Guide to Item Analysis

Introduction

Item Analysis (a.k.a. Test Question Analysis) is a useful means of discovering how well individual test items assess what students have learned. For instance, it helps us to answer the following questions.

- Is a particular question as difficult, complex, or rigorous as you intend it to be?
- Does the item do a good job of separating students who know the content from those who may merely either guess the right answer or apply test-taking strategies to eliminate the wrong answers?
- Which items should be eliminated or revised before use in subsequent administrations of the test?

With this process, you can improve test score validity and reliability by analyzing item performance over time and making necessary adjustments. Test items can be systematically analyzed regardless of whether they are administered as a Canvas assignment or if they are submitted as "bubble sheets" to Scanning Services.

With this guide, you'll be able to

- Define and explain the indices related to item analysis.
- Locate each index of interest within Scanning Services' Exam Analysis reports.
- Identify target values for each index, depending upon your testing intentions.
- Make informed decisions about whether to retain, revise, or remove test items.

Anatomy of a Test Item

In this guide, we refer to the following terms to describe the items (or questions) that make up multiple-choice tests.

1. **Stem** refers to the portion of the item that presents a problem for the respondents (students) to solve
2. **Options** refers to the various ways the problem might be solved, from which respondents select the best answer.
    a. **Distractor** is an incorrect option.
    b. **Key** is a correct option.

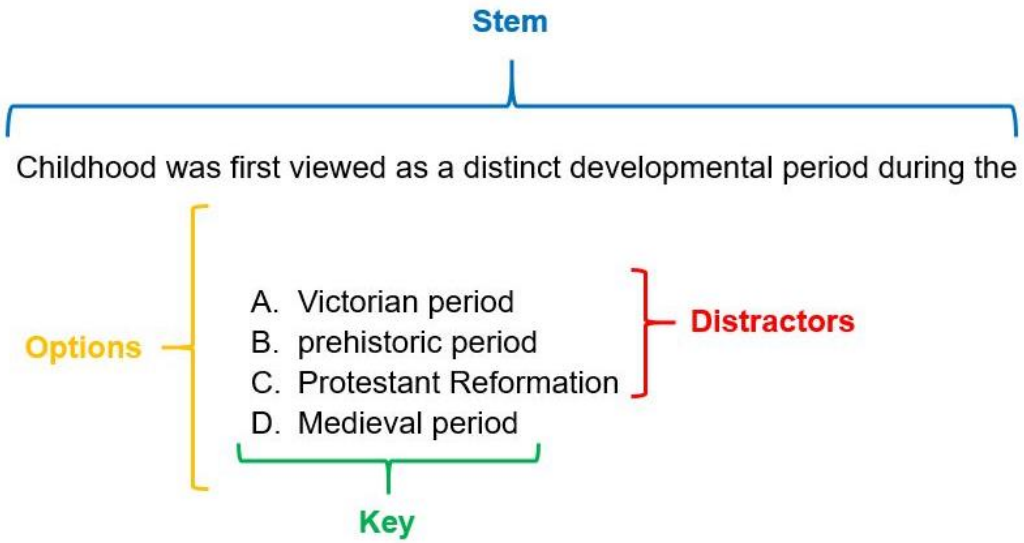


Figure 1: Anatomy of a test item

Item Analysis in Canvas

By default, the quiz summary function in Canvas shows average score, high score, low score, standard deviation (how far the values are spread across the entire score range), and average time of quiz completion. This means that, after the quiz has been administered, you automatically have access to those results, and you can sort those results by Student Analysis or Item Analysis. The Canvas Doc Team offers a number of guides on using these functions in that learning management system. Click on Search the Canvas Guides under the Help menu and enter "Item Analysis" for the most current information.

Item Analysis in Scanning Services

Scanning Services offers an Exam Analysis Report (see example) through its Instructor Tools web site. Learn how to generate and download the report at Scanning Services Instructor Tools Help.

Four Steps to Item Analysis

Item analysis typically focuses on four major pieces of information: test score reliability, item difficulty, item discrimination, and distractor information. No single piece should be examined independent of the others. In fact, understanding how to put them all together to help you make a decision about the item's future viability is critical.

Reliability

Test Score Reliability is an index of the likelihood that scores would remain consistent over time if the same test was administered repeatedly to the same learners. Scanning Services' Exam Analysis Report uses the Cronbach's Alpha measure of internal consistency, which provides reliability information about items scored dichotomously (i.e., correct/incorrect), such as multiple choice items. A test showing a Chronbach's Alpha score of .80 and higher has less measurement error and is thus said to have very good reliability. A value below .50 is considered to have low reliability.

Item Reliability is an indication of the extent to which your test measures learning about a single topic, such as "knowledge of the battle of Gettysburg" or "skill in solving accounting problems." Measures of internal consistency indicate how well the questions on the test consistently and collectively address a common topic or construct.

In Scanning Services' **Exam Analysis Report**, next to each item number is the percentage of students who answered the item correctly.

To the right of that column, you'll see a breakdown of the percentage of students who selected each of the various **options** provided to them, including the **key** (in dark grey) and the **distractors** (A, B, C, D, etc.). Under each **option**, the **Total (TTL)** indicates the total number of students who selected that option. The **Reliability coefficient (R)** value shows the mean score (%) and Standard Deviation of scores for a particular distractor.

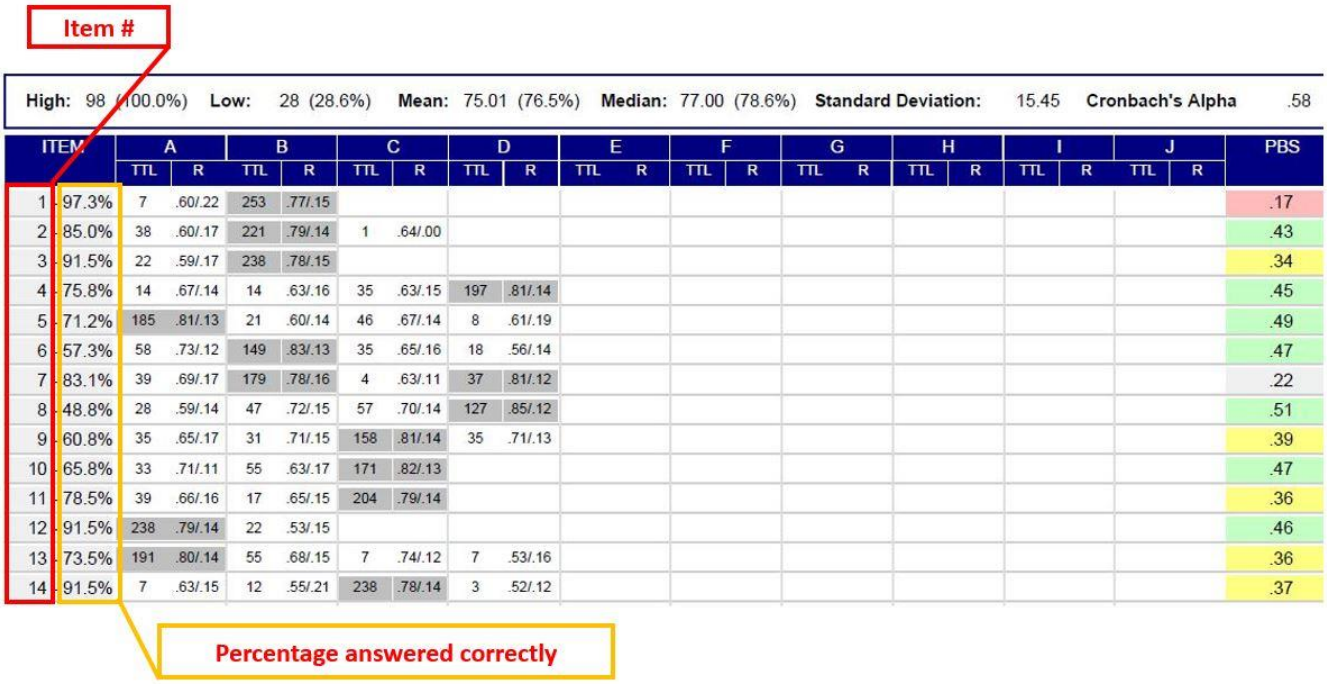


Figure 2: Item number and percentage answered correctly on Exam Analysis Report

How would you use this information?

Score Reliability is dependent upon a number of factors, including some that you can control and some that you can't.

Factor	Why it's important
Length of the test	Reliability improves as more items are included.
Proportion of students responding correctly and incorrectly to each item	Helps determine item reliability.
Item difficulty	Very easy and very difficult items do not discriminate well and will lower the reliability estimate.
Homogeneity of item content	Reliability on a particular topic improves as more items on that topic are included. This can present a challenge when a test seeks to assess a lot of topics. In that case, ask questions that are varied enough to survey the topics, but similar enough to collectively represent a given topic.
Number of test takers	Reliability improves as more students are tested using the same pool of items.
Factors that influence any individual test taker on any given day	Preparedness, distraction, physical wellness, test anxiety, etc. can affect students' ability to choose the correct option.

What should you aim for?

Reliability coefficients range from 0.00 to 1.00. Ideally, score reliability should be above 0.80. Coefficients in the range 0.80-0.90 are considered to be very good for course and licensure assessments.

Difficulty

Item Difficulty represents the percentage of students who answered a test item correctly. This means that low item difficulty values (e.g., 28, 56) indicate difficult items, since only a small percentage of students got the item correct. Conversely, high item difficulty values (e.g., 84, 96) indicate easier items, as a greater percentage of students got the item correct.

As indicated earlier, in Scanning Services' Exam Analysis Report, there are two numbers in the Item column: item number and the percentage of students who answered the item correctly. A higher percentage indicates an easier item; a lower percentage indicates a more difficult item. It helps to gauge this difficulty index against what you expect and how difficult you'd like the item to be. You should find a higher percentage of students correctly answering items you think should be easy and a lower percentage correctly answering items you think should be difficult.

Item difficulty is also important as you try to determine how well an item "worked" to separate students who know the content from those who do not (see **Item Discrimination** below). Certain items do not discriminate well. Very easy questions and very difficult questions, for example, are poor discriminators. That is, when most students get the answer correct, or when most answer incorrectly, it is difficult to ascertain who really knows the content, versus those who are guessing.

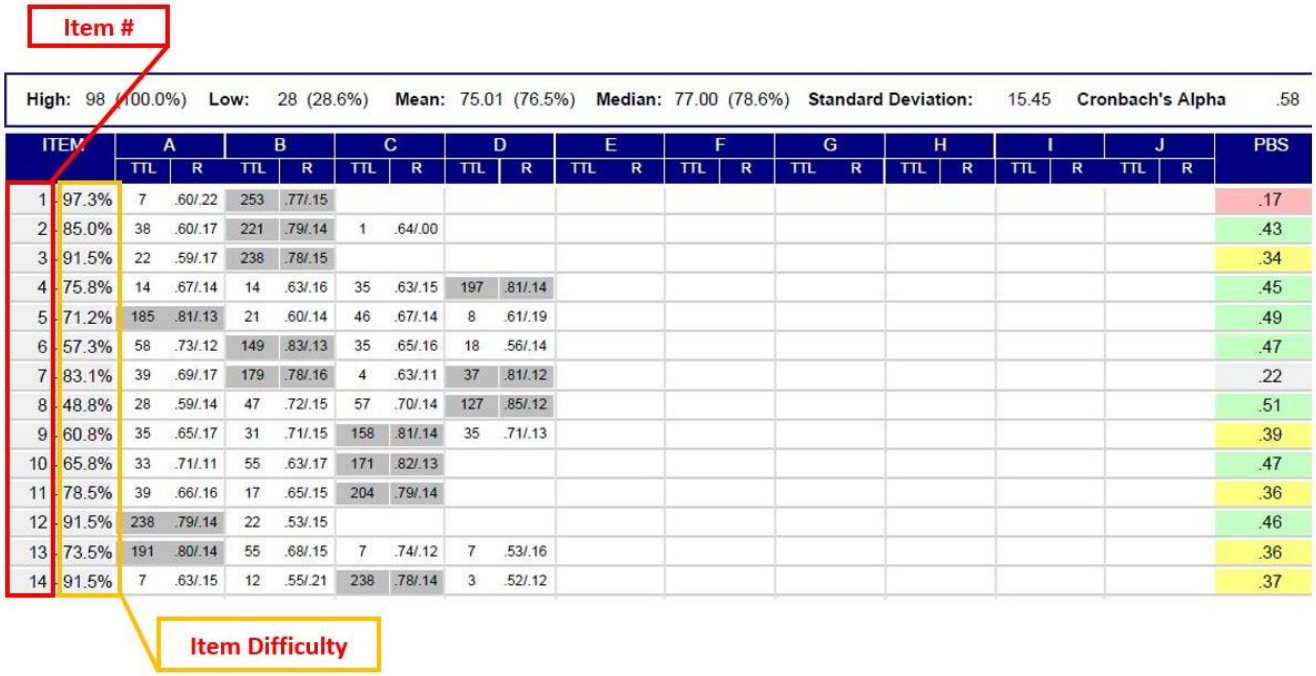


Figure 3: Item number and item difficulty on Exam Analysis Report

How should you use this information?

As you examine the difficulty of the items on your test, consider the following.

1. Which items did students find to be easy; which did they find to be difficult? Do those items match the items you *thought* would be easy/difficult for students? Sometimes, for example, an instructor may put an item on a test believing it to be one of the easier on the exam when, in fact, students find it to be challenging.
2. Very easy items and very difficult items don't do a good job of discriminating between students who know the content and those who do not. (The section on **Item Discrimination** discusses this further.) However, you may have very good reason for putting either type of question on your exam. For example, some instructors deliberately start their exam with an easy question or two to settle down anxious test takers or to help students feel some early success with the exam.

What should you aim for?

Popular consensus suggests that the best approach is to aim for a mix of difficulties. That is, a few very difficult, some difficult, some moderately difficult, and a few easy. However, the level of difficulty should be consistent with the degree of difficulty of the concepts being assessed. The Testing Center provides the following guidelines.

% Correct	Item Difficulty Designation
0 – 20	Very difficult
21 – 60	Difficult
61 – 90	Moderately difficult
91 – 100	Easy

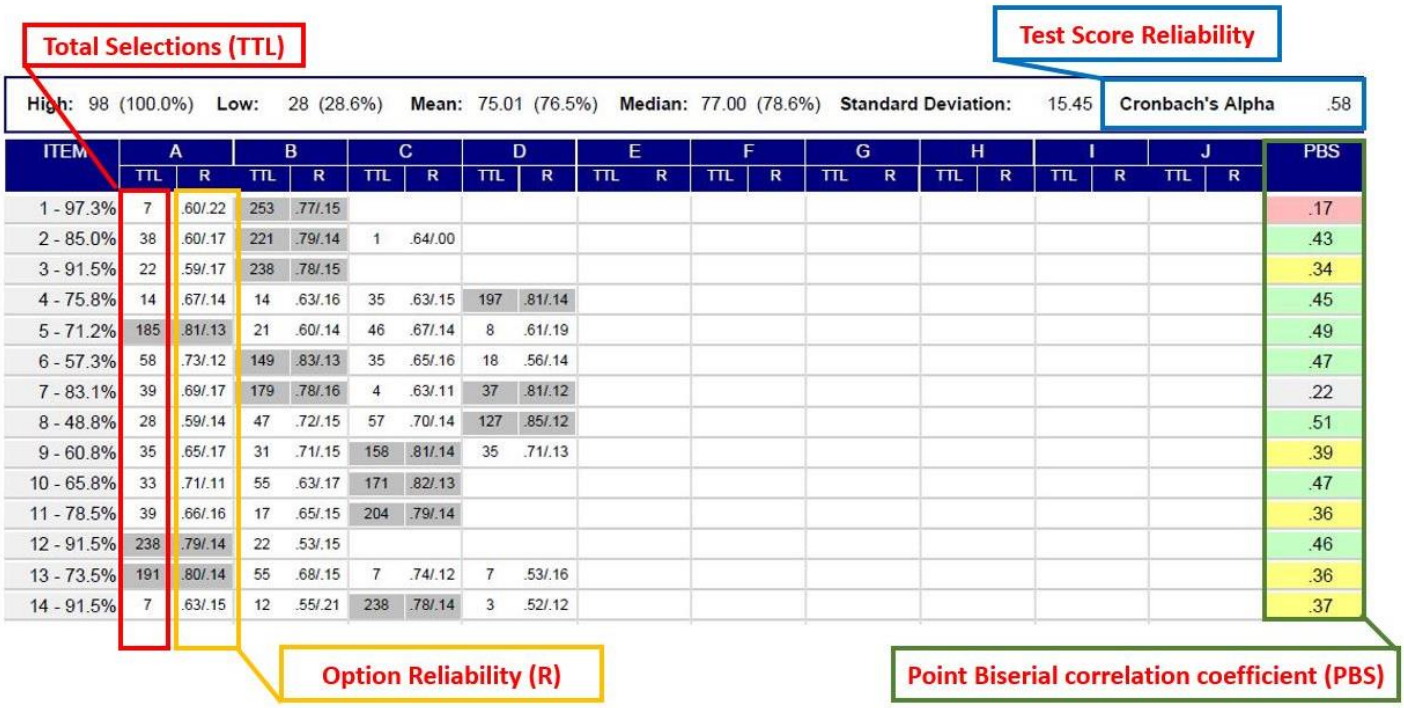
Discrimination

Item Discrimination is the degree to which students with high overall exam scores also got a particular item correct. It is often referred to as Item Effect, since it is an index of an item's effectiveness at discriminating those who know the content from those who do not.

The **Point Biserial correlation coefficient (PBS)** provides this discrimination index. Its possible range is -1.00 to 1.00. A strong and positive correlation suggests that students who get a given question correct also have a relatively high score on the overall exam. Theoretically, this makes sense. Students who know the content and who perform well on the test overall should be the ones who know the content. There's a problem, however, if students are getting correct answers on a test and they don't actually know the content.

One would expect that the students who did well on the exam selected the correct response, thus generating a higher mean score and a higher PBS which shows the correlation between a high overall exam score to a given correct response to an item. Conversely, cases where an incorrect response distracted students who did well on the exam, exhibited by a high R value, should result in a lower PBS score.

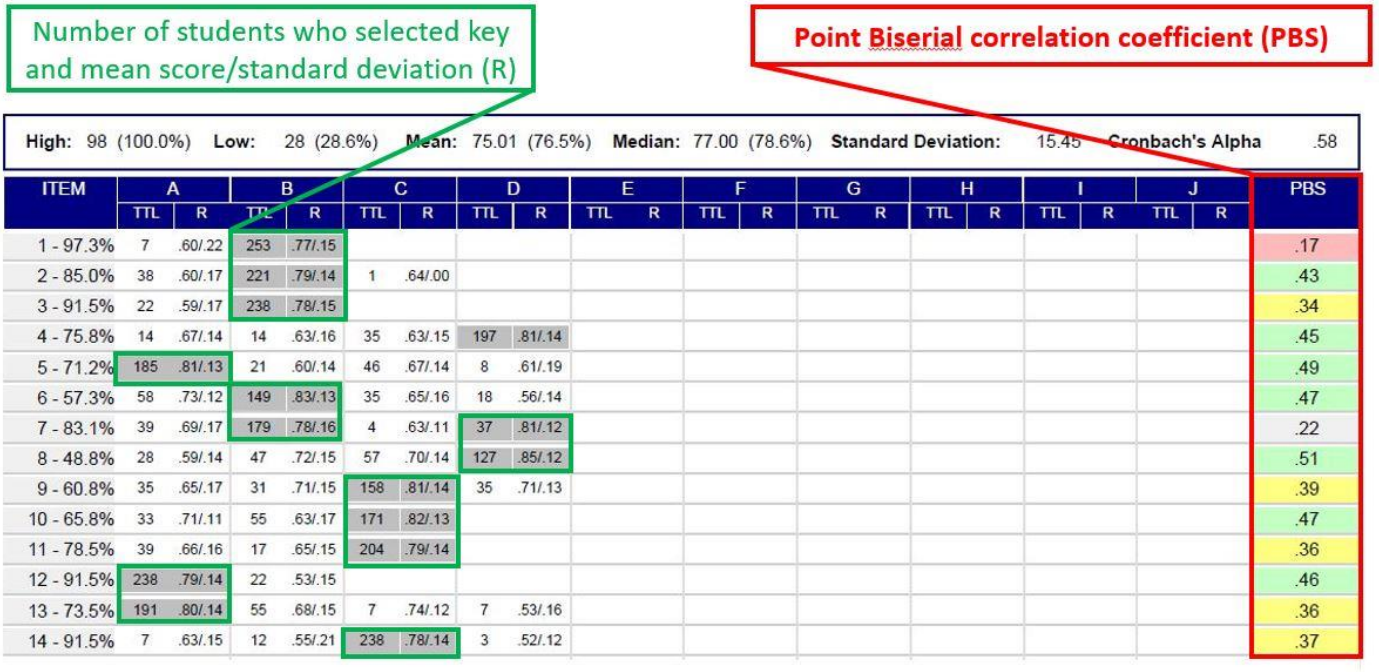
PBS score ranges from -1.0 to 1.0, with a minimum desired score greater than 0.15. If a single test is weighted heavily as part of students' grades, reliability must be high. Low score reliability is an indication that, if students took the same exam again, they might get a different score. Optimally, we would expect to see consistent scores on repeated administrations of the same test.



ITEM	A		B		C		D		E		F		G		H		I		J		PBS
	TTL	R	TTL	R	TTL	R	TTL	R	TTL	R	TTL	R	TTL	R	TTL	R	TTL	R	TTL	R	
1 - 97.3%	7	.60/.22	253	.77/.15																	.17
2 - 85.0%	38	.60/.17	221	.79/.14	1	.64/.00															.43
3 - 91.5%	22	.59/.17	238	.78/.15																	.34
4 - 75.8%	14	.67/.14	14	.63/.16	35	.63/.15	197	.81/.14													.45
5 - 71.2%	185	.81/.13	21	.60/.14	46	.67/.14	8	.61/.19													.49
6 - 57.3%	58	.73/.12	149	.83/.13	35	.65/.16	18	.56/.14													.47
7 - 83.1%	39	.69/.17	179	.78/.16	4	.63/.11	37	.81/.12													.22
8 - 48.8%	28	.59/.14	47	.72/.15	57	.70/.14	127	.85/.12													.51
9 - 60.8%	35	.65/.17	31	.71/.15	158	.81/.14	35	.71/.13													.39
10 - 65.8%	33	.71/.11	55	.63/.17	171	.82/.13															.47
11 - 78.5%	39	.66/.16	17	.65/.15	204	.79/.14															.36
12 - 91.5%	238	.79/.14	22	.53/.15																	.46
13 - 73.5%	191	.80/.14	55	.68/.15	7	.74/.12	7	.53/.16													.36
14 - 91.5%	7	.63/.15	12	.55/.21	238	.78/.14	3	.52/.12													.37

Figure 4: Total selections (TTL), option reliability (R), and point biserial correlation coefficient (PBS) on Exam Analysis Report

In Scanning Services' Exam Analysis Report, you'll find the PBS final column, color-coded so you can easily distinguish the items that may require revision. Likewise, the key for each item is color-coded grey.



ITEM	A		B		C		D		E		F		G		H		I		J		PBS
	TTL	R	TTL	R	TTL	R	TTL	R	TTL	R	TTL	R	TTL	R	TTL	R	TTL	R	TTL	R	
1 - 97.3%	7	.60/.22	253	.77/.15																	.17
2 - 85.0%	38	.60/.17	221	.79/.14	1	.64/.00															.43
3 - 91.5%	22	.59/.17	238	.78/.15																	.34
4 - 75.8%	14	.67/.14	14	.63/.16	35	.63/.15	197	.81/.14													.45
5 - 71.2%	185	.81/.13	21	.60/.14	46	.67/.14	8	.61/.19													.49
6 - 57.3%	58	.73/.12	149	.83/.13	35	.65/.16	18	.56/.14													.47
7 - 83.1%	39	.69/.17	179	.78/.16	4	.63/.11	37	.81/.12													.22
8 - 48.8%	28	.59/.14	47	.72/.15	57	.70/.14	127	.85/.12													.51
9 - 60.8%	35	.65/.17	31	.71/.15	158	.81/.14	35	.71/.13													.39
10 - 65.8%	33	.71/.11	55	.63/.17	171	.82/.13															.47
11 - 78.5%	39	.66/.16	17	.65/.15	204	.79/.14															.36
12 - 91.5%	238	.79/.14	22	.53/.15																	.46
13 - 73.5%	191	.80/.14	55	.68/.15	7	.74/.12	7	.53/.16													.36
14 - 91.5%	7	.63/.15	12	.55/.21	238	.78/.14	3	.52/.12													.37

Figure 3: Number of students who selected key and mean score/standard deviation, and point biserial correlation coefficient (PBS)

Figure 4: Color legend

Color Legend

TTL Dark Grey = Key (Correct Option), with number of students who selected it (TTL), and mean score/standard deviation (R)

TTL White = Distractor (Incorrect Option), with number of students who selected it (TTL), and mean score/standard deviation (R)

PBS Red = Review, revision likely necessary

PBS Light Grey = Marginal, revision may be necessary

PBS Yellow = Better, some revision may be necessary

PBS Green = Best, no revision necessary

How should you use this information?

As you examine item discrimination, there are a number of things you should consider.

1. Very easy or very difficult items are not good discriminators. If an item is so easy (e.g., difficulty = 98) that nearly everyone gets it correct or so difficult (e.g., difficulty = 12) that nearly everyone gets it wrong, then it becomes very difficult to discriminate those who actually know the content from those who do not.
2. That does not mean that all very easy and very difficult items should be eliminated. In fact, they are viable as long you are aware that they will not discriminate well and if putting them on the test matches your intention to either really challenge students or to make certain that everyone knows a certain bit of content.
3. Nevertheless, a poorly written item will have little ability to discriminate.

What should you aim for?

It is typically recommended that item discrimination be at least 0.15 It's best to aim even higher. Items with a negative discrimination are theoretically indicating that either the students who performed poorly on the test overall got the question correct or that students with high overall test performance did *not* get the item correct. Thus, the index could signal a number of problems.

1. There is a mistake on the scoring key.
2. Poorly prepared students are guessing correctly.
3. Well prepared students are somehow justifying or misled by the wrong answer.

In those cases, action must be taken. Items with negative item difficulty must be addressed. Also, reliability improves when students are provided with a total of 4 options that include only 1 key and 3 distractors. Be certain that there is only one possible answer, that the question and options are written clearly, and that your answer key is correct.

Distractors

Distractors are the multiple choice response options that are *not* the correct answer. They are plausible but incorrect options that are often developed based upon students' common misconceptions or miscalculations to see if they've moved beyond them. As you examine distractors, there are a number of things you should consider.

1. Are there at least some respondents for each distractor? If you have 4 possible options for each item but students are selecting from between only one or two of them, it is an indication that the other distractors are ineffective. Even low-knowledge students can reduce the "real" options to one or two, so the odds are now good that they will choose correctly. In short, this situation increases the possibility of students without knowledge guessing correctly.
2. It is not necessary to revisit every single "0" in the response table. Instead, be mindful, and responsive, where it looks as if distractors are ineffective. Typically, this is where there are two or more distractors selected by no one.
3. Are the distractors overly complex, vaguely worded, or contain obviously wrong, "jokey" or "punny" content? Distractors should be not be mini-tests in themselves, nor should they be a waste of effort. Ineffective distractors can be a strong demotivating factor for students that last beyond their initial appearance, since they often lead to questions of fairness.
4. Theoretically, distractors should not be selected by students who have a good understanding of the material. So it's important to determine whether your best students (on the exam) are, for some reason, being drawn to an incorrect answer. Either the distractor is ambiguous or unclear, or the topic deserves more attention during instruction.

What should you aim for?

Distractors should be plausible options. Test writers often use students' misconceptions, mistakes on homework, or missed quiz questions as fodder for crafting distractors. When this is the approach to distractor writing, information about student understanding can be gleaned even from their selection of *wrong* answers.

For best practices on developing effective multiple-choice items, schedule a SITE consultation or see the following resources.

14 Rules for Writing Multiple Choice Questions

Developing Test Items for Course Examinations, Idea Paper 70

Constructing Written Test Questions For the Basic and Clinical Sciences

---

Conclusion

Item analysis is an empowering process. Knowledge of score reliability, item difficulty, item discrimination, and crafting effective distractors can help an instructor make decisions about whether to retain items for future administrations, revise them, or eliminate them from the test item pool. Item analysis can also help an instructor to determine whether a particular portion of course content should be revisited. In any case, all indices should be considered together before making decisions or revisions. One important thing to always keep in mind is that decisions about item revision should be based on the extent to which item performance matches your intent for the item and your intent for the overall exam.

Resources

Suskie, L. (2017). "Making Multiple Choice Tests More Effective." Schreyer Institute for Teaching Excellence. The Pennsylvania State University.

Understanding Item Analyses. (2018). Office of Educational Assessment. University of Washington. Retrieved from http://www.washington.edu/assessment/scanning-scoring/scoring/reports/item-analysis/.

Mary Ann Tobin, PhD, Assistant Research Professor and Instructional Consultant Schreyer Institute for Teaching Excellence
The Pennsylvania State University
site@psu.edu • https://www.schreyerinstitute.psu.edu • (814) 863-2599