپوشه و بر اساس قضاوت خود از عملکرد دانشجو تصمیم‌گیری کردند.

۵. ارزیابی پایانی

بعد از تکرار تعداد مناسبی از چرخه‌های وظایف یادگیری، فعالیت‌های ارزیابی و حمایتی و ارزیابی میانی، یک ارزیابی نهایی در همان زمانی که تصمیم‌گیری در مورد پیشرفت فراگیران در جریان است رخ می‌دهد. این یک تصمیم‌گیری حساس با عواقب مهم برای فراگیران است. تصمیم‌گیری توسط همان کمیته‌ای که ارزیابی میانی را انجام داده است، ولی با پروسیجرهای تا حد امکان دقیق‌تر صورت می‌گیرد. به این دلیل که این افراد به طور مرتب با نتایج عملکرد دانشجویان سر و کار داشته‌اند و با سوابق آن‌ها آشنایی دارند و در نتیجه می‌توانند قضاوت قطعی‌تری ارائه دهند.

مثال‌هایی از پروسیجرهای دقیق‌تری که برای ارزیابی پایانی به کار می‌روند عبارتند از: پروسیجرهای استیناف، استفاده از اطلاعات ورودی منتورها و مربیان، و فراگیران، آموزش و استاندارد نمودن قضاوت ارزیابان، استفاده از استانداردها و روبریک‌های سنجش عملکرد، افزایش تعداد افراد کمیته، گسترش بحث، مشورت و مستندسازی در فرایند تصمیم‌گیری کمیته، و در نهایت ورود همه اطلاعات دوره‌های قبلی از جمله ارزشیابی میانی.

در حالت ایده‌آل، تصمیمات مربوط به ارزشیابی پایانی باید قابل دفاع باشد. تصمیم‌گیری تنها محدود به رد یا قبولی نیست، بلکه برتری خاص در عملکرد را نیز نشان می‌دهد. هر چند لزومی به جزئی نمودن قضاوت در سطح نمرات نیست. طبقه‌بندی بیشتر عملکرد به صورت نمرات احتمال خطا در طبقه‌بندی و مشکلات قضاوت را بیشتر می‌کند.

اگر سیستم ارزیابی خوب عمل کند، تصمیمات اخذ شده موجب تعجب فراگیران (یا مربیان و منتورها) نمی‌شود. در موارد کمی، تصمیم اخذ شده با انتظارات فراگیر در تقابل است و تکرار وقوع این نوع تصمیمات ممکن است به اعتبار کمیته خدشه وارد سازد. بر اساس ماهیت پیشرفت دانشجو، کمیته می‌تواند توصیه‌هایی را برای آموزش و اصلاح عملکرد ارائه دهد. در مجموع، تصمیم نهایی دقیق است و بر اساس اطلاعات غنی و آزمون‌های متعدد است. اگر تصمیمات با چالش مواجه شود باید قابل دفاع باشد. مدلی که در شکل ۲ نشان داده شده است یک دوره یادگیری مشخص را به تصویر می‌کشد که با تصمیم‌گیری در مورد ارتقای فراگیر به اتمام می‌رسد ولی یک برنامه درسی را به طور کامل نشان نمی‌دهد. با توجه به برنامه درسی، دوره یادگیری می‌تواند به صورت چرخه‌ای تکرار شود تا برنامه درسی را کامل کند. لزومی به برابر بودن طول هر چرخه نیست و تعداد و طول چرخه‌ها با توجه به ماهیت برنامه درسی و زمان ارزیابی و تصمیم‌گیری نهایی می‌تواند متفاوت باشد.

تجربه دانشکده پزشکی، دانشگاه ماستریخ (دریسن و همکاران ۲۰۱۲)

• ارزیابی پایانی
مسؤول ارزیابی پایانی، کمیته ارزیابی دانشکده پزشکی است. به این دلیل که این ارزیابی مهمترین ارزیابی سال آخر است، چندین شاخص برای تضمین کیفیت این ارزیابی مهم در نظر گرفته شده است:

۱. کمیته از روبریک ارزیابی که در آن استانداردهای عملکرد در هر توانمندی تعریف شده است استفاده می‌کند.
۲. دانشجویان و منتورها پیشنهادات خود را در ارتباط با ارزیابی به کمیته ارسال می‌کنند. به این منظور فرمی تعبیه شده است که مشابه فرم‌های مورد استفاده توسط کمیته است. بنابراین کمیته می‌تواند ارزیابی خود را با توجه به نظرات دانشجویان و منتورها تکمیل کند.
۳. چرخه بازخورد بر اساس ارزیابی میانی و جلسات بررسی پیشرفت تکمیل می‌شود.
۴. پروسیجر استیناف برای دانشجویان در نظر گرفته شده است.
۵. آموزش و هماهنگ نمودن ارزیابان به صورت دوره‌ای انجام می‌شود.
۶. کمیته ارزیابی از تعداد محدود اعضا (۶ نفر) تشکیل شده است.

تجربه دانشکده دامپزشکی، دانشگاه یوتریچ (بوک و همکارن ۲۰۱۳)

- فعالیت‌های ارزیابی
  - ارزیابی دوره‌ای پس از گذشت شش ماه از آموزش بالینی، توسط کمیته مستقل مرور کارپوشه انجام می‌شود.
  - ارزیابی بر اساس استانداردهای عملکرد از پیش تعیین شده انجام می‌شود.
  - اطلاعات ارزیابی مربوط به شش ماه به منظور یک ارزیابی روا و پایا با هم تجمیع می‌شود.
  - فرم‌های استانداردی برای قضاوت توسط کمیته مرور کارپوشه استفاده می‌شود.
  - یک قضاوت کیفی به فراگیران ارائه می‌شود که در صورت لزوم با پیشنهادات اصلاحی همراه است.
  - تعداد حداقل مورد نیاز از هر یک از ابزارهای ارزیابی برای ارزیابی دوره‌ای (۶ ماه) عبارتند از:
    - ۱۲ آزمون mini-CEX (مشاهده توسط همگنان یا معلمان)
    - یک مورد ارزیابی ۳۶۰ درجه
    - ۲ مورد گزارش موارد بیماری مبتنی بر شواهد (EBCR)
    - ۲ مورد برنامه پیشرفت فردی[1] (PDP)
    - ۲ مورد آزمون پیشرفت تحصیلی

1. Personal Development Plan

مدل ارائه شده در بالا باعث ارتقای کیفیت می‌شود و ارزش یادگیری را در برنامه ارزیابی به حداکثر می‌رساند. بدون این‌که خدشه‌ای به معنی‌دار بودن اطلاعات برنامه ارزیابی وارد شود، تصمیم‌گیری نهایی از دقت و اعتبار کافی برای پاسخگویی درونی و بیرونی در برابر کیفیت دانش‌آموختگان برخوردار است. از دیگر اهداف برنامه ارزیابی، ارائه اطلاعات به منظور ارزشیابی فرایند و پیامد برنامه درسی است.

تجربه دانشگاه علوم پزشکی تهران (ایمانی‌پور و همکاران ۲۰۱۳):

- طراحی، اجرا و ارزشیابی برنامه ارزیابی در واحد کارورزی مراقبت‌های ویژه دانشکده پرستاری

مرحله طراحی: به منظور طراحی برنامه ارزیابی مراحل زیر طی شد:

۱. ابتدا فهرست کاملی از اهداف آموزشی و رفتاری دوره کارورزی ویژه، بر اساس سرفصل دروس پرستاری ویژه مصوب شورای عالی برنامه‌ریزی درسی تهیه شد.

۲. اهداف بر اساس نظر تعدادی از اساتید با تجربه در امر تدریس نظری و عملی دروس پرستاری ویژه از دانشگاه‌های مختلف شهر تهران و تعدادی از پرستاران با تجربه شاغل به کار
در بخش‌های ویژه به سه گروه اهداف مربوط به مهارتهای بالینی،
دانش شناختی و ویژگی‌های حرفه‌ای‌گری تقسیم شدند.

۳. روش‌های ارزیابی مناسب برای هر یک از این گروه‌های اهداف به کمک بررسی متون علمی مرتبط تعیین شد و از پانلی از متخصصان آموزش پزشکی و ارزشیابی و اساتید پرستاری خواسته شد تا مناسب‌ترین روش ارزیابی برای هر گروه از اهداف را با توجه به معیارهای وندرولوتن «کاربردی بودن، مقرون به صرفه بودن، داشتن تأثیر آموزشی بر یادگیری و عملکرد فراگیران» مشخص کنند. بعد از بحث، تبادل نظر و وزن‌دهی به روش‌های ارزیابی توسط اعضای پانل، مناسب‌ترین روش ارزیابی برای هر گروه از اهداف یادگیری مشخص گردید. نتیجه اجماع نظر اعضای پانل عبارت بود از: آزمون شفاهی برای اهداف دانش شناختی، فرم درجه بندی گلوبال[1] (GRF) و آزمون مشاهده مستقیم مهارت‌های پروسیجرال[2] (DOPS) برای اهداف مهارت‌های بالینی و فرم نمونه کار بالینی[3] (CWS) برای ارزشیابی ویژگی‌های حرفه‌ای‌گری.

۴. در گام بعد، ابزار ارزیابی هر یک از روش‌های تعیین شده به کمک بررسی متون طراحی شد. پس از چندین بار بررسی و بازنگری توسط متخصصان آموزش پزشکی، فرم نهایی آنها آماده و روایی محتوایی آنها به روش محاسبه نسبت و شاخص اعتبار محتوا تأیید گردید.

مرحله اجرا: قبل از اجرای روش ارزیابی برنامه‌ای، ابتدا نحوه استفاده از هر ابزار و به کارگیری این روش، طی یک جلسه توجیهی و به طور مجزا برای دانشجویان و اساتید توضیح داده شد. ضمن این که راهنما و دستورالعمل به کارگیری آن به صورت کتبی نیز در اختیار آزمون‌گر (استادان) و آزمون‌شونده (دانشجویان) قرار گرفت. این روش ارزیابی، اولاً همراه با ارائه بازخورد به دانشجو بود و ثانیاً هر دو نوع ارزیابی تکوینی و پایانی را شامل می‌شد. لازم به توضیح است پایایی این روش نیز بعد از اجرا در عرصه و به روش تعیین همسانی درونی تأیید گردید (آلفای کرونباخ ۰/۹۰). کلیه دانشجویان شرکت کننده در مطالعه، در طول گذراندن دوره کارورزی بخش‌های ویژه با این روش، ارزیابی می‌شدند.

مرحله ارزشیابی برنامه: بعد از اتمام دوره کارورزی نظرات اساتید و دانشجویان راجع به روش ارزیابی بالینی برنامه‌ای به کمک پرسشنامه محقق ساخته بررسی شد. پرسشنامه اساتید شامل ۱۰ سؤال و مربوط به دانشجویان شامل ۸ سؤال با مقیاس لیکرت پنج‌تایی بود. این پرسشنامه همچنین شامل چند سؤال باز پاسخ در مورد نقاط قوت و ضعف روش ارزیابی طراحی شده بود. در مجموع ۸۷/۵ درصد اساتید و ۹۷/۳ درصد دانشجویان با استفاده از این روش ارزیابی برای سنجش عملکرد بالینی دانشجویان موافق بودند. نقاط قوت این روش از نظر دو گروه عبارت بودند از: عینی بودن، پرهیز از سوگیری، ارائه بازخورد، پوشش دقیق اهداف یادگیری و اختصاصی شدن ارزیابی. ۸۷/۵ استادان و ۸۹/۴۷ درصد دانشجویان معتقد بودند ارزیابی با رویکرد برنامه‌ای بر فرایند یادگیری تأثیر آموزشی مثبت داشته است.

1. Global Rating Form
2. Direct Observation of Procedural Skills
3. Clinical Work Sampling

چالش‌ها و فرصت‌های نظام ارزیابی

برخی از مهمترین چالش‌های نظام ارزیابی عبارت هستند از (وندرولوتن و همکاران ۲۰۱۲):

- اولین چالش مهم، منابع مورد نیاز برای اجرای برنامه ارزیابی از جمله منابع مالی است. به منظور کاهش هزینه، انجام تعداد کمتری از فعالیت‌ها با کیفیت بالا بهتر از انجام تعداد زیاد آن با کیفیت پایین است. دلیلی برای جمع‌آوری داده‌های وسیع که اطلاعاتی در اختیار ما قرار نمی‌دهند، وجود ندارد. در صورتی که برنامه ارزیابی جزئی جداناپذیر از برنامه یادگیری باشد، مانند آن‌چه در الگوهای این فصل ارائه شد، در صرفه‌جویی منابع کمک خواهد کرد. برخی از فعالیت‌های ارزیابی، به ویژه انواع با حساسیت کمتر می‌توانند با هزینه کمتری انجام شود. برای مثال، در برخی حوزه‌ها می‌توان بانک سؤال را به صورت اینترنتی در اختیار دانشجویان قرار داد تا دانش خود را در حوزه‌های مشخص ارزیابی کنند. علاوه بر این، به اشتراک گذاشتن مستندات آزمون‌ها بین موسسات آموزشی راهبرد مناسبی است. برخی از توانمندی‌ها مانند تعهد حرفه‌ای یا مهارت برقراری ارتباط از طریق ارزیابی توسط همگنان به خوبی ارزیابی می‌شود. همچنین می‌توان بین عناصر مشخصی در مدل یا در دوره‌های خاصی از برنامه درسی بسته به تعادل بین اهمیت و حساسیت ارزیابی و منابع، مصالحه ایجاد کرد. برای مثال، منتورینگ یا مربی‌گری می‌تواند در بخش‌های مشخصی در برنامه درسی و نه در همه موارد انجام شود.
- دومین چالش بزرگی که باید مستقیماً با آن روبه رو شد بروکراسی، بیهودگی و تقلیل‌گرایی است. بیهودگی یا بی‌اهمیت شدن آزمون از این نظر مهم است که همه جا در کمین است. به محض این‌که یک ابزار ارزیابی، راهبرد ارزیابی یا پروسیجر ارزیابی از هدف اصلی که برای آن طراحی شده است مهم‌تر شود، بیهودگی چهره زشت خود را نشان می‌دهد. فراگیران ترفندهایی را برای قبولی در امتحان اجرا می‌کنند، معلمان فرم‌ها را با یک حرکت قلم پر می‌کنند (بنا بر ضرورت اجرایی و بر اساس قضاوت بدون مفهوم)، مدیران بدون هیچ دلیلی به این دلیل که همیشه این‌گونه بوده است یا به دلیل سیاست‌های مؤسسه بر پروسیجر متمرکز می‌شوند (تنها نمرات را می‌خواهند زیرا نمرات عینی و پاسخگوی جامعه هستند). به محض این‌که ما متوجه مبادله نمونه سؤالات در بازار سیاه یا منابع اینترنتی برای آمادگی سریع در امتحان می‌شویم مطمئن می‌شویم که فرایند ارزیابی دچار بیهوده‌گرایی شده است. همه دست‌اندرکاران در برنامه ارزیابی باید وظایف خود و دلایل انجام آن را درک کنند. با این‌حال خطر از دست دادن بینش درست نسبت به هدف واقعی ارزیابی و گیر افتادن در پروسیجرهای بروکراتیک و فعالیت‌های بدون مفهوم وجود دارد. اداره کردن و مقابله با بیهوده‌گرایی احتمالاً سخت‌ترین ولی ضروری‌ترین وظیفه‌ای است که به عنوان مدافع نظام ارزیابی باید انجام دهیم. به منظور پیشگیری از بروکراسی، باید اجرای فرایند را تسهیل کنیم. یکی از روش‌ها می‌تواند استفاده از فناوری کامپیوتری باشد.
- سومین چالش، محدودیت‌های قانونی است. برنامه‌های درسی باید منطبق بر آیین‌نامه‌های دانشگاه و مقررات ملی باشند. این آیین‌نامه‌ها و قوانین معمولاً بسیار محافظه‌کارانه و در راستای رویکرد سنتی به ارزیابی هستند که به ارزیابی مبتنی بر حد تسلط به یادگیری و قضاوت در مورد عملکرد فراگیران بر اساس نمرات و امتیازات بها می‌دهد. بسیاری از این موارد با اصول نظام ارزیابی در تضاد است.
- چالش چهارم، جدید بودن و ناشناخته‌بودن نظام ارزیابی است. برنامه ارزیابی بسیار متفاوت از ارزیابی‌های پایانی است که برای اکثر ما از نظر تجربه فردی به عنوان معلم و فراگیر آشنا است. بسیاری این ایراد را به برنامه ارزیابی وارد می‌دانند که رویکرد بسیار ملایمی به ارزیابی، به ویژه به دلیل تکیه آن بر اطلاعات قضاوت غیرعینی، دارد. این وظیفه اجرا کنندگان برنامه ارزیابی است که نشان دهند فرایند تصمیم‌گیری کاملاً دقیق است و همه دست اندرکاران دلیل و هدف کارشان را می‌دانند.

تجربه دانشکده پزشکی، دانشگاه ماستریخ (دریسن و همکاران ۲۰۱۲)

• ارزشیابی برنامه ارزیابی
بین سال‌های ۲۰۰۷ تا ۲۰۰۹، ۶۷۰ دانشجو در برنامه شرکت کردند. همه دانشجویان پرسشنامه‌ای که مربوط به ارزیابی درک دانشجویان از برنامه ارزیابی بود را تکمیل نمودند. معیارهای ارزشیابی عبارت بودند از:
۱. احتمال مشارکت در فعالیت‌های یادگیری
۲. بازخورد از طریق ابزارهای ارزیابی
۳. دقت ارزیابی پایانی
۴. کیفیت فعالیت‌های حمایتی
نتایج نظرسنجی انجام شده مثبت بود. نتایج بحث گروهی با دانشجویان نیز نشان داد دلایل موفقیت سیستم ارزیابی عبارت بودند از:
۱. ارائه بازخورد به دانشجو و استفاده از آن به منظور هدایت یادگیری به کمک منتورها
۲. ساختار ساده برنامه و اجرای آسان برای دانشجویان و مدرسان
۳. حجم کم روش‌های ارزیابی مورد استفاده
۴. آزادی عمل دانشجویان برای موقعیت‌های مورد مشاهده و ارائه بازخورد
۵. استفاده از داده‌های کیفی در فرایند ارزیابی
۶. آشناسازی اعضای هیات علمی و دانشجویان با فرایند ارزیابی

از نظر تئوری، فرصت‌ها در رویکرد نظام ارزیابی بسیار بیشتر هستند. هر چند این ادعاها باید با طراحی و اجرای برنامه‌های ارزیابی بیشتر در عمل ثابت شوند. برنامه ارزیابی، علاوه بر این‌که برای ارزیابی با هدف یادگیری ارزش قائل است نشان می‌دهد که استفاده از ارزیابی با هدف یادگیری و تصمیم‌گیری در مورد عملکرد فراگیران به صورت همزمان امکان‌پذیر است.

تجربه دانشکده دامپزشکی، دانشگاه یوتریچ (بوک ۲۰۱۳)

• ارزشیابی برنامه ارزیابی
۸۵ دانشجو در سال ۲۰۱۰ وارد دوره بالینی شدند و برنامه ارزیابی به صورت آزمایشی برای آن‌ها اجرا شد. ۱۹۸ پرسشنامه بین سال‌های ۲۰۱۱ و ۲۰۱۲ توسط دانشجویان در خصوص کیفیت برنامه ارزیابی تکمیل شد. سپس ۱۸ دانشجو و ۸ مدرس بالینی در جلسات بحث گروهی که با هدف ارزشیابی کیفیت برنامه ارزیابی برگزار شد، شرکت کردند.
نتایج نشان داد با وجود اقرار دانشجویان به ارزش روش‌های ارزیابی مبتنی بر محل کار در ارائه بازخورد و ارتقای یادگیری، آن‌ها ارزیابی‌های مذکور را به عنوان تراکمی در نظر می‌گرفتند. دانشجویان اظهار داشتند در نهایت همه ارزیابی‌ها در ارزیابی نهایی تاثیر داشت. علاوه بر این، برنامه در دستیابی به هدف خود یعنی «ارزیابی برای یادگیری» و نه «ارزیابی یادگیری» موفقیت کامل نداشت. مدرسان بالینی اعتقاد داشتند به دلیل محدودیت‌های زمانی قادر به ارائه بازخورد با کیفیت به دانشجویان نبودند. دانشجویان، به دلیل در نظر داشتن آزمون‌های مبتنی بر محل کار به عنوان آزمون تراکمی و نیز مشغله اساتید، از درخواست ارزیابی‌های همراه با بازخورد امتناع می‌ورزیدند. بازخورد توسط همگنان به عنوان ابزار باارزشی در نظر گرفته شد.

منابع

1. Beard J, Strachan A, Davies H, Patterson F, Stark P, Ball S, et al. Developing an education and assessment framework for the Foundation Programme. Med Educ. 2005;39(8):841–51.
2. Bok HG, Teunissen PW, Favier RP, Rietbroek NJ, Theyse LF, Brommer H, et al. Programmatic assessment of competency-based workplace learning: when theory meets practice. BMC Med Educ. 2013;13(1):1–9.
3. Dijkstra J, Galbraith R, Hodges BD, McAvoy PA, McCrorie P, Southgate LJ, et al. Expert validation of fit-for-purpose guidelines for designing programmes of assessment. BMC Med Educ. 2012;12(1):20.
4. Dijkstra J, Van der Vleuten C, Schuwirth L. A new framework for designing programmes of assessment. Adv Heal Sci Educ. 2010;15:379–93.
5. Driessen EW, van Tartwijk J, Govaerts M, Teunissen P, van der Vleuten CP. The use of programmatic assessment in the clinical workplace: a Maastricht case report. Med Teach. 2012;34(3):226–31.
6. Holmboe ES, Hawkins RE. Practical guide to the evaluation of clinical competence. Philadelphia: Mosby/Elsevier; 2008.
7. Imanipour M, Jalili M, Mirzazadeh A, Dehghan Nayeri N, Haghani H. Viewpoints of Nursing Students and Faculties about Clinical Performance Assessment Using Programmatic Approach. Iranian Journal of Medical Education. 2013;12(10):743–55
8. Gandomkar R, Jalili M, Mirzazadeh A. Developing comprehensive student assessment guidelines: The first step towards programmatic approach to assessment in Tehran University of Medical Sciences. Iranian Journal of Medical Education. 2015;14(12):1107–10.
9. Schuwirth LW, Van der Vleuten CP. Programmatic assessment: From assessment of learning to assessment for learning. Med Teach. 2011;33(6):478–85.
10. Smith SR, Dollase RH, Boss JA. Assessing students' performances in a competency-based curriculum. Acad Med. 2003;78(1):97–107.
11. Van der Vleuten CPM, Schuwirth LWT, Driessen EW, Dijkstra J, Tigelaar D, Baartman LKJ, et al. A model for programmatic assessment fit for purpose. Med Teach. 2012;34(3):205-14.
12. van der Vleuten CPM, Dannefer EF. Towards a systems approach to assessment. Med Teach. 2012; 34(3); 185-186.
13. Van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: from methods to programmes. Med Educ. 2005;39(3):309-17.
14. van der Vleuten CP, Schuwirth LWT, Driessen EW, Govaerts MJB, Heeneman S.. 12 Tips for programmatic assessment. Med Teach. 2015;37(7):641-6.
15. Wilkinson TJ, Tweed MJ, Egan TG, Ali AN, McKenzie JM, Moore M, et al. Joining the dots: Conditional pass and programmatic assessment enhances recognition of problems with professionalism and factors hampering student progress. BMC Med Educ. 2011;11(1):29.

فصل ۳۸

ارزشیابی نظام ارزیابی

مقدمه

با تغییر رویکرد به ارزیابی دانشجو، ضرورت بازنگری و تغییر معیارهای بررسی کیفیت ارزیابی نیز احساس می‌شود. اگر چه در بخش‌های قبل کتاب دیدیم که نظریه تست کلاسیک و روش‌های سنتی ارزیابی مورد انتقاد قرار گرفته‌اند ولی روش‌ها و رویکردهای جدید نیز بدون مشکل نیستند. بنابراین، لازم است کیفیت برنامه ارزیابی بر اساس معیارهای کیفی متناسب با اصول زیر بنایی آن مورد آزمایش قرار گیرد.

یکی از اصول زیربنایی ارزشیابی برنامه ارزیابی این است که هر دو نوع معیارهای کیفیت سنتی و جدید برای ارزیابی مورد نیاز هستند زیرا برنامه ارزیابی شامل هر دو نوع روش‌های ارزیابی سنتی و جدید است. حال این سؤال مطرح می‌شود که آیا روش‌های روان‌سنجی مرسوم می‌توانند کیفیت برنامه ارزیابی را تضمین کنند. به بیان دیگر، آیا مفهوم پایایی و روایی به همان روشی که برای آزمون‌های سنتی به کار می‌رود، برای برنامه ارزیابی نیز کاربرد دارد یا خیر.

برخی متخصصان معتقدند که اصول زیر بنایی نظریه تست کلاسیک می‌تواند به منظور سنجش کیفی توانمندی‌ها به کار رود. در حالی که برخی دیگر معتقدند روش‌های آماری مورد استفاده در این دیدگاه مناسب نیستند و باید به دنبال معیارهایی متناسب با اصول زیربنایی نظام ارزیابی باشیم. به طور خاص در خصوص پایایی باید گفت هر چند، ایده پایایی به خودی خود اشتباه نیست ولی همان‌طور که قبلاً اشاره شد، مشکلاتی در استفاده از مفهوم پایایی بر اساس دیدگاه سنتی در برنامه ارزیابی وجود دارد:

- اولاً به طور سنتی، پایایی یعنی ثبات اندازه‌گیری در موقعیت‌های مکرر با آزمون شونده ثابت یا ارتباط یک سؤال واحد با کل آزمون. مورد اول مناسب اندازه‌گیری ویژگی‌های ثابت است در حالی که توانمندی ویژگی در حال تغییر است. مورد دوم در آزمون‌های طولانی که از سؤالات مجزا تشکیل شده است کاربرد دارد، در حالی که در ارزیابی عملکرد کلی یک فعالیت مصداق ندارد.
- ثانیاً اغلب پایایی با مفاهیم عینیت و استانداردسازی به صورت مترادف به کار می‌رود. این در حالی است که در دیدگاه نظام ارزیابی پایایی می‌تواند حتی با آزمون‌های کمتر استاندارد و قضاوت‌های بیشتر ذهنی مانند مشاهده عملکرد به دست آید.

در مجموع، ایده پایایی در برنامه ارزیابی از اهمیت خاصی برخوردار است، ولی لازم است به روشی متفاوت از دیدگاه کلاسیک تعریف و بررسی شود (بارتمن[1] و همکاران ۲۰۰۶). به عنوان مثال، پایایی بین ارزیابان و همچنین ثبات بین قضاوت در مواجهه‌های مختلف در برنامه ارزیابی اهمیت زیادی دارد.

در مورد روایی شرایط به وضوح پایایی نیست. به این دلیل که تعاریف متنوعی از روایی و تقسیم‌بندی‌های مختلفی برای

1. Baartman

آن وجود دارد. یکی از انواع معروف این تقسیم‌بندی‌ها را مسیک[1] (۱۹۹۴) ارائه داده است. در چارچوب مسیک، روایی سازه[2] مطرح‌کننده مفهوم اصلی روایی است و شش روایی دیگر شامل روایی محتوا[3]، روایی اساسی[4]، روایی ساختاری[5]، روایی پی‌آیندی[6]، روایی بیرونی و تعمیم‌پذیری است. در حال حاضر، مفهوم روایی در برنامه ارزیابی فراتر از ارزیابی سازه مورد نظر است و مواردی مانند استفاده صحیح از ابزار و نحوه استفاده ابزار توسط استفاده‌کنندگان را نیز در بر می‌گیرد (بارتمن و همکاران ۲۰۰۷).

لین[7] (۱۹۹۱) معتقد است که با ظهور ابزارهای جدید ارزیابی و همچنین تغییر پارادایم حاکم بر ارزیابی دانشجو و مشکلاتی که مفاهیمی مانند روایی و پایایی برای استفاده در سیستم ارزیابی توانمندی دارند منطقی است که مفاهیم مذکور به صورت دیگری عملیاتی شوند، یا با معیارهای دیگری که برای روش‌های جدید پیشنهاد شده‌اند، جایگزین شوند و همچنین معیارهای بیشتری در نظر گرفته شوند. در این بخش برخی از رویکردهایی را که به منظور کیفیت برنامه ارزیابی پیشنهاد شده‌اند، مرور می‌کنیم.

بیگز[8] (۱۹۹۶) با طرح مفهوم هم‌راستایی سازنده[9]، ضرورت هم‌راستا بودن اهداف یادگیری، آموزش و ارزیابی را بیان می‌کند. بر این اساس لازم است در یک برنامه آموزشی مبتنی بر توانمندی، ارزیابی دانشجو نیز در راستای ارزیابی میزان دستیابی به توانمندی‌ها باشد. هر چند به لحاظ نظری هم‌راستایی سازنده باید در همه ابعاد یک ارزیابی مبتنی بر توانمندی یا برنامه ارزیابی جاری باشد ولی در عمل با تهیه بلوپرینت ارزیابی، در حد انطباق محتوای برنامه ارزیابی با اهداف برنامه یا توانمندی‌ها باقی مانده است (وب[10]، ۲۰۰۷). رویکرد دیگر به کیفیت برنامه ارزیابی کاربرد رویکرد سنتی روان‌سنجی ولی با تمرکز بر ترکیب ابزارها بوده است که به نوبه خود منجر به تدوین چارچوبی برای تحلیل برنامه ارزیابی با تمرکز بر «دیدگاه یکپارچه به روایی» یا «پایایی ترکیبی» شده است (واس و همکاران[11] ۲۰۰۱).

هیچ‌یک از موارد ذکر شده در بالا رویکردی جامع به برنامه ارزیابی نداشته‌اند. هر چند کیفیت روان‌سنجی مهم است، ولی قابلیت اجرای برنامه ارزیابی، هزینه و اثربخشی آن و محیط و شرایطی که در آن ارزیابی رخ می‌دهد نیز از اهمیت زیادی برخوردار هستند. لین و همکاران (۱۹۹۱) معیارهایی مانند نتایج، انتقال و تعمیم‌پذیری، عدالت، پیچیدگی شناختی، معنی‌داری، کیفیت محتوا، پوشش محتوا و هزینه و مقرون به صرفه بودن را مطرح می‌کنند. آلنبک[12] (۲۰۰۲) معیارهای دیگری را اضافه بر معیارهای کلاسیک شامل واقعی بودن، کیفیت محتوا، پوشش محتوایی، قابلیت مقایسه، تاثیر و قابلیت اجرا پیشنهاد می‌کند (بارتمن ۲۰۰۶).

برخی معتقد هستند استفاده از معیارهای کیفیت به کاربرد آزمون، یعنی تکوینی یا تراکمی بودن، آن بستگی دارد. در ارزیابی تکوینی، مقایسه‌پذیر بودن و تعمیم پذیری اهمیت کمتر و کارایی از این نظر که بازخورد در اغلب موارد و به صورت مؤثر ارائه می‌شود، اهمیت بسیاری دارد. در حالی‌که در ارزیابی تراکمی، پروسیجرهای استاندارد برای اطمینان از مقایسه‌پذیری، قابلیت تکرار و رعایت عدالت لازم است. انتقال این مفاهیم به برنامه ارزیابی به این معنی است که لازم نیست هر روش ارزیابی به تنهایی در برنامه ارزیابی همه معیارهای کیفیت را برآورده کند. هر چند در مورد معیارهای مقایسه‌پذیری و قابلیت تکرار در مورد ارزیابی تکوینی با سهولت بیشتری برخورد می‌کنیم، ارزیابی تراکمی باید همه معیارهای کیفیت از جمله معیارهای مربوط به یادگیری و بازخورد مانند معنی‌دار بودن و نتایج آموزشی را برآورده کند. در مجموع می‌توان گفت کیفیت برنامه ارزیابی در کل در نظر گرفته می‌شود. به این معنی که یک برنامه ارزیابی باید در مجموع همه معیارهای کیفیت را برآورده کند. برای مثال، نمره بالا در صحت و اعتبار نمی‌تواند نقص‌های عمده در پیچیدگی شناختی را جبران کند.

---

1. Messick
2. Construct validity
3. Content validity
4. Substantial validity
5. Structural validity
6. Consequential validity
7. Linn
8. Biggs
9. Constructive alignment
10. Webb
11. Wass etal.
12. Uhlenbeck

معیارهای کیفیت نظام ارزیابی

هر چند در قسمت قبل مروری کوتاه بر برخی از معیارهای کیفیت برنامه ارزیابی شد اما در مورد نحوه تعیین کیفیت یک برنامه ارزیابی شناخت کمی وجود دارد و شواهد بیشتری به منظور بررسی سودمندی برنامه‌های ارزیابی مورد نیاز است. یکی از مدل‌هایی که به منظور تحلیل برنامه ارزیابی و تضمین کیفیت آن ارائه شده است توسط بارتمن و همکاران در سال ۲۰۰۶ تحت عنوان «چرخ ارزیابی توانمندی[1]» مطرح شد. معیارهایی که در این مدل به عنوان معیارهای کیفیت برنامه ارزیابی مطرح شده است عبارتند از (شکل ۱-۲۶):

۱. تناسب با هدف

تناسب با هدف[2] در مرکز چرخ ارزیابی توانمندی قرار دارد و مبنایی برای طراحی کل برنامه ارزیابی است. این معیار قابل قیاس با مفهوم «هم‌راستایی سازنده» است که باید کل برنامه ارزیابی با اهداف یادگیری و آموزش ارائه شده هم‌راستا باشد.

لایه بعدی و داخلی چرخ شامل معیارهای کیفیتی همچون قابلیت مقایسه[3]، قابلیت تکرار تصمیمات[4]، مقبولیت[5] و شفافیت[6] است.

۲. قابلیت مقایسه

قابلیت مقایسه به معنی انجام برنامه ارزیابی به روشی با ثبات و پاسخگو است. موقعیت‌هایی که در آن ارزیابی انجام می‌شود باید تا حد امکان برای همه فراگیران یکسان باشد و نمره‌گذاری با استفاده از معیارهای یکسان برای همه فراگیران انجام شود. روش‌های افزایش قابلیت مقایسه شامل نمونه‌گیری دقیق از شرایط مختلف، و نمونه‌گیری وسیع از محتوا، و موقعیت‌های توانمندی‌های مهم است. قابلیت مقایسه معادل پایایی در دیدگاه کلاسیک است.

۳. قابلیت تکرار تصمیمات

قابلیت تکرار تصمیمات به معنی دقیق و پایدار بودن تصمیمات اخذ شده بر اساس نتایج برنامه ارزیابی در طول زمان و بین ارزیابان متفاوت است. این به این معنی نیست که برنامه باید عینی باشد. در ارزیابی عملکرد، ارزیابان به صورت ذهنی عملکرد فراگیران را مورد قضاوت قرار می‌دهند. مهم این است که تصمیمات دقیق باشند و به ارزیابان یا موقعیت ارزیابی خاصی وابسته نباشند. می‌توان گفت قابلیت تکرار تصمیمات نیز معادل جنبه‌هایی از پایایی در دیدگاه کلاسیک ارزیابی است. به طور خلاصه قابلیت تکرار تصمیمات در برنامه ارزیابی یعنی تصمیمات نهایی و مهم در مورد عملکرد دانشجو باید بر اساس ارزیابان متعدد، مواجهه‌های متعدد، شرایط مختلف و روش‌های ارزیابی متنوع باشد.

۴. مقبولیت

ارزیابی باید توسط افرادی که در آن حرفه مشغول هستند پذیرفته شده باشد. به این معنی که نگرش و دیدگاه مثبتی نسبت به آن داشته باشند. همچنین معیارهای ارزیابی و نحوه اجرای برنامه ارزیابی مورد قبول همه ذی‌نفعان شامل دانشجویان، استادان و کارکنان باشد و از کیفیت آن مطمئن باشند.

---

1. The Wheel of Competency Assessment
2. Fitness for Purpose
3. Comparability
4. Reproducibility of decisions
5. Acceptability
6. Transparency

۵. شفافیت

شفافیت به این معنی است که آیا برنامه ارزیابی شفاف است و برای همه ذی‌نفعان قابل درک است. فراگیران باید معیارهای نمره‌دهی را بدانند، ارزیابان چه کسانی هستند و هدف ارزیابی چیست. یکی از شاخص‌های شفافیت آزمون این است که آیا فراگیران خود و دیگر فراگیران را با همان دقت ارزیابان آموزش دیده مورد قضاوت قرار می‌دهند.

لایه بیرونی چرخ ارزیابی توانمندی شامل معیارهای عادلانه بودن[1]، واقعی بودن[2]، پیچیدگی شناختی[3]، معنی‌دار بودن[4] و مناسب بودن برای خود ارزیابی[5] است. این معیارها جدیدتر هستند و از تغییر فرهنگ ارزیابی نشأت می‌گیرند. انتظار می‌رود این معیارها در عمل کمتر از معیارهای لایه درونی استفاده شود. معیارهای لایه درونی پیش نیاز معیارهای لایه بیرونی هستند. در زمان طراحی برنامه ارزیابی احتمالاً معیارهای لایه درونی برآورده می‌شوند و سپس معیارهای لایه بیرونی بر اساس معیارهای لایه درونی بنا می‌شوند. به عنوان مثال، یک برنامه ارزیابی نمی‌تواند بدون قابل مقایسه و تکرار پذیر بودن، عادلانه باشد و باید شفاف باشد قبل از آن‌که معنی‌دار باشد.

۶. عادلانه بودن

برنامه ارزیابی نباید به سمت فراگیران گروه‌های خاصی سوگیری داشته باشد، منعکس‌کننده دانش، مهارت و نگرش مربوط به توانمندی‌های مهم باشد و واریانس نامربوط را حذف کند. علل احتمالی سوگیری عبارتند از عدم تناسب با سطح فراگیران یا استفاده از محتوایی که به عنوان مثال شامل جنبه‌های فرهنگی است که همه فراگیران با آن آشنا نیستند.

۷. واقعی بودن

واقعی بودن به معنی میزان همانندی برنامه ارزیابی با زندگی حرفه‌ای آینده فراگیران است. برنامه ارزیابی باید توانمندی‌های مورد نیاز برای کار در محیط طبابت آینده را ارزیابی کند. ۵ حوزه مشخص می‌تواند در واقعی بودن برنامه ارزیابی تاثیرگذار باشد: فعالیت مورد ارزیابی، محیط فیزیکی، محیط اجتماعی، نتایج ارزیابی یا روش مورد استفاده برای ارزیابی و معیارهای ارزیابی. واقعی بودن تا حدودی با روایی محتوا در چارچوب مسیک نزدیک است ولی در عین حال چیزی فراتر از آن است، به طوری که محیط کار و شرایط اجتماعی که نشان دهنده شرایط کاری آینده فراگیر است را نیز در بر می‌گیرد.

۸. پیچیدگی شناختی

پیچیدگی شناختی از این جهت که به فرایندهای به کار رفته در زندگی حرفه‌ای آینده مربوط می‌شود، شبیه واقعی بودن است ولی مستقیماً بر این واقعیت استوار است که تکلیف ارزیابی باید مهارت‌های شناختی سطوح بالا را نیز ارزیابی کند. بسته به مرحله آموزشی، یک فعالیت ارزیابی باید فرایندهای تفکر مورد استفاده توسط متخصصان جهت حل مشکلات پیچیده در حوزه کاری آن‌ها را مورد بررسی قرار دهد. بنابراین، ارزیابی عملکرد لزوماً تضمین کننده ارزیابی مهارت‌های شناختی سطوح بالا نیست و این مهارت‌ها باید به طور کامل ارزیابی شود.

پیچیدگی شناختی معادل روایی اساسی در چارچوب مسیک است. روایی اساسی به این معنی است که فرایندهای فکری که در ارزیابی به کار گرفته می‌شوند تا چه حد منعکس کننده فرایندهای فکری است که توسط افراد در محیط کار به کار گرفته می‌شود. تنها تفاوت در نحوه سنجش این دو است. به این صورت که روایی اساسی بر اساس تحلیل وظیفه

1. Fairness
2. Authenticity
3. Cognitive complexity
4. Meaningfulness
5. Fitness for Self-assessment

و پروتکل‌های بلند فکر کردن ارزیابی می‌شود. پیچیدگی شناختی علاوه بر این دو به اطمینان از فرایند فکری در زمان آزمون نیز مربوط می‌شود. به عنوان مثال، ممکن است از ارزیابی‌شونده بخواهیم در مورد انتخاب‌های مختلفی که دارد، توضیح دهد.

۹. معنی‌دار بودن

برنامه ارزیابی باید برای هر دو گروه استادان و فراگیران دارای ارزش و معنی مشخص باشد. یک روش امکان‌پذیر برای افزایش معنای ارزیابی، درگیر نمودن فراگیران در فرایند ارزیابی است. فراگیران در صورتی یک فعالیت ارزیابی را با معنی می‌دانند که با علایق فردی آن‌ها ارتباط داشته باشد. همچنین از نظر آن‌ها آزمون زمانی معنی‌دار است که آن‌ها آمادگی کافی برای آزمون را کسب نموده باشند تا بیشترین فایده را از آزمون ببرند.

۱۰. مناسب بودن برای خودارزیابی

برنامه ارزیابی باید یادگیری خودتنظیمی را در دانشجویان با به کارگیری روش‌هایی مانند تمرین خودارزیابی و ارائه و دریافت بازخورد ترغیب کند.

چهارگوش اطراف چرخ ارزیابی توانمندی، فضای آموزشی گسترده‌تری که در آن ارزیابی رخ می‌دهد را نشان می‌دهد که شامل دو معیار پیامدهای آموزشی[1] و هزینه و مقرون به صرفه بودن[2] است.

۱۱. پیامدهای آموزشی

معیار فوق به نتایج آموزشی که برنامه آموزشی بر تدریس و یادگیری دارد اشاره می‌کند. مجموعه‌ای از شواهد در مورد اثرات مطلوب یا غیرمطلوب و مثبت یا منفی ارزیابی مورد نیاز است تا مشخص شود فراگیران و معلمان از اهداف آموزشی چه برداشتی دارند و چگونه فعالیت‌های یادگیری خود را با آن منطبق می‌کنند. مطالعات بعدی در مورد نتایج آموزشی باید در پاسخ به این سؤال باشد که آیا همه معیارهای کیفیت برای رسیدن به اثر مثبت بر یادگیری لازم هستند یا برخی از معیارها از تاثیر بیشتری نسبت به بقیه برخوردار هستند.

۱۲. هزینه و مقرون به صرفه بودن

در سیستم کلی آموزشی زمان و هزینه باید به همه بخش‌های آموزش که ارزیابی نیز یکی از بخش‌های آن است اختصاص یابد. اگر یک برنامه ارزیابی به صورت درستی طراحی شود و همه معیارها را در نظر بگیرد ولی به دلیل هزینه بسیار بالا و کارایی پایین نتواند اجرا شود، طراحی آن اتلاف وقت است. هزینه و مقرون به صرفه بودن بسیار مهم هستند زیرا برنامه ارزیابی عموماً پیچیده‌تر از آزمون‌های کلاسیک است و اجرای آن نیز مشکل‌تر است. این معیار به زمان و منابع مورد نیاز برای طراحی و اجرای برنامه ارزیابی توانمندی در مقایسه با منافع آن اشاره دارد. سرمایه‌گذاری بیشتر بر زمان و منابع در برابر اثرات مثبت بیشتر مانند ارتقای یادگیری و تدریس توجیه می‌شود[3].

---

1. Educational consequences
2. Costs and efficiency

۳. در چارچوب اولیه‌ای که بارتمن در مطالعه خود در نظر گرفت یکی از معیارهای کیفیت برنامه ارزیابی صراحت (Directness) بود. صراحت به میزانی که معلمان و ارزیابان می‌توانند تفسیر فوری از نتایج ارزیابی بدون ترجمه آن‌ها از تئوری به عمل داشته باشند اشاره دارد. یک آزمون نظری نمی‌تواند به طور سریع نشان دهد که یک فراگیر در موقعیت شغلی خود موفق است در حالی که یک آزمون عملکردی می‌تواند. می‌توان شواهدی یافت مبنی بر این که روش‌های مستقیم بهتر از روش‌های غیرمستقیم موفقیت فرد در کار را پیش‌بینی می‌کند. البته این به این معنی نیست که روش‌های غیر مستقیم مانند آزمون‌های دانشی نباید در برنامه ارزیابی گنجانده نشود. این معیار در چارچوب نهایی در دیگر معیارها ادغام شد.

شکل ۱-۳۸: چرخ ارزیابی توانمندی (بارتمن و همکاران ۲۰۰۶)

بارتمن و همکاران (۲۰۰۷) به منظور عملیاتی سازی معیارهای فوق شاخص‌هایی را برای هر معیار تعریف کردند (جدول ۱-۳۸). شایان ذکر است پژوهش‌ها در زمینه کیفیت برنامه ارزیابی انگشت شمار است و در اغلب موارد در حد پیشنهاد معیارها و شاخص‌ها است. لازم است مطالعات بیشتری با هدف بررسی سودمندی چارچوب‌ها، معیارها و شاخص‌های پیشنهادی در حوزه آموزش پزشکی انجام شود.

جدول ۱-۳۸: معیارها و شاخص‌های کیفیت برنامه ارزیابی (بارتمن و همکارن ۲۰۰۷)

تناسب با هدف	مقبولیت
۱. پوشش پروفایل توانمندی ۲. ارزیابی تلفیقی دانش/مهارت/نگرش ۳. استفاده از روش‌های ارزیابی مختلف ۴. استفاده از ارزیابی تکوینی و تراکمی هردو ۵. تناسب روش‌های ارزیابی با اهداف آموزشی	۱. پذیرش معیارهای ارزیابی توسط فراگیران ۲. پذیرش پروسیجر ارزیابی توسط فراگیران ۳. پذیرش برنامه ارزیابی توان‌مندی توسط مدرسان ۴. پذیرش برنامه ارزیابی توان‌مندی توسط کارکنان ۵. اعتماد به کیفیت برنامه ارزیابی توان‌مندی
مقایسه پذیر بودن	**شفافیت**
۱. مقایسه‌پذیر بودن تکالیف ارزیابی ۲. مقایسه‌پذیر بودن شرایط کاری ۳. مقایسه‌پذیر بودن معیارهای ارزیابی ۴. مقایسه‌پذیر بودن پروسیجرهای ارزیابی	۱. آگاهی دانشجویان از کابرد ارزیابی (تکوینی یا تراکمی) ۲. آگاهی دانشجویان از معیارها ۳. آگاهی دانشجویان از پروسیجرها ۴. آگاهی و درک استادان از برنامه ارزیابی ۵. آگاهی و درک کارکنان از برنامه ارزیابی ۶. امکان ممیزی بیرونی

ادامه جدول ۱-۳۸ معیارها و شاخص‌های کیفیت برنامه ارزیابی (بارتمن و همکارن ۲۰۰۷)

قابلیت تکرار تصمیمات	واقعی بودن
۱. ارزیابی در چندین زمان ۲. استفاده از چندین ارزیاب ۳. استفاده از ارزیابان با پیش‌زمینه‌های مختلف ۴. تبادل نظر و بحث کافی بین ارزیابان ۵. استفاده از ارزیابان آموزش دیده و توانمند ۶. ارزیابی در چندین موقعیت کاری	۱. تکالیف ارزیابی نمایان‌گر فعالیت/حرفه/شغل فراگیر است. ۲. شرایط انجام ارزیابی نمایان‌گر فعالیت/حرفه/شغل فراگیر است. ۳. محیط اجتماعی که ارزیابی در آن انجام می‌شود نمایان‌گر فعالیت/حرفه/شغل فراگیر است. ۴. معیارهای ارزیابی نمایان‌گر فعالیت/حرفه/شغل فراگیر است.
پیچیدگی شناختی	**رعایت عدالت**
۱. تکالیف ارزیابی مراحل تفکر را به حرکت در می‌آورد. ۲. موقعیت‌های مختلف را تبیین می‌کند. ۳. معیارهای ارزیابی مراحل تفکر را به حرکت در می‌آورد. ۴. تکالیف ارزیابی نیازمند سطوح مختلف تفکر است.	۱. پروسیجرهایی برای اصلاح اشتباهات تدوین شده است. ۲. وزن عناصر ارزیابی بر اساس اهمیت آن است. ۳. عدم تعصب ارزیابان ۴. استفاده از انواع مختلفی از وظایف ارزیابی ۵. دانشجویان برنامه ارزیابی توان‌مندی را عادلانه می‌دانند.
تناسب برای خود ارزیابی	**معنی‌دار بودن**
۱. امکان خود ارزیابی و ارزیابی توسط همگنان ۲. ارائه و دریافت بازخورد ۳. بازاندیشی بر ارتقای فردی ۴. فرمول‌بندی اهداف یادگیری فردی	۱. مفید بودن بازخورد ارزیابی تکوینی ۲. مفید بودن بازخورد ارزیابی تراکمی ۳. ارزیابی فرصتی برای یادگیری است ۴. از نظر دانشجویان معیارها بامفهوم است ۵. از نظر مدرسان و کارکنان معیارها بامفهوم است
پیامدهای آموزشی	**هزینه و مقرون به صرفه بودن**
۱. ترغیب به سمت فرایندهای یادگیری مطلوب ۲. تاثیر مثبت بر دانشجویان ۳. تاثیر مثبت بر مدرسان ۴. بهبود نتایج منفی (در صورت وجود) ۵. تغییر و بازنگری برنامه درسی	۱. زمان و هزینه تخمین زده شده ۲. ترکیب آگاهانه از ابزارهای ارزیابی ۳. ارزشیابی سالانه کارآیی ۴. اثرات مثبت مزید بر سرمایه صرف شده

به صورت خلاصه باید گفت برنامه ارزیابی مانند هر برنامه آموزشی دیگری نیز نیاز به ارزشیابی دارد. لازم است برنامه ارزیابی به طور سیستماتیک مورد پایش قرار گیرد و فعالیت‌ها و فرایند آن در صورتی که به طور مناسب اجرا نمی شود و دچار بیهوده‌گرایی شده است اصلاح شود. علاوه بر این، دستیابی به پیامدهای مطلوب، پیامدهای ناخواسته و عوارض جانبی مورد بررسی قرار گیرد (وندرولوتن و همکاران ۲۰۱۴).

ارزشیابی برنامه ارزیابی می‌تواند شامل برنامه یک فاز (مانند دوره علوم پایه)، یک نیم‌سال یا سال تحصیلی، یک موضوع خاص (مانند بیولوژی) یا حتی برای یک برنامه آموزشی کامل (مانند برنامه پرستاری) باشد. ارزشیابی برنامه ارزیابی می‌تواند درونی یا بیرونی باشد. در صورت ارزشیابی درونی افرادی از درون همان برنامه ارزشیابی را انجام می‌دهند. ذی‌نفعان مختلف برنامه از جمله منتورها می‌توانند منبع بسیار مهمی برای جمع‌آوری اطلاعات باشند. روش‌های گردآوری اطلاعات نیز مختلف بوده و شامل تحلیل نمرات آزمون، پرسشنامه‌های نظر خواهی از دانشجویان، مصاحبه با ذی‌نفعان، مستندات برنامه و ... است. استفاده از مدل‌های شناخته شده ارزشیابی برنامه می‌تواند در تمام مراحل ارزشیابی یک برنامه ارزیابی حتی در زمانی که برنامه در حال تدوین است کمک کننده باشد (گندم‌کار و همکاران ۲۰۱۵)

بارتمن و همکاران (۲۰۰۷)

بارتمن و همکارن (۲۰۰۶) ۱۲ معیار را برای کیفیت برنامه ارزیابی تدوین کردند. سپس در مطالعه مذکور برای هر معیار ۴ تا ۶ شاخص تعریف شد. پژوهشگران یک خود ارزیابی[1] از برنامه‌های ارزیابی ۸ دبیرستان فنی حرفه‌ای در هلند بر اساس شاخص‌های تهیه شده انجام دادند. ابزار ارزشیابی مبتنی بر وب بود و برای شرکت کنندگان که شامل مدیر گروه فن‌آوری آزمایشگاهی، یک نفر از بورد امتحانات و یکی از معلمان گروه از هر دبیرستان بودند ارسال شد. ابتدا افراد توصیفی از برنامه ارزیابی توانمندی را تکمیل کردند که شامل سال، سطح آموزش و روش‌های ارزیابی مورد استفاده بود. سپس برنامه ارزیابی را بر اساس شاخص‌ها ارزشیابی کردند. علاوه بر شاخص‌های مذکور، دو امکان هم برای شاخص‌های بیشتر در نظر گرفته شده بود. از شرکت کنندگان خواسته شد یک مثال یا مستنداتی که نمره‌دهی ایشان به هر شاخص را حمایت کند یادداشت کنند. نتایج مرحله اول به عنوان مبنایی برای مصاحبه گروهی مورد استفاده قرار گرفت تا نقاط قوت و ضعف برنامه ارزیابی مشخص شود. تحلیل داده‌ها نشان داد آلفای کرونباخ در خصوص برخی از معیارها بالاتر از ۰/۷۰ و در مورد برخی دیگر (مقبولیت، هزینه و کارآیی، رعایت عدالت، قابلیت تکرار تصمیمات و شفافیت) کمتر از ۰/۰۷ بود. نویسندگان پیشنهاد دادند که با توجه به حجم نمونه پایین مطالعه، مطالعات بیشتری مورد نیاز است. بیشترین میزان نمره بالا به برنامه ارزیابی دبیرستان‌ها مربوط به معیار قابلیت مقایسه بود و کمترین مربوط به معیار معنی دار بودن، هزینه و کارآیی و پیچیدگی شناختی بود.

بارتمن و همکاران (۲۰۱۱)

در این مطالعه به منظور اطمینان ازاعتبار معیارهای برنامه ارزیابی، نتایج خودارزیابی دو دبیرستان کم تجربه و باتجربه در برنامه ارزیابی با هم مقایسه شدند. نتایج نشان داد که دو مدرسه از رویکردهای متفاوتی به منظور اطمینان از کیفیت ارزیابی استفاده کردند. مدرسه باتجربه به نظر می‌رسید که از نقاط ضعف و قوت خود آگاه است و ذی‌نفعان مختلف را در برنامه ارزیابی مشارکت داده است. همچنین از دورنمای موسسه آگاهی بیشتری داشته و ارزیابی را در راستای آن تدوین نموده است.

1. Self-study

منابع

1. Baartman LK, Bastiaens TJ, Kirschner PA, Van der Vleuten CPM. The wheel of competency assessment: Presenting quality criteria for competency assessment programs. Stud Educ Eval. 2006;32(2):153–70.
2. Baartman LK, Bastiaens TJ, Kirschner PA, Van der Vleuten CPM. Evaluating assessment quality in competence-based education: A qualitative comparison of two frameworks. Educ Res Rev. 2007;2(2):114–29.
3. Baartman LK, Prins FJ, Kirschner PA, Van der Vleuten CPM. Determining the quality of competence assessment programs: A self-evaluation procedure. Stud Educ Eval. 2007;33(3):258–81.
4. Baartman LK, Prins FJ, Kirschner PA, Van der Vleuten CPM. Self-evaluation of assessment programs: A cross-case analysis. Evalu Program Plann. 2011;34(3):206-16.
5. Biggs J. Enhancing teaching through constructive alignment. High Educ. 1996;32(3):347-64.
6. Cantillon P, Wood D. ABC of Learning and Teaching in Medicine. 2nd ed. West Sussex: John Wiely & Sons; 2010.
7. Gandomkar R, Jalili M, Mirzazadeh A. Evaluating assessment programmes using programme evaluation models. Med Teach. 2015;37(8):792-3.
8. Holmboe ES, Hawkins RE. Practical guide to the evaluation of clinical competence. Philadelphia: Mosby/Elsevier; 2008.
9. Linn RL, Baker J, Dunbar SB. Complex, performance-based assessment: Expectations and validation criteria. Educ Res. 1991;20(8):15-21.
10. Messick S. The interplay of evidence and consequences in the validation of performance assessments. Educ Res. 1994;23(2):13–23.
11. Schuwirth LWT, van der Vleuten CPM. Programmatic assessment: from assessment of learning to assessment for learning. Med Teach. 2011;33(6):478–85.
12. Swanwick T. Uderstanding Medical Education: Evidence, Theory and Practice. West Sussex: John Wiely & Sons; 2010.
13. van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: from methods to programmes. Med Educ. 2005;39(3):309–17.
14. van der Vleuten CP, Schuwirth LWT, Driessen EW, Govaerts MJB, Heeneman S.. 12 Tips for programmatic assessment. Med Teach. 2015;37(7):641-6
15. Wass V, McGibbon D, van der Vleuten CPM. Composite undergraduate clinical examinations: How should the components be combined to maximize reliability? Medical Education. 2001:35(4);326–30.
16. Webb NL. Issues related to judging the alignment of curriculum standards and assessments. Applied Measurement in Education. 2007;20(1):7-25.

نمایه

و

A

C

D

G

K

O