Kennwerte psychologischer Testverfahren

Publikationsgestützte Dissertation zur Erlangung des akademischen Grades

eines Doktors der Philosophie im Fachbereich Human- und

Gesundheitswissenschaften der Universität Bremen

vorgelegt von

Jörg Michael Müller

an der Universität Bremen

Bremen, 2001

1. Gutachter: Prof. Dr. Hans-Jörg Henning, Universität Bremen

Danksagung

An dieser Stelle möchte ich mich bei all denen bedanken, die mich im Laufe der Zeit geduldig bei der Entwicklung der Arbeit begleitet und unterstützt haben. Folgenden Personen möchte ich in umgekehrter alphabetischer Reihenfolge danken:

Dr. Hans-Jörg Walter
Prof. Dr. Rolf Steyer
Herbert Scheithauer
Prof. Dr. Jürgen Rost
Norbert Richter
Prof. Dr. Franz Petermann
Gerd Ostermann
Dr. Andreas Müller
Prof. Dr. Ralf Mittenecker
Susanne Meyer
Ulrike Lange
Miriam Körner
Prof. Dr. Hans Jörg Henning
Sandra Helfert
Gunter Groen
Prof. Dr. Georg Gittler
Jochen Geist
Dr. Klaus Freidel
Patricia Eitemüller
Frank Eckhardt
Dr. Patric Becker
Katharina Becker
Dr. Andre Beauducel
Prof. Dr. Frank Baumgärtel
Prof. Dr. Manfred Amelang

Bremen, den 30.3.2001 JMM

INHALTSVERZEICHNIS

2. ARTIKEL:

DAS PERSONENUNTERSCHEIDUNGSVERMÖGEN EINES DIAGNOSTISCHEN TESTS UNTER BERÜCKSICHTIGUNG DER MESSWERTVERTEILUNG 47-77

3. ARTIKEL:

VARIATIONSBREITE PSYCHOLOGISCHER EIGENSCHAFTEN: DEFINITION UND MESSUNG ÜBER DIE RASCHSKALIERUNG 78-100

EINLEITUNG

EINLEITUNG

‚Kennwerte psychologischer Testverfahren' stellt das verbindende Thema dreier Artikel dar, welche jeweils eine bestimmte wissenschaftliche Fragestellung vertieft behandeln. Die Einleitung soll in die Thematik einführen und die Artikel in einen Zusammenhang setzen. In der Einleitung wird insbesondere darauf eingegangen, was Testkennwerte und was psychologische Testverfahren sind, wozu Testkennwerte genutzt werden und welche Probleme in Zusammenhang mit Testkennwerten auftreten. Die letzte Frage leitet auf eine kurze Inhaltsangabe des ersten Artikels über, wobei sich Kurzzusammenfassungen der beiden verbleibenden Artikel anschließen.

Was sind Testkennwerte?

Testkennwerte sind - allgemein formuliert - Maßzahlen, die einen bestimmten Testaspekt abbilden. Hierzu zählen die klassischen Gütekriterien eines Tests, wie der Reliabilitätskoeffizient, der Validitätskoeffizient und die Objektivität (insbesondere der Aspekt der Beurteilerübereinstimmung). Neben diesen eher statistischen Größen gibt es noch rein deskriptive Angaben zur benötigten Zeit einer Testanwendung oder die Anzahl enthaltener Fragen und Aufgaben. Die Menge an denkbaren Testkennwerten ist prinzipiell unbegrenzt und bedeutsame Neuerungen haben sich im Zuge der Item-Response-Theorien (Rost, 1996) durch ‚Modell-Fit-Indizes' ergeben. Diese Maßzahl steht für die ‚Passung' des Messmodells und der empirischen Datengrundlage. Aus dem Ansatz der Item-Response-Theorie kündigen sich weitere Testkennwerte an (Rost, 2000). Testkennwerte gibt es, seit die psychometrischen Grundlagen der Messfehlertheorie Anfang des 20. Jahrhunderts entwickelt wurden. Testkennwerte

sind als wissenschaftliches Arbeitsgebiet nur einem langsamen Veränderungsprozess unterworfen, vergleicht man es mit inhaltlich definierten Feldern, wie beispielsweise der klinischen Psychologie. Ein Grund hierfür könnte in ihrer grundlegenden Beschreibungsfunktion von psychologischen Testverfahren liegen. Mit der Entwicklung neuer Messmodelle (Rost, 1999) ergibt sich jedoch zunehmend mehr die Notwendigkeit, die neuen Möglichkeiten zu nutzen. Hinzu kommt, dass ein Teil der Testaspekte, die vom Testkuratorium (1986) für die Beschreibung als relevant erachtet werden, noch nicht durch Kennwerte repräsentiert werden. Es ist somit zu erwarten, dass sich zukünftig substantielle Veränderungen im Gebiet der Testkennwerte ergeben werden.

Was sind psychologische Testverfahren?

Psychologische Testverfahren sind ein Resultat der Forschungsbemühungen einer psychologischen Diagnostik. Die psychologische Diagnostik umfasst alle Methoden und deren Anwendung, welche zur Messung und Beschreibung psychologischer Unterschiede verwendet werden (Dorsch, 1991). In der Öffentlichkeit wahrgenommene Beispiele sind Verfahren zur Intelligenzmessung und Fragebögen zur Erfassung weiterer Persönlichkeitseigenschaften. Psychologische Testverfahren gehören mit zu den erfolgreichsten Produktentwicklungen der Psychologie und werden zum Beispiel für die Auswahl von Bewerbern oder zur Diagnostik psychischer Auffälligkeiten verwendet. Testverfahren werden darüber hinaus für die Erstellung von Gutachten herangezogen, aber auch außerhalb der psychologischen Diagnostik als ökonomische Verfahren der Datenerhebung innerhalb empirischer

Studien. Psychologische Testverfahren (und Testkennwerte) unterliegen damit *praktischen* als auch *wissenschaftlichen* Ansprüchen.

Wozu werden Testkennwerte benötigt?

Testkennwerte sind für Testanwender nützliche Informationen, denn sie beschreiben in prägnanter Weise wesentliche Eigenschaften von Testverfahren. Die Beurteilung der ‚Qualität' eines psychologischen Testverfahrens ist immer abhängig vom Anwendungskontext und der Fragestellung. Aus diesem Grund können immer verschiedene Eigenschaften eines Tests interessant sein. Testkennwerte werden für die Beurteilung der Eignung eines Testverfahrens verwendet und unterstützen damit eine möglichst objektive *Testauswahl*. Darüber hinaus bieten Testkennwerte während der *Testkonstruktion* eine wichtige Entscheidungshilfe bei der Itemsauswahl. Items, die sich negativ auf die Kennwerte auswirken, können hierdurch identifiziert und ausgeschlossen werden. Kennwerte werden schließen auch im Anschluss an eine Testung herangezogen, um die *statistische Bedeutsamkeit von Einzelbefunden* und Personenunterschieden zu belegen. Je nach Kontext übernehmen Testkennwerte demnach für verschiedene Anwendergruppen unterschiedliche und mehrfache Funktionen.

Testkennwerte finden sich - entsprechend der Standards für psychologisches und pädagogisches Testen (Häcker, Leutner & Amelang, 1998) - in annährend jedem Testmanual. Kurzbeschreibungen und Testrezensionen enthalten ebenfalls diese präzise und aussagekräftige Form der Testbeschreibung, womit sie eine zentrale Position innerhalb der gesamten Testbeschreibung einnehmen.

Was macht einen Kennwert praxisnah?

Testkennwerte können für einige der oben genannten Funktionen mehr oder weniger praktisch ‚gestaltet' sein. Der erste Artikel beschäftigt sich mit der Benutzbarkeit von Testkennwerten. Die Messung einer ‚Benutzbarkeit' erfolgt auf der Basis von insgesamt 14 Aspekten, die eine mehr oder weniger ‚umständliche' Verwendung anzeigen sollen. Beispielhaft werden zwei Kennwerte der Messgenauigkeit, der Reliabilitätskoeffizient und die Differenziertheit, bewertet, wobei sich deutliche Unterschiede ergeben. Der Artikel versucht damit das ‚Unbehagen', welches hinter der Vernachlässigung von Testkennwerten liegt, zu konkretisieren und damit eine konstruktive Verbesserung zu ermöglichen.

Beispiel für einen neuen Testkennwert

Im zweiten Artikel wird ein neuer Testkennwert, das Personenunterscheidungsvermögen, vorgestellt. Dieser verbindet die praxisrelevante Information der Messwertverteilung mit der Information der Messgenauigkeit. Um den Einfluss der Messwerteverteilung unter Berücksichtigung verschiedener Reliabilitäten und Stichprobengrößen zu veranschaulichen, wurden umfangreiche Simulationsstudien durchgeführt. Ein vom Autor erstelltes Softwareprogramm (SAS-Makro) soll die Berechnung des Koeffizienten durch andere Testautoren erleichtern; eine Anwendung findet sich bei Dr. Beauducel für die Analyse des IST-2000.

Die psychologische Bedeutung der Erstreckung einer Raschskalierung

Der dritte Artikel vertieft eine Überlegung zu einer bislang unbeachteten Eigenschaft einer Raschskalierung. Ausgangspunkt war die für die psychologische Diagnostik

grundlegende Fragestellung, inwieweit sich Personen mehr oder weniger ausgeprägt in einer psychologischen Eigenschaft unterscheiden. Die Überlegungen führen auf die Feststellung, dass die Erstreckung einer Raschskalierung als ein Maß für die Ausdehnung einer psychologischen Eigenschaft interpretiert werden kann. Die Möglichkeit der Übertragung des Ansatzes auf andere Item-Response-Modelle wird derzeit mit Kollegen diskutiert. Des Weiteren wird diskutiert, welche – nach statistischen und inhaltlichen Gesichtspunkten – ergänzenden Kennwerte abgeleitet werden können. In Gesprächen mit Prof. Dr. Steyer erscheint die einfache (nicht quadrierte) mittlere Abweichung vom Durchschnitt den statistischen Ansprüchen (u.a. erschöpfend an der Gesamtstichprobe bestimmt) zu genügen, als auch inhaltlich gut interpretierbar.

Überblick zu den entwickelten Testkennwerten

Das an die Einleitung angefügte Poster bietet nochmals eine graphische Zusammenfassung der drei Artikel, sowie zweier weitere Kennwerte (siehe Ausblick). Die Motivation zur Beschäftigung mit Testkennwerten wird mit drei Gründen benannt: zum ersten die kritisierte Praxis der Testauswahl, zum zweiten, aus dem themenübergreifenden Bemühen um eine Qualitätssicherung in der psychologischen Diagnostik und zum dritten aus neueren Entwicklungen und Ansätzen aus der Testtheorie. Das Vorgehen konzentriert sich zunächst auf eine ‚Ursachenanalyse' der geringen Beachtung psychometrischer Eigenschaften von Tests. Anschließend wurden aus der Ursachenanalyse heraus Zielvorgaben formuliert, die wiederum grundlegend für die Bewertung von Kennwerten (vgl. die Benutzbarkeitskriterien) sind. Des Weiteren wurden praxisrelevante Aspekte (u.a.

die Messwertverteilung) ausgewählt, deren Information in neue Kennwerte (Personenunterscheidungsvermögen) eingehen sollen. Zuletzt wird in der graphischen Übersicht versucht, die Vorschläge in Bezug zu den traditionellen Kennwerte einzuordnen. Ein mögliches Gliederungsmerkmal besteht in der Unterscheidung von sogenannten ‚Basisinformationen', die primär wissenschaftlich hinreichende Informationen darstellen, und hiervon zu trennen, die nach den ‚ergonomischen Leitlinien' optimierten Kennwerten. Diese Gliederung ist aber in jedem Falle vorläufig und keinesfalls zwingend, sondern sollte lediglich einen übergreifenden Aspekt der Konsequenzen aus den Überlegungen zu den benutzerfreundlichen Kennwerten andeuten.

Die in dieser Gesamtschrift vorliegenden Ausarbeitungen sind identisch mit den bei den Fachzeitschriften eingereichten Manuskripten. Es wurden lediglich formale Veränderungen vorgenommen, die eine bessere Lesbarkeit gewährleisten.

Zur Entstehungsgeschichte der Arbeit

Die ersten Überlegungen zu den Testkennwerten sind aus der Diplomarbeit entstanden, die eine mögliche Vorgehensweise (mit Hilfe von zeitreihenanalytischen Modellen) für eine einzelfallorientierte Testkonstruktion diskutiert. Aus den Überlegungen bezüglich der Änderungssensitivität eines Tests resultierten erste Ansätze für einen Kennwert zur Informativität. Die Bemühungen um eine Verbindung von Messtheorie und der Informationstheorie von Shannon und Weaver (1949) dauert noch an und wird zusammen mit Experten der Informationstheorie (Prof. Dr. Mittenecker) fortgesetzt. Aus der Beschäftigung mit dieser Thematik folgt der Kennwert des Personenunterscheidungsvermögens, der trotz der unterschiedlichen Konzeption

eine sehr hohe Verwandtschaft aufweist (beide verbinden die Messgenauigkeit und die Verteilungsform der Messwerte in einem Kennwert). In Gesprächen mit Kollegen zeigten sich Schwierigkeiten, die Vorteile des Personenunterscheidungsvermögens – beispielsweise gegenüber dem Reliabilitätskoeffizienten – darzustellen. Dies führte wiederum auf die Ausformulierung der Benutzbarkeitskriterien, die nun am Anfang der Gesamtarbeit stehen. Im Artikel zum Personenunterscheidungsvermögen waren es noch 5, im Artikel zur Benutzbarkeit von Testkennwerten schon 14 Kriterien. Aus diesem Grund ist die Bewertung des Personenunterscheidungsvermögens nur auf die dort genannten Kriterien begrenzt. Ebenso wurde der Begriff der ‚ergonomischen Leitlinien' durch den Begriff der ‚Benutzbarkeit' ersetzt.

An dieser Stelle werden einige Probleme einer publikationsgestützten Promotion erkennbar. Viele Detailprobleme, wie eine ausführliche Diskussion der Probleme um den Reliabilitätskoeffizienten, zu der es eine Vielzahl an kritischen Bemerkungen in der Literatur gibt, konnten nicht ausführlich dargelegt werden. In einem Artikel muss notgedrungen - aus Platzgründen - einiges voraussetzt werden, um hinreichend Raum für das Neue freizuhalten und dieses möglichst ausführlich diskutieren zu können.

Ausblick

Es liegt in der Natur des Themas, dass dieses mit den aufgeführten Artikeln noch nicht abschließend bearbeitet ist. Vielmehr kündigen sich weitere Beispiel für die Darstellung von praxisrelevanten Aspekten durch Kennwerte an (vgl. auch die Bemühungen innerhalb der Item-Response-Theorie um ein neues Maß der Messgenauigkeit auf Basis der Informationsfunktion; Rost, 2000). An dieser Stelle sei

abschließend auf zwei Kennwerte des Autors verwiesen, die konzeptuell schon auf dem Kongress der DGPS 2000 in Jena vorgestellt wurden: Der Ausschöpfungsquotient, der den für eine Stichprobe genutzten Messwertbereich zum Gesamtmesswertbereich als Prozentangabe abbildet, und die Operationalisierung des Verhältnisses von Messaufwand zur Messgenauigkeit (vgl. Poster). Für diese und weitere Kennwerte von psychologischen Testverfahren kann zukünftig gefordert werden, dass sie die Kriterien der Benutzbarkeit nach Möglichkeit berücksichtigen, um Neuentwicklungen nicht am Testanwender ‚vorbei' einzuführen.

Literatur

Dorsch, F. (1991): *Psychologisches Wörterbuch*. Bern: Huber.

Häcker, H., Leutner, D. & Amelang, M. (Hrsg.) (1998): Standards für pädagogisches und psychologisches Testen. *Diagnostica*, Suppl.1.

Müller, J. M. (2000): *Neue Leistungs- und Effizienzkennwerte für psychologische Testverfahren: Breite, Differenziertheit, Personenunterscheidungsvermögen, Effizienz und Ausschöpfungsquotient*. Poster auf dem Kongress der DGPS in Jena.

Rost, J. (1996): *Lehrbuch der Testtheorie, Testkonstruktion*. Bern: Huber.

Rost, J. (1999): Was ist aus dem Rasch-Modell geworden? *Psychologische Rundschau*, *50,* 140-156.

Rost, J. (2000): Haben ordinale Raschmodelle variierende Trennschärfen? Eine Antwort auf die Wiener Repliken. *Psychologische Rundschau*, *51*, 36-37.

Shannon, C. E. & Weaver, W. (1949): *The mathematical theory of communication*. Urbana: University of Illinois Press.

Testkuratorium der Föderation für die Testbeurteilung (1986): Beschreibung der einzelnen Kriterien für die Testbeurteilung, *Diagnostica*, 32, 358-360.

Poster zum 34. Kongress der Deutschen Gesellschaft für Psychologie in Jena, 2000

Neue Leistungs- und Effizienzkennwerte für psychologische Testverfahren: Breite, Differenziertheit, Personenunterscheidungsvermögen, Effizienz und Ausschöpfungsquotient

Jörg M. Müller
Universität Bremen

Motivation I

Verwendung von Testkennwerten in der Testpraxis

- Umfragen bei Testanwendern zeigen, dass die Gütekriterien (Testkuratorium, 1986) zur Testbeschreibung bei der Auswahl von Testverfahren nur eine untergeordnete Rolle spielen (Steck, 1997; Schorr, 1995; Wade & Baker, 1977, für den amerikanischen Sprachraum).

Motivation II

Qualitätssicherung

- Standardisierte Form der Testbeschreibung
- Umsetzung der Bedürfnisse der Testpraxis in Kennwerte
- Berücksichtigung ergonomische Leitlinien
- Gewährleistung der Nutzung der Testbeschreibung für eine angemessene Testauswahl

Motivation III

Vernachlässigte Informationen:

- Informationsmenge einer Testung
- Anomalien der Messwertverteilung
- Besondere Merkmale der Eigenschaft
- Ausnutzung des Messaufwandes

Konsequenz I

Ursachenanalyse psychometrischer Kennwerte

- *Indirekter Anwendungsbezug* - Notwendigkeit zur Transformation
- *Problematische Skalierung*
  - Keine ‚sprechende' Maßeinheit
  - Unhandliche Skalierung (nicht linear und Dezimalbrüche)
- *Theoretischer Vorwissen* beim Testanwender
- *Zum Fremdenweg von Messungsfreiheit* ungeeignet

Konsequenz II

Ergonomische Leitlinien für Testkennwerte

- *Praxisbezug*: ohne weitere Umrechnungen, direkt aussagekräftig
- *Einheitlich-Vergleichbar*: unabhängig von test- und eigenschafts-spezifischen Besonderheiten bestimmbar
- *Skalierung des Kennwertes*: Äquidistant , aussagekräftige Maßeinheit, fehlerfreie Handhabung
- *Minimale Voraussetzungen* beim Testanwender

Konsequenz III

Berücksichtigung neuer Information

- *Test-Subjektsynchron-Passung* Ausnutzung der Messwertskala
- *Verteilungsform*: der Messwerte und deren Einfluss auf die Messmöglichkeiten in der Praxis
- *‚Ausdehnung'*: einer Eigenschaft (s. u.; im Titel ‚Breite')
- *Aufwand*: einer Messung
- *Informationsmenge*: einer Messung

5 neue Testkennwerte

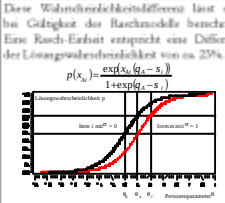
Ausdehnung

Verwechselung der ‚Ausdehnung' anhand der Schach-Analogie

Zwei Personen spielen mehrere Schachpartien. Liegt die Gewinnwahrscheinlichkeit von Spieler A mit Spieler B bei 2:1, so definieren wir Spieler A, der im Durchschnitt zwei von drei Partien gewinnt, als eine Stufe besser als Spieler B. Lässt man nun in gleicher Weise die gesamte schachspielende Population mehrfach gegeneinander antreten, dann würden sich ‚Stufen' von ‚Spieler X besser als Spieler Y' bilden (=Skalierung einer Eigenschaft über Gewinnwahrscheinlichkeiten) bei festgelegter Wahrscheinlichkeitsdifferenz als Maßeinheit. Die Anzahl der Wahrscheinlichkeitsdifferenzen ist für jede psychologische Eigenschaft unterschiedlich und ein Maß für die Ausdehnung der Eigenschaft.

Diese Wahrscheinlichkeitsdifferenz lässt sich bei Gültigkeit der Raschmodelle berechnen. Eine Rasch-Einheit entspricht eine Differenz der Lösungswahrscheinlichkeit von ca. 25%.

$$p(x_{si}) = \frac{\exp(x_{si}(q_A - s_i))}{1+\exp(q_A - s_i)}$$



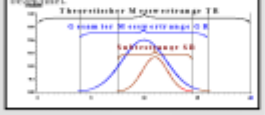
Die ‚Ausdehnung einer Eigenschaft bestimmt sich über die Erstreckung einer Raschskalierung.

$$A = q_{max} - q_{min}$$

Ausschöpfung

Die Vorgabe der Standardversion eines Tests stellt trotz der Ansätze um ein adaptives Testen noch immer die übliche Art der Testvorgabe dar, obwohl Subgruppen (z.B. im klinischen Bereich oder in der Leistungsmessung) zu viele ‚zu leichte' oder ‚zu schwere' Items bearbeiten. Wie kann die Ausnutzung des Aufwandes für eine Untergruppe in einem Kennwert erfasst werden?

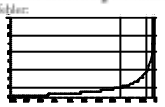
Charakteristisch für die praktisch Verwendung von Tests ist, dass bestimmte Subgruppen nur ein Teil der Messwertskala nutzen. Diese ‚Passung' zwischen Subjektgruppe und Test kann über das Verhältnis von *SR* zu *TR* ausgedrückt werden und wird als Ausschöpfungsquotient (abgekürzt *AQ*) bezeichnet.



$$AQ = \frac{SR}{TR} * 100$$

Differenziertheit

Der *nicht linearer* Zusammenhang von Reliabilität und Messfehler:



Die Differenziertheit ist eine lineare Funktion des Messfehlers (s.o.).

Die Differenziertheit bestimmt sich über die Teilung der Range der Testskala (R) durch die Distanz der kritischen Differenz (k) (vgl. auch Lehl & Kienzel, 1973; sowie Wright & Masters ‚Number of Person Strata', 1982). Sie verfügt über eine ‚sinnvolle' und praktisch interpretierbare Skaleneinheit. Sie ist zudem äquidistant und eine Maß für den Effekt der Reliabilität auf den Messfehler. Das Maß *k* wird innerhalb der Klassischen oder Probabilistischen Testtheorie bestimmt.

$$D = \frac{R}{k}$$

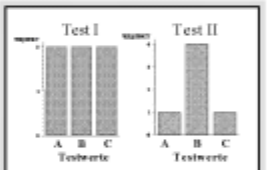
Mess-Effizienz

Die Messung der Effizienz definiert sich über die Leistung je Aufwand. Transport-Analogie: Im Bereich des Transportwesens kann die Leistung z. B. in einer zurückgelegten Distanz bestehen, oder in der Beanspruchung von Energie (kWh bzw. PS). Der Aufwand kann im Verbrauch der Treibstoffe für diese Leistung gemessen werden, wodurch eine Schätzung der Effizienz möglich wird. Wenn besteht die ‚Leistung' bzw. der Aufwand eines psychologischen Tests?

Innerhalb einer Kommunikationssituation (als solche kann eine Testung aufgefasst werden) interessiert die Informationsmenge als ‚Leistung'. Neben anderen Alternativen (zum Beispiel die Informativität nach Müller, 1998) sollen zwei Basisinformationen berücksichtigt werden: Die Differenziertheit (D) und die Ausschöpfung (*AQ*), um auf eine Schätzung der erhaltenen Information zu gelangen. Der Zeitaufwand (*T*) dient als einfacher Schätzer für den Aufwand.

$$E = \frac{D * AQ}{T_{[min]}}$$

Personen-unterscheidungs-vermögen



Es wird innerhalb jeder Verteilung ein vollständiger Vergleich aller Personen (kU) durchgeführt. Zählt man die Anzahl der signifikanten Paarvergleiche zusammen (sU), dann ergeben sich für die Verteilungsform I 12 und für Verteilungsform II nur 9 Unterscheidungen.

Das Personenunterscheidungsvermögen gibt eine Antwort auf die Frage:

‚Wie groß ist die Wahrscheinlichkeit, mit Hilfe eines Tests zwischen zwei zufällig ausgewählten Testpersonen einen signifikanten Unterschied festzustellen?'

Die kritische Differenz *k* kann innerhalb der Klassischen oder Probabilistischen Testtheorie ermittelt werden, weshalb sich die angestrebte Unabhängigkeit ergibt. Mit Hilfe des PUV-Wertes kann ein Aspekt der Messmöglichkeiten eines Tests unter Berücksichtigung der Verteilungsform bestimmt werden.

$$PUV = \frac{sU}{tU} * 100$$

$$tU = \frac{n*(n-1)}{2}$$

$$sU = \sum_{i,j} s_{ij} \begin{cases} s_{ij} = 1, \text{ wenn } x_i - x_j \le k \\ s_{ij} = 0, \text{ wenn } x_i - x_j > k \end{cases}$$

Kennwerte der 1. Generation:

Diese Art Kennwert enthält Basisinformationen über einen Test aus der Testtheorie (z.B. die Reliabilität) und rein deskriptive Merkmale, wie Verteilungsform; Range/Streuung; Anzahl der Items; Zeitaufwand, etc.

Kennwerte der 2. Generation:

Diese Art Kennwert basieren auf den Kennwerten der 1. Generation und bereiten deren Informationen für die Praxis nach ergonomischen Leitlinien auf, wie zum Beispiel im Falle des PUV (Reliabilität und Verteilungsform).

Bewertung/Diskussion

Zum momentanen Zeitpunkt ist lediglich eine theoretische Bewertung möglich, wenngleich eine Evaluation in der Testpraxis die Anwenderfreundlichkeit und Nutzung der Kennwerte belegen muss. Eine ergonomisch optimierte Skalierung (Linearität; bedeutungshaltige Maßeinheiten) der Kennwerte konnte, mit Ausnahme der Effizienz, bei allen Kennwerten erreicht werden. Die Ausdehnung bleibt (noch) auf raschskalierte Tests beschränkt, da eine Verallgemeinerung auf weitere Item-Response-Theorien aussteht. Die vorgestellten Kennwerte eröffnen neue Wege und Ansätze Tests in der Praxis gezielt zu beschreiben, um letztlich eine Qualitätssicherung im Sinne der Testanwender zu fördern.

Literatur:
Lehl, S. & Kienzl, W. (1973): Die Standardbreite der kritischen Differenz. Diagnostica, 19, 75-88.
Müller, J.M. (1998): Erörterungen zur Informativität und deren Implikationen für die Konstruktion von psychologischen Merkmalsparametern. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 19, 43.
Schorr, A. (1995): Stand und Perspektiven diagnostischer Verfahren in der Praxis. Ergebnisse einer repräsentativen Befragung westdeutscher Psychologen. Diagnostica, 41, 3-20.
Steck, P. (1997): Zur der Arbeit der Testkuratorium. Psychologische Testverfahren in der Praxis. Diagnostica, 43, 267-284.
Testkuratorium der Föderation für die Testanwendung (1986): Erweiterung der erweiterten Kriterien für die Testanwendung. Diagnostica, 32, 358-360.
Wade, T. C. & Baker, T. B. (1977): Opinions and use of psychological test. *American Psychologist*, 32, 874-882.
Wright, B. & Masters, G. (1982): *Rating Scale Analysis. Rasch Measurement*. Chicago: MESA Press.

Die Benutzbarkeit von Testkennwerten

am Beispiel der Messgenauigkeit

ABSTRACT

Surveys in using psychodiagnostic tests shows that testusers do not concern much about the psychometric properties of a test. A theoretical founded analysis of the psychometric coefficients is recommended. The approach of ergonomie (Murrell, 1971) is used to compile a list of 14 criteria to evaluate the usability (Mayhew, 1999) of psychometric admeasurements, refering to the foundation, scale and interpretation of a coefficient. Two coefficient of the error of measurement, the reliability coefficient and the Differenziertheit, are evaluated on the basis of these criteria. The assets and drawbacks of the approach are discussed and consequences for the conceptualization of psychometric coefficient are shown.

Keywords: error of measurement, usability, test selection, diagnostic, reliability.

ZUSAMMENFASSUNG

Umfragen bei Testanwendern zeigen, dass die psychometrische Qualität eines Tests die Testauswahl nur wenig beeinflusst. Ein Erklärungsversuch für die geringe Beachtung der psychometrischen Eigenschaften bei der Testauswahl erfolgt mit Hilfe des Ansatzes der Ergonomie (Murrell, 1971). Insgesamt werden 14 Kriterien bezüglich der Grundlagen, Skalierung und Interpretation von Testkennwerten aufgestellt, welche für die Bewertung der Benutzbarkeit (Usability, Mayhew, 1999) herangezogen werden können. Zwei Maße der Messgenauigkeit, der Reliabilitätskoeffizient und die Differenziertheit, werden anhand dieser Kriterien analysiert. Der Ansatz der Benutzbarkeit sowie die Konsequenzen für die Konzeption von Testkennwerten werden diskutiert.

Stichwörter: Messfehler, Testkennwerte, Testtheorie, Testpraxis, Ergonomie, Qualitätssicherung

DIE PRAXIS DER TESTAUSWAHL

Umfragen bei Testanwendern weisen seit längerem darauf hin, dass die Gütekriterien (Testkuratorium, 1986) bei der Auswahl von Testverfahren in der Praxis nur eine nachgeordnete Rolle spielen (Grubitzsch & Rexilius, 1978; Kubinger, 1997; Tent, 1991; Schorr, 1995; Schulz, Schuler & Stehle, 1985; Steck, 1997; Stoll, 1978). Eine vergleichbare Situation zeigt sich auch außerhalb des deutschen Sprachraums (Archer, Maruish, Imhof & Piotrowski, 1991; Frauenhoffer, Ross, Gfeller, Searright & Piotrowski, 1998; Piotrowski, Belter & Keller, 1998; Piotrowski & Keller, 1992; Wade & Baker, 1977). Diese Bestandsaufnahme weist auf mögliche grundlegende Schwierigkeiten im Umgang mit den Kennwerten hin.

Im vorliegenden Beitrag soll ein möglicher Grund für die geringe Beachtung der Testkennwerte durch die Testanwender analysiert[1] werden. Vorarbeiten bezüglich der Einbeziehung des Testanwenders in die Gestaltung und Beschreibung von Tests existieren von Müller-Böling (1991), Lehfeld und Erzigkeit (1993) und Steck (1991). Die Lesbarkeit (Readability) von Testmanualen behandeln Schinka und Borum (1994; Junga, 1979). Es werden in den genannten Arbeiten jedoch keine konkreten Kriterien für die Bewertung der Benutzbarkeit von Testkennwerten formulierten. Derartige Kriterien sind jedoch unerlässlich, wenn spezifische Schwächen eines Kennwertes dargestellt werden sollen. Um die benötigten Kriterien aufzustellen, wird zu Beginn der Ansatz der Ergonomie (Murrell, 1971) erläutert. Mit den konzeptuellen

[1] Gründe, die den Einsatz eines Tests trotz Kritik (Angleitner, 1997) rechtfertigen, werden nicht besprochen (vgl. die Verwendung des MMPI als Kommunikationshilfe, Engels, 1997).

Vorarbeiten aus anderen Gebieten der Ergonomie werden anschließend Benutzbarkeitskriterien formuliert. Zwei Maße der Messgenauigkeit, der Reliabilitätskoeffizient und die Differenziertheit, werden anhand dieser Kriterien genauer betrachtet.

ERGONOMIE IN DER TESTBESCHREIBUNG

Der Begriff der Ergonomie entstammt ursprünglich aus den Arbeitswissenschaften (Behr, 1996; Bubb & Bubb, 1996; Murrell, 1971; Schmidtke, 1993) und wurde darauffolgend in der Softwarebranche (Mayhew, 1999) eingeführt, um Gestaltungsrichtlinien für die Programmentwicklung aufzustellen. In beiden Bereichen wurde eine Mensch-Technik-Schnittstelle zugrunde gelegt, die prinzipiell auch in der psychologischen Diagnostik (Testanwender-Test-Interaktion) vorliegt. Es stellt sich damit die Frage, inwieweit Testkennwerte den Testanwender in funktionaler Weise bei der Testauswahl unterstützen. So wurde zum Beispiel der Reliabilitätskoeffizient ursprünglich nicht für diesen Anwendungsbereich konzipiert (Rost, 1996). Testkennwerte resultierten in einigen Fällen aus dem zugrunde gelegten Messmodell und genügen in erster Linie wissenschaftlichen Ansprüchen. Der Ergonomiegedanke hingegen betont die Bedeutung des Testanwenders bei der Testgestaltung (und den Testkennwerten) und soll die Bemühungen zur Reduzierung von Anwendungsfehlern unterstützen.

Die ‚mangelhafte Vermarktung' psychologischer Produkte wurde von Spada (1997; ebenso Kluwe, 2001) in ihren Berichten zur Lage der Psychologie betont. Dies muss

in gleicher Weise auch innerhalb der Psychologie gelten, denn Testkennwerte sind in gewisser Weise ein Produkt psychologischer Methodenforschung, welche von Testanwendern genutzt werden. Die bekannten Gütekriterien müssen dabei nicht grundsätzlich in Frage gestellt werden, aber mögliche Schwächen in der Interpretier- und Benutzbarkeit sollten analysiert und – falls notwendig - entsprechende Alternativen oder Ergänzungen erarbeitet werden.

BENUTZBARKEIT VON TESTKENNWERTEN

Grundlegend für die Analyse von Testkennwerten ist die Annahme, dass Kennwerte mehr oder weniger dem Testanwender helfen, sich schnell über relevante Aspekte eines Tests zu informieren. Umrechnungen eines Kennwertes machen zum Beispiel die Verwendung umständlich, zeitintensiv und fehleranfällig. Es sollte deshalb möglich sein, eine Reihe von Kriterien aufzustellen, die derartige Probleme im Umgang mit Kennwerten nachgehen und letztlich messbar machen.

Die Bewertung der ‚Benutzbarkeit' ist nur für *eine* Verwendungsweise, in unserem Falle die Testauswahl in der Praxis, aussagekräftig. In anderen Kontexten (z. B. der Verwendung von Tests innerhalb empirischer Studien) werden wahrscheinlich andere Ansprüche an Kennwerte gestellt. Die Erstellung von Kriterien der Benutzbarkeit (Wandmacher, 1993; engl. Usability; Bennet, 1894; Eason, 1984; Nielsen & Mack, 1994) lehnt sich u.a. an die Arbeiten von Mayhew (1999) und Galer (1987) an und berücksichtigt für die Aufstellung von Kriterien der Benutzbarkeit drei Perspektiven:

a) der Anwender (User-Profile),

b) die Aufgabe bzw. die Funktion der Testkennwerte (Task-Analysis) und

c) die Rahmenbedingungen(Platform-Constraints).

Aus allen drei Betrachtungsebenen ergeben sich Anforderungen für die Gestaltung von Kennwerten. Der Begriff der ‚Benutzerfreundlichkeit' entspricht damit dem Verständnis von Wandmacher (1993, S.200; Czida, Herda & Itzfeld, 1978) als messbare Dimension, dem verschiedene Einzelaspekte zugeordnet sind.

Anwenderprofil

Die Gruppe der *Testanwender* (vorwiegend Psychologen) wurde bei den Umfragen von Steck (1997) und Schorr (1995) beschrieben. Es blieb in diesen Umfragen allerdings offen, über welche Grundlagenkenntnisse die Testanwender verfügen. Ohne dies im Einzelnen analysieren zu müssen, ist anzunehmen, dass die Vorkenntnisse eine bedeutende Rolle für das Verständnis von Kennwerten spielen. Prinzipiell sollten Kennwerte möglichst einfach und verständlich sein – da mit der Menge notwendigen Wissens die Gefahr der fehlerhaften Verwendung steigt. Die Anforderungen an die ‚Interpretierbarkeit eines Kennwertes' berücksichtigen diesen Punkt.

Aufgabenanalyse

Die *Aufgabenanalyse* kann über die Beschreibung des Auswahlprozesses (Abb. 1) relevante Aspekte der Testauswahl aufdecken, wobei am Ende des Prozesses schließlich zwei Entscheidungssituationen auftreten können:

a) *Einzelbewertung*: ‚Welche Eigenschaften muss ein Test mindestens aufweisen?' und die

b) *Vergleichende Bewertung* von Tests: ‚Welches ist der bessere Test?'.

Aus beiden Entscheidungssituationen folgen Anforderungen an die Vergleichbarkeit von Testkennwerten und deren Skalierung (vgl. Kasten 1 und 2).

Rahmenbedingungen

Die Rahmenbedingungen in Abbildung 1 zeigen, dass psychometrischen Eigenschaften ein Kriterium neben anderen sind. Testkennwerte scheinen innerhalb des Informationsangebotes (Testmanuale, Testrezensionen, Handbuch psychologischer und pädagogischer Tests von Brickenkamp, 1997; Testkatalog der Testzentrale, 2000; ZIS von Glöckner-Rist & Schmidt, 1999) eine geeignete – weil präzise, objektive und valide - Ergänzung der Testbeschreibung darzustellen.

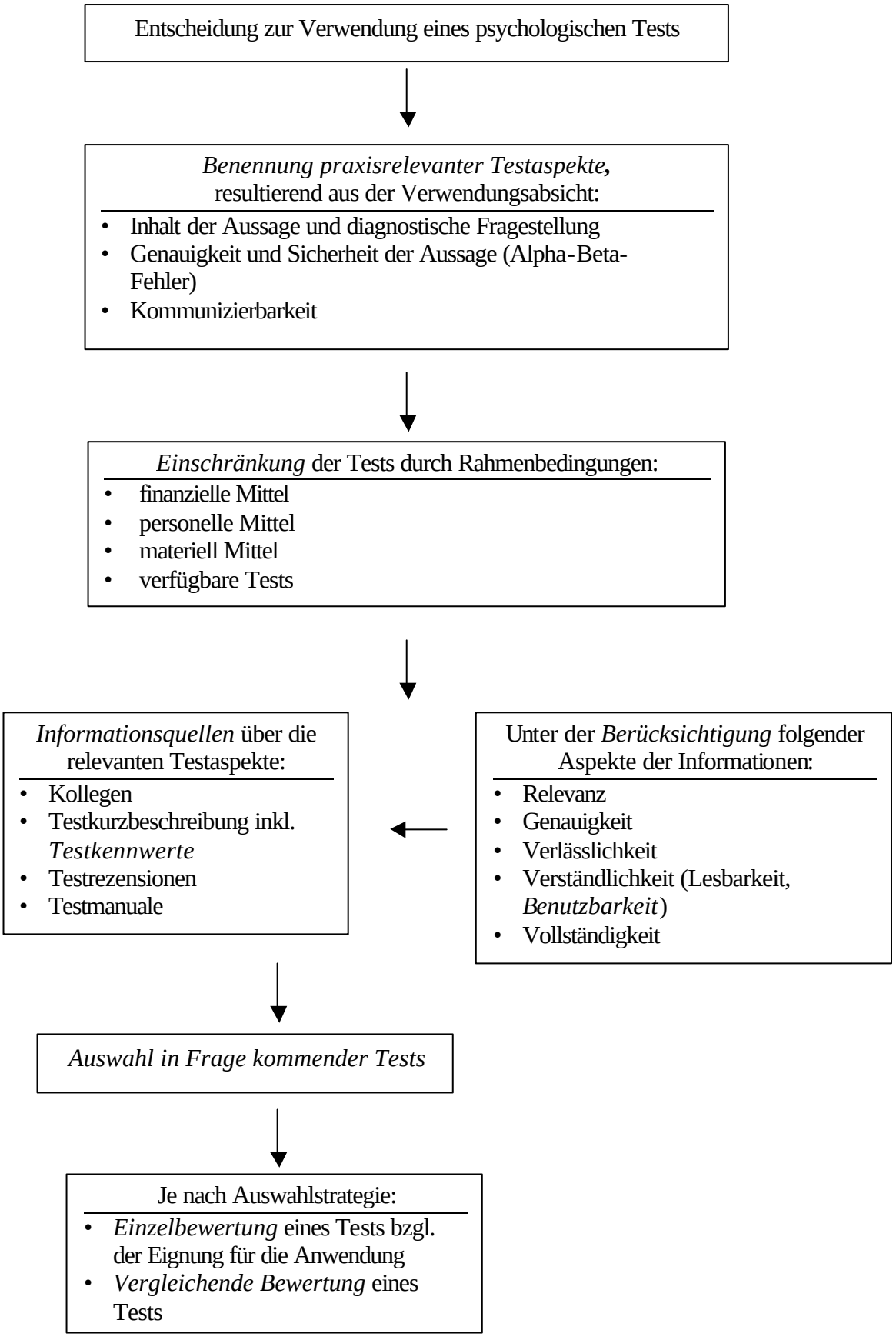
Auf der Grundlage der Betrachtungen auf allen drei Ebenen folgt nun die Aufstellung von Benutzbarkeitskriterien, die ein Testkennwert erfüllen sollte. Die Richtlinien sind in drei Bereiche unterteilt:

a) Theoretische Fundierung und Grundlage des Kennwertes (Kasten 1)

b) Skalierung des Kennwertes (Kasten 2)

c) Interpretierbarkeit des Kennwertes (Kasten 3)

Abbildung 1. Auswahlprozess für Testverfahren



Die wissenschaftliche Basis oder Fundierung eines Testkennwertes basiert zumeist auf einer Messtheorie, die die ‚logische Struktur' des psychologischen Konstruktes berücksichtigt. Steyer und Eid (1993) unterscheiden hier ‚klassifikatorische', ‚komparatorische' und ‚metrische' Begriffe, wobei letztere häufig in der Persönlichkeitspsychologie (z.B. Intelligenz) anzutreffen sind und klassifikatorische Begriffe meist im klinischen Bereich (z.B. ICD-10 der WHO, 1991) eine besondere Rolle spielen. Hieraus resultieren Probleme für den Anwender, was die Vergleichbarkeit der Grundlagen eines Kennwertes angeht. Die Richtlinien berücksichtigen deshalb zum einen die Fundierung eines Kennwertes, versuchen aber auch die Notwendigkeit der Vergleichbarkeit und methodische Aspekte der Schätzung herauszustellen.

KASTEN 1: ANFORDERUNGEN AN DIE GRUNDLAGEN FÜR DIE VERGLEICHBARKEIT VON KENNWERTEN

1. Eindeutigkeit: Der Kennwert sollte eine eindeutige Berechnungsgrundlage besitzen. Diese Eigenschaft ist erfüllt, wenn ein eindeutiger Algorithmus vorliegt.

2. Vergleichbarkeit: Der Kennwert sollte für möglichst viele Tests bzw. Messmodelle ermittelt werden können.

3. Relevante Abhängigkeit: Alle praxisrelevanten Einflüsse der Testanwendung sollten den Kennwert beeinflussen.

4. Unabhängigkeit von irrelevanten Einflüssen: Alle sonstigen Bedingungen sollten den Kennwert nicht beeinflussen.

Es folgen nun allgemeine Kriterien bezüglich der Skalierung eines Kennwertes (vgl. Kasten 2). Die Skalierung eines Kennwertes ergibt sich zumeist direkt aus dem Messmodell. Transformationen des Kennwertes (für eine höhere Benutzbarkeit) erscheinen aus mathematischer Sicht überflüssig, da sie keinen Informationsgewinn darstellen. Anders fällt das Urteil aus, wenn der Benutzer in die Überlegungen mit einbezogen wird und positive Eigenschaften einer Skalierung benannt werden, die eine Transformation rechtfertigen können. Es ist zum Beispiel nicht davon auszugehen, dass nummerische Ausprägungen eines Kennwertes bei der Testauswahl vollständig ‚mental repräsentiert' und verarbeitet werden, sondern menschliche Fehler der Informationsverarbeitung auch in diesem Kontext eine Rolle spielen.

Obwohl die Skalierung mit dem ergonomischen Begriff des Displays (Schmidtke, 1993) eine gewisse Verwandtschaft aufweist, finden sich wenig übertragbare Richtlinien, wann eine Skalierung eine hohe Benutzbarkeit aufweist. Ebenso wenig ergiebig zeigte sich die Suche nach DIN oder ISO-Normen, die Richtlinien für die (nach wahrnehmungs- und kognitionspsychologischen Gesichtspunkten) optimierte Skalierung von Kennwerten zum Inhalt haben. Dutke (1994) weist zumindest auf die Bedeutung der ‚Analogie' für den Erwerb und Erlernbarkeit von mentalen Modellen hin. Demzufolge wäre wahrscheinlich der ‚Meterstab' der Prototyp einer Skalierung, den jeder kennt und dessen Eigenschaft (möglicherweise vorschnell) auf neue Skalierungen übertragen wird – und daher zu Fehlschlüssen verleitet.

Aufgrund dieses Mangels sollen Vorschläge über die Eigenschaften einer benutzerfreundlichen Skalierung gemacht werden (vgl. Kasten 2). Es ist noch einmal

zu betonen, dass die Richtlinien sehr allgemeine Anforderungen darstellen, die im speziellen Fall unterschiedlich bedeutsam sein können.

KASTEN 2: ANFORDERUNGEN AN DIE SKALIERUNG DES TESTKENNWERTES

5. Positive Zahlen sind negativen vorzuziehen.

6. Ganze Zahlen sind Dezimalbrüchen vorzuziehen.

7. Die Anzahl möglicher Ausprägungen sollte überschaubar bleiben. Die Vorgabe von 100 Ausprägungen soll als Kompromiss von zu wenigen vs. zu vielen Ausprägungen aufgefasst werden.

8. Intervallskalenniveau: Der Kennwert sollte mindestens *intervallskaliert* bezüglich des interessierenden Aspektes sein. Diese Eigenschaft liegt vor, wenn keine Umrechnungen erforderlich werden.

9. Signifikante Einheiten: Nummerische Unterschiede der Testkennwerte sollten bei einem Vergleich von Tests nicht auf die Ungenauigkeit der Schätzung des Kennwertes zurückgeführt werden können. Die Eigenschaft ist erfüllt, wenn die kleinstmöglichen nummerischen Unterschiede auch *statistisch signifikant* sind.

Über die Anforderungen an die Skalierung des Kennwertes können noch weitere wünschenswerte Eigenschaften bezüglich der Interpretierbarkeit des Kennwertes aufgestellt werden (vgl. Kasten 3).

KASTEN 3: RICHTLINIEN FÜR DIE INTERPRETIERBARKEIT VON TESTKENNWERTEN

10. Relevanz: Der Testkennwert sollte einen praxisrelevanten Inhalt abdecken.

11. Unmittelbarer Bezug: Der Kennwert sollte direkt über die relevante Testeigenschaft informieren. Die Eigenschaft ist erfüllt, wenn *keine Transformationen* notwendig werden.

12. Maßeinheit: Der Kennwert sollte eine *interpretierbare Maßeinheit* aufweisen. Dies lässt sich dadurch feststellen, dass keine sinnerhaltende Transformation existiert[2].

13. Angabe der notwendigen Höhe: Der Kennwert sollte sich dazu eignen, die zur Bearbeitung einer diagnostischen Fragestellung *notwendige Höhe* anzugeben.

14. Erlernbarkeit: Die Grundlagen eines Testkennwertes sollten möglichst einfach und leicht verständlich sein (vgl. die Begriffsbestimmung der Erlernbarkeit von Softwareprogrammen nach Polson & Lewis, 1990). Folgende (subjektiv bewertete) Voraussetzungsstufen werden unterschieden:

a) Grundlegende mathematische Kenntnisse

b) Höhere Mathematik

c) Messfehlertheorie

d) Item-Response-Theorien (IRT)

[2] Prozentangaben können zum Beispiel nicht transformiert werden, ohne ihre Aussage zu verlieren.

Nachdem Richtlinien für die Bewertung der Benutzbarkeit von Testkennwerten aufgestellt wurden, sollen nun verschiedene Möglichkeiten zur Darstellung der Messgenauigkeit angesprochen und hiervon zwei bezüglich ihrer Benutzbarkeit bewertet werden. Alle vorgeschlagenen Kriterien werden dabei gleich gewichtet, da die Bedeutung für die resultierende Benutzbarkeit bislang ungeklärt ist.

VERSCHIEDENE MAßE DER MESSGENAUIGKEIT

Mit dem Begriff der Messgenauigkeit verbindet sich zunächst noch kein konkretes Maß, da verschiedene Möglichkeiten für die Bildung von Kennwerten bestehen. Der Reliabilitätskoeffizient ist damit nur eine unter mehreren der folgenden Alternativen. Rost schreibt hierzu (1996, S.34): ‚*Reliabilität im engeren Sinne* meint jedoch eine bestimmte Definition von *Meßgenauigkeit*, die nicht die einzig mögliche ist und auch nicht bei jedem Testmodell Sinn macht.' (Hervorhebung im Original). Demzufolge gibt es mehrere Möglichkeiten zur Bildung von Maßen bezüglich der Messgenauigkeit:

a) Der *Reliabilitätskoeffizient*

- Variante a: Klassisch nach Kelley (1921, 1942; Gulliksen, 1950; Kuder & Richardson, 1937; vgl. Formel 1; m_w = wahrer Wert, m_x = empirischer Wert, m_e = Fehler, r = Reliabilität) innerhalb der Messfehlertheorie oder
- Variante b: Nach Formel 2 mit einer über die Personen gemittelten, personenspezifischen Messgenauigkeit innerhalb der IRT (nach Rost,

1996, S. 354; vgl. Formel 3 und 4; $E_?$ = Erwartungswert der Parameters der Person *v*, N = Stichprobengröße, p_{vi} = Lösungswahrscheinlichkeit der Person *v* bei Item *i*).

Formel 1

$$r = \frac{\mathrm{var}(m_w)}{\mathrm{var}(m_x)}$$

Formel 2

$$\mathrm{Re}\, l(q) = 1 - \frac{\sum_{v=1}^{N} Var(E_{q_v})}{N \cdot Var(\hat{q})}$$

Formel 3

$$Var(E_{q\,v}) = \frac{1}{\sum_{i=1}^{k} p_{vi}(1 - p_{vi})}$$

Formel 4

$$Var(E_q) = \frac{\sum_{v=1}^{N} Var(E_{q_V})}{N}$$

b) Eine standardisierte Form des *Standardmessfehlers* m_e:

- Variante a: In Formel 5 wird die Standardabweichung auf eins gesetzt.

Formel 5

$$s_{m_e} = s_{m_x}\sqrt{1 - r}$$

- Variante b: Standardisierung nach Wright und Masters (1982; Person-Seperation-Index; vgl. Formel 6).

Formel 6

$$G_P = \frac{\sqrt{\mathrm{var}(m_w) - \mathrm{var}(m_e)}}{\sqrt{\mathrm{var}(m_e)}}$$

c) Die *Anzahl unterscheidbarer Testergebnisse (AuT);* Messmodell übergreifende Konzeption (vgl. die Ausführungen zur Differenziertheit im Text).

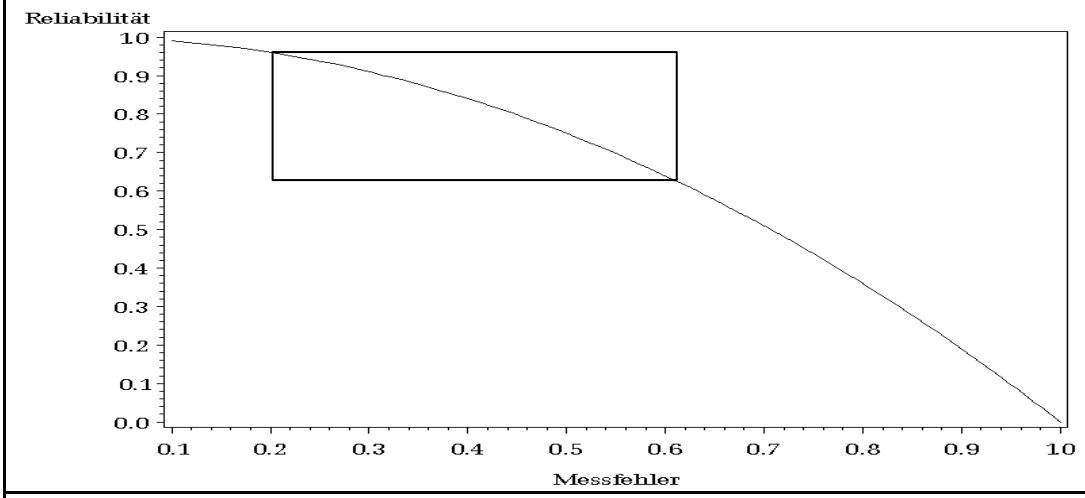
d) Das *Personenunterscheidungsvermögen* nach Müller (2000), Messmodell übergreifende Konzeption;

e) Die *Informativität* nach Müller (2000), Messmodell übergreifende Konzeption.

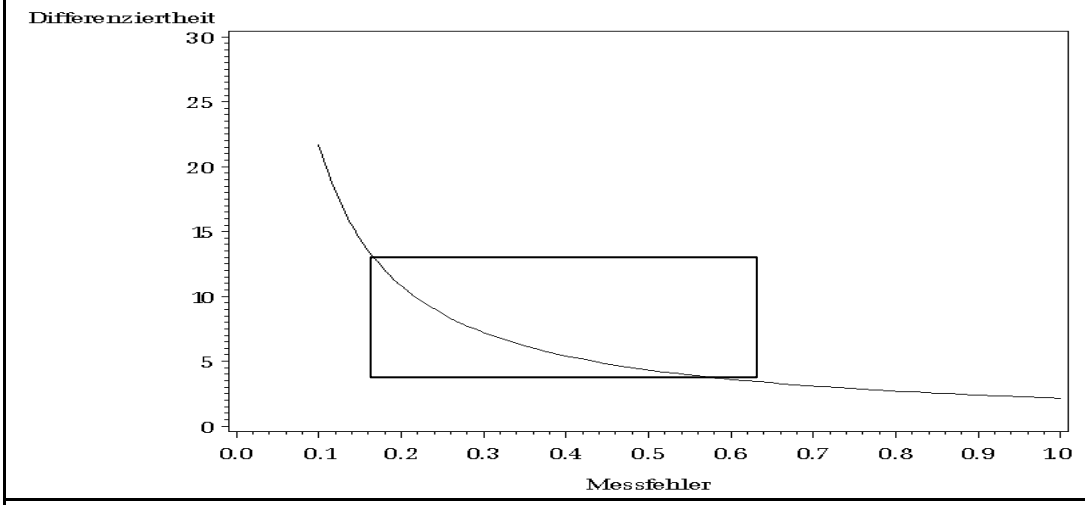
f) *Summation der Informationswerte* pro Person je Item (vgl. Rost, 2000; bislang nicht formal definiert) innerhalb der IRT.

Die Betrachtung der Zusammenhänge der drei erstgenannten Indikatoren für die Messgenauigkeit zeigt, dass diese nicht linear sind (vgl. Abb. 2a,b,c). Für einen Testanwender sind die verschiedenen Maße deshalb nicht einfach ineinander überführbar. Eine Präferenz für die eine oder andere Alternative könnte deshalb die verschiedenen Aspekte der Benutzbarkeit berücksichtigen.

a) Zusammenhang von Reliabilität und Standardmessfehler



b) Zusammenhang von Differenziertheit und Messfehler



c) Zusammenhang von Differenziertheit und Reliabilität

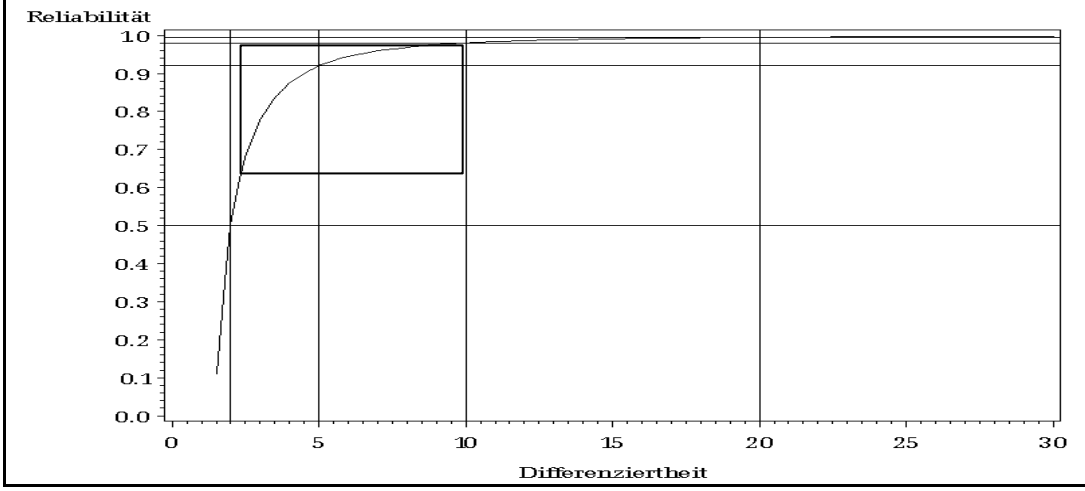


Abbildung 2a,b,c. a (in allen Abbildungen ist das praxisrelevante Spektrum durch ein graues Rechteck hervorgehoben).

BEWERTUNG DES RELIABILITÄTSKOEFFIZIENTEN

Der Reliabilitätskoeffizient als bekanntester Vertreter zur Darstellung der Messgenauigkeit wird im Folgenden in seiner klassischen Variante bezüglich seiner Benutzbarkeit evaluiert:

ad 1+2 (Eindeutige Berechnungsgrundlage, Vergleichbarkeit)

Der Reliabilitätskoeffizient ist trotz seiner Definition (Formel 1) ein ‚Gattungsbegriff' (Häcker, Leutner & Amelang, 1998; Gulliksen, 1950), da die Schätzung mit unterschiedlichen Schwerpunktsetzungen (Innere Konsistenz, Testhalbierung, Retestung sowie Längenkorrekturen) auf verschiedene interpretierbare Koeffizienten führt. Für kategoriale Konstrukte gibt es darüber hinaus keine vergleichbare einheitliche Definition der Messgenauigkeit, sondern lediglich die über alle Kategorien gemittelte Zuordnungssicherheit (Rost, 1996, S.37, 361; Fricke, 1972; Klauer, 1987). Da keine verbindliche Empfehlung für die Wahl eines Schätzers vorliegt, und zumeist auch nicht alle Varianten geschätzt werden, ist in der Regel keine Vergleichbarkeit gegeben.

ad 3 (Relevante Abhängigkeit)

Es gibt bislang keinen Konsens darüber, welche praxisrelevanten Faktoren den Kennwert beeinflussen und welcher Einfluss anzugeben ist. Rückert (1993) nennt zumindest einige für die Messgenauigkeit in Frage kommende Faktoren:

a) Objektivität:

- Einfluss des räumlichen Settings: Einzel- vs. Gruppentestung
- Standardisierung der Instruktion (PC vs. mündliche Instruktion)
- Standardisierung der Auswertung

b) Validität:

- Veränderliche Eigenschaft

c) Methodenaspekte:

- Testlänge
- Verteilung der Messwerte (z.B. Boden- und Deckeneffekte)

Testanwender wissen in der Regel um den Einfluss der genannten Faktoren. Es ist dennoch für den Testanwender nur schwer abschätzbar, wie massiv sich die Messgenauigkeit unter den situativen Bedingungen in der Praxis verändert. Prinzipiell gehen diese Faktoren auch bei der Normierung und Bestimmung des Reliabilitätskoeffizienten ein (wenngleich meist nur getrennte Schätzungen für verschiedene Personengruppe angegeben werden). Grundsätzlich besteht eine sinnvolle Abhängigkeit für die Mehrzahl der angeführten Faktoren. Die Anforderung wird als erfüllt gewertet.

ad 4 (Unabhängigkeit von irrelevanten Einflüssen)

Die Höhe des Reliabilitätskoeffizienten ist auch von Aspekten abhängig, die weniger praktisch relevante Einflüsse darstellen:

a) Varianz der Personenstichprobe (Stichprobenabhängigkeit)

b) Itemhomogenität

c) Längenkorrekturen

Diese Abhängigkeiten führen für den Praktiker zu nicht interpretierbaren Verzerrungen des Koeffizienten, womit die Anforderung nicht erfüllt ist.

ad 5 (Positive Zahlen)

Reliabilitätskoeffizienten können nur positive Werte annehmen, womit die Anforderung erfüllt ist.

ad 6 (Ganze Zahlen)

Diese Anforderung wird nicht erfüllt (hinzu kommt die ungewöhnliche Form der Darstellung ‚.xx').

ad 7 (Die Anzahl möglicher Ausprägungen)

Durch die Begrenzung auf zwei Nachkommastellen ist diese Bedingung erfüllt.

ad 8 (Intervallskalenniveau)

Um Reliabilitätskoeffizienten (geschätzt über Korrelationen) zu vergleichen, müssen sie Fischer-Z transformiert werden (vgl. Formel 7), womit die Bedingung nicht erfüllt ist.

Formel 7

$$Z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

ad 9 (Signifikante Einheiten)

Es ist bislang generell unüblich, den Standardfehler eines Testkennwertes anzugeben. Die Forderung wäre theoretisch umsetzbar, ist aber praktisch nicht erfüllt.

ad 10 (Relevanz)

Messgenauigkeit ist für den Testanwender ein relevantes Testmerkmal.

ad 11 (Unmittelbarer Bezug)

Der Reliabilitätskoeffizient weist nur einen mittelbaren und darüber hinaus nicht-linearen Bezug (vgl. Abbildung 2a) zu dem – für den Testanwender interessanteren - Standardmessfehler oder zur Anzahl unterscheidbarer Testergebnisse (AuT) auf. Der Anteil aufgeklärter Varianz ist für den Testanwender nur von nachgeordnetem Interesse, weshalb die Bedingung nicht erfüllt ist.

ad 12 (Maßeinheit)

Die Maßeinheit ist definiert als ‚Varianz der Messwerte'. Diese ist für den Testanwender nicht unmittelbar bedeutsam.

ad 13 (Angabe der notwendigen Genauigkeit)

Es ist aufgrund der nicht-linearen Zusammenhänge zu anderen Maßen der Messgenauigkeit schwierig, aus der Fragestellung heraus die notwendige Höhe des Reliabilitätskoeffizienten anzugeben (vgl. ad 13 zur Differenziertheit).

ad 14 (Erlernbarkeit)

Um den Reliabilitätskoeffizienten interpretieren zu können, muss in der Regel die Messfehlertheorie beherrscht werden. Eine leichte Erlernbarkeit scheint deshalb nicht gegeben.

Der Reliabilitätskoeffizient schneidet nach den hier vorgeschlagenen Kriterien mit 4 von 14 möglichen Punkten für die Benutzbarkeit insgesamt unbefriedigend ab. Für den Testanwender in der Praxis scheint der Reliabilitätskoeffizient damit letztlich für die Testauswahl nicht optimal benutzbar.

DIE DIFFERENZIERTHEIT ALS MASS DER MESSGENAUIGKEIT

Im Folgenden wird ein Kennwert vorgestellt, der mehrfach in verschiedenen Varianten und unabhängig voneinander vorgeschlagen wurde, sich aber bislang nicht als Standard durchsetzen konnte: Die Differenziertheit (Jäger, 1974, 1976; Lehrl & Kinzel, 1973; Rückert, 1987). Wright und Masters definieren den sehr ähnlichen Kennwert ‚Number of Person Strata' (1982, Formel 5.5.4), wobei die Autoren den Kennwert als ‚number of statistically distinct *person strata*' (S.106; Hervorhebung im Original) bezeichnen. Die Absicht von Wright und Masters bestand in der Bestimmung der *Anzahl unterscheidbarer Testergebnisse (AuT),* ohne jedoch auf die vorteilhafte testmodellübergreifende Konzeptualisierung der Messgenauigkeit hinzuweisen. Dieses Konzept scheint nach Ansicht des Autors in ein ‚anwendernahes' Maß zur Veranschaulichung der Messgenauigkeit zu führen. Die von Lehrl und Kinzel definierte Differenziertheit kann als eine (anders genormte) Schätzung der Anzahl unterscheidbarer Testergebnisse aufgefasst werden. Obwohl der Kennwert ursprünglich innerhalb metrischer Konstrukte eingeführt wurde, besteht doch prinzipiell die Möglichkeit, auch die Anzahl unterscheidbarer Testergebnisse im Falle von klassifikatorischen Konstrukten darzustellen (Anzahl unterscheidbarer Typen oder Störungen – wenngleich das Maß noch durch die Zuordnungssicherheit bzw. die Irrtumswahrscheinlichkeit korrigiert werden muss). Das Maß der Anzahl unterscheidbarer Testergebnisse wäre damit – neben den Begrenzungen aus der Skalierung (Anzahl der Klassen) - eine Funktion aus der Messgenauigkeit und der Messsicherheit (Irrtumswahrscheinlichkeit bzw. Zuordnungssicherheit).

Die Schätzung der Anzahl unterscheidbarer Testergebnisse für metrische Konstrukte kann über die Differenziertheit erfolgen. Die Differenziertheit ist nach Lehrl und Kinzel (1973) über die Teilung der Testskala (R) (vgl. Formel 9) durch die kritische Differenz (k) (vgl. Formel 8) definiert.

Formel 8

$$(x_2 - x_1)_{0.05} = 1{,}96 * s_x * \sqrt{2(1-r)}$$

Formel 9

$$D = \frac{R}{k}$$

Der Zusammenhang zwischen der Differenziertheit (der Range der Messwertverteilung wird auf 6 z-Einheiten begrenzt) und der Reliabilität ist graphisch in Abbildung 2c dargestellt. Die Schätzung der AuT über die Differenziertheit (innerhalb der Messfehlertheorie) ist damit nur eine Transformation des Reliabilitätskoeffizienten. Es besteht – um die Probleme des Reliabilitätskoeffizienten zu vermeiden - die Notwendigkeit sich innerhalb der Messfehlertheorie auf einen Reliabilitätsschätzer festzulegen, um eine eindeutige Vergleichsbasis herzustellen. Die Innere Konsistenz nach Cronbach stellt eine sehr häufig angegebene Schätzung der Messgenauigkeit dar. Hinzu kommt, dass dieser Schätzer auf nur einem Erhebungszeitpunkt beruht, wodurch Lern- und Reifeprozesse den Kennwert nicht beeinflussen. Es scheint aber unvermeidbar, dass mit der Festlegung auf einen bestimmten Schätzer andere Aspekte zugunsten der Vergleichbarkeit zurücktreten müssen.

BEWERTUNG DER BENUTZBARKEIT DER DIFFERENZIERTHEIT

ad 1 (Eindeutige Berechnungsgrundlage)

Die Anforderung ist erfüllt.

ad 2 (Vergleichbarkeit)

Die Operationalisierung der AuT über die Differenziertheit führt bislang zur Einschränkung auf metrische Konstrukte. Dies trifft jedoch auf die Mehrheit psychologischer Testverfahren zu, womit eine hohe (wenngleich noch verbesserungswürdige) Vergleichbarkeit gegeben ist.

ad 3 (Relevante Abhängigkeit)

Vergleichbar mit den Ausführungen zur Reliabilität, weshalb die Anforderung erfüllt ist.

ad 4 (Unabhängigkeit von irrelevanten Einflüssen)

Vergleichbar mit den Ausführungen zur Reliabilität, weshalb die Anforderung nicht erfüllt ist.

ad 5 (Positive Zahlen)

Die Anforderung ist erfüllt.

ad 6 (Ganze Zahlen)

Die Anforderung ist nicht erfüllt.

ad 7 (Die Anzahl möglicher Ausprägungen)

Aufgrund einer realistischen Begrenzung der Messgenauigkeit werden Differenziertheitswerte in der Regel in einem Intervall von 2 bis 10 liegen (vgl. Abb. 2c).

ad 8 (Intervallskalenniveau)

Die Skala ist mit der konstanten Größe der kritischen Differenz intervallskaliert.

ad 9 (Signifikante Einheiten)

Um diese Anforderung zu erfüllen, müsste die Stichprobengröße bestimmt werden, ab welcher der doppelte Standardfehler der Differenziertheit die kritische Differenz übertrifft. Dieser Standard wurde bislang nicht berechnet, weshalb die Bedingung derzeit nicht erfüllt ist.

ad 10 (Relevanz)

Diese Bedingung ist erfüllt.

ad 11 (Unmittelbarer Bezug)

Die Differenziertheit macht eine Aussage über die Anzahl unterscheidbarer Testergebnisse und bietet für den Testanwender ein gut handhabbares Maß für die Messgenauigkeit an, wodurch Transformationen nicht mehr notwendig sind.

ad 12 (Maßeinheit)

Die Differenziertheit verfügt über eine interpretierbare Maßeinheit: Die kritische Differenz.

ad 13 (Angabe der notwendigen Genauigkeit)

Ein Beispiel soll die ‚Verwendungsweise' des Kennwertes veranschaulichen: Ein Praktiker möchte lediglich eine sehr grobe Aussage innerhalb eines Screenings machen, weshalb für diese diagnostische Fragestellung A eine Differenziertheit von 2 genügt. In einer anderen Situation B möchte der Testanwender eine genauere Aussage machen und verlangt deshalb eine Differenziertheit von 5, während er für eine dritte Fragestellung C eine sehr genaue Aussage erreichen möchte, weshalb er eine Differenziertheit von 10 benötigt. Die unterschiedlichen Ansprüche lassen sich

graphisch in Abbildung 3 darstellen, wobei die Analogie zu Messgeräten (z.B. Lineal) für die Längenmessung beabsichtigt ist.

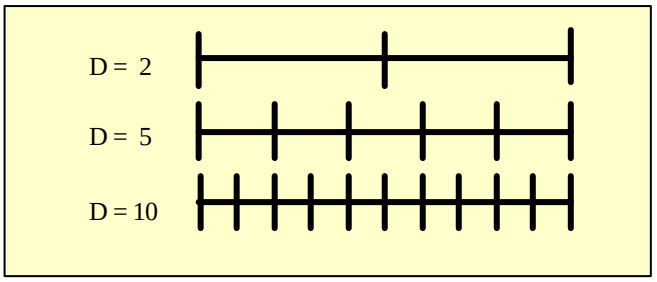


Abbildung 3. Graphische Veranschaulichung der Differenziertheit

Es erscheint ungleich schwieriger, die geforderte Messgenauigkeit über den Reliabilitätskoeffizienten anzugeben (die nummerischen Werte finden sich im Anhang A).

ad 14 (Erlernbarkeit)

Obwohl - wie im Falle des Reliabilitätskoeffizienten - die Messfehlertheorie zugrunde gelegt wird, benötigt ein Testanwender aufgrund des Konzeptes der Anzahl unterscheidbarer Testergebnisse nach Ansicht des Autors kein tiefgreifendes Verständnis probabilistischer Grundkonzepte, sondern lediglich grundlegende mathematische Kenntnisse. Die Anforderung wird als erfüllt gewertet.

Insgesamt schneidet die Differenziertheit mit 11 von 14 Punkten in der Benutzbarkeitsskala im Vergleich zum Reliabilitätskoeffizienten wesentlich zufriedenstellender ab.

BEWERTUNG

Die Bewertungen der eigenen Vorgehensweise beziehen sich erstens auf die Kriterien der Benutzbarkeit und zweitens darauf, ob aufgrund der dargestellten Analysen eine Präferenz eines Messgenauigkeitsmaßes erfolgen kann.

Das Ziel der Entwicklung der Anforderungen an Kennwerte lag in der Messung der Benutzbarkeit, wie sie in der Softwareergonomie erfolgte (ISO 9241/10, 1991; Willumeit, Gediga & Hamborg, 1995). Die Benutzbarkeitskriterien bezogen sich auf die Grundlagen, Skalierung und Interpretation von Testkennwerten. Die Anforderungen an die *Grundlagen* eines Kennwertes könnten noch weitere statistische Gesichtspunkte mit aufnehmen, wie einen möglichst geringen Standardfehler oder eine Robustheit gegenüber Abweichungen von der Normalverteilung (allgemein hierzu die Fisher-Kriterien der Parameterschätzung). Die Frage nach den ‚positiven' Eigenschaften einer *Skalierung* kann derzeit nur mit Hilfe von Plausibilitätsargumenten gestützt werden, weshalb weiterführende Untersuchungen in diesem Bereich notwendig erscheinen. Die Kriterien und Operationalisierung zur *Interpretation* besitzen ebenfalls nur einen vorläufigen Charakter (z.B. die Erlernbarkeit). Die Liste ist deshalb zunächst als Anregung zu verstehen. Entsprechend kann noch kein abschließendes Urteil über die Präferenz der Differenziertheit gegenüber dem Reliabilitätskoeffizienten erfolgen, wenngleich deutliche Schwächen des Reliabilitätskoeffizienten aufgedeckt wurden. Darüber hinaus existieren noch weitere Maße der Messgenauigkeit, deren Bewertung noch aussteht.

AUSBLICK

Die Kriterien der Benutzbarkeit verweisen auf weiteren Forschungsbedarf. Das Kriterium der ‚optimalen' Skalierung verlangt nach einer theoretischen Fundierung, beispielsweise über kognitive Ansätze. Eine Fragestellung wäre zum Beispiel, wie nummerische Werte mental repräsentiert werden. Erste Ansätze legen hierbei keine rationale Verarbeitung nahe (vgl. die Prominenzstruktur des Dezimalsystems; Albers & Albers, 1983; Henss, 1989; Schmale, 1996). Ebenso können Handlungsmodelle über das Verhalten eines Testanwenders bei der Testauswahl aufgestellt werden. Offen blieb auch die Frage nach der ‚besten' Auswahlstrategie. Bei 95 primären, situationsabhängigen und sekundären Aspekten (Standards für pädagogisches und psychologisches Testen; Häcker et al., 1998), die für die Testauswahl herangezogen werden können, müssen die Möglichkeiten der Praxis evtl. verstärkt berücksichtigt werden. Es fehlt zudem eine Messung der ‚Benutzbarkeit' für das *Gesamt* aller in der Testbeschreibung vorkommenden Informationen.

Inwieweit alternative Testkennwerte die Testauswahl tatsächlich verbessern können, müssten Evaluationsstudien zeigen. Diese könnte sich an folgenden Dimensionen (in Anlehnung an das Vorgehen in der Softwareergonomie) orientieren:

A) Die *Häufigkeit* der Heranziehung des Kennwertes bei der Testauswahl durch den Testanwender (diese ist bislang gering).

B) Die *Nützlichkeit*, die dem Kennwert innerhalb der Testauswahl relativ zu anderen zugeschrieben wird.

C) Die *Zufriedenheit* des Testanwenders mit dem Angebot an Kennwerten vor dem Hintergrund von praxisrelevanten Aspekten.

Die Anforderungen an die Grundlagen eines Kennwertes machen deutlich, dass die Konzeption von Testkennwerten diskutiert werden sollte, wenn eine breite Vergleichbarkeit gewährleistet werden soll. Testkennwerte sollten demnach konzeptionell derart definiert sein, dass die Operationalisierung in verschiedenen Messtheorien möglich und vergleichbar bleibt.

Die Interessen des Testanwenders sollten mehr als bisher bei der Entwicklung von Testkennwerten berücksichtigt werden. Die Umfragen von Steck (1997) und Schorr (1995) weisen zum Beispiel darauf hin, dass die Testanwender an der Ökonomie bzw. Messeffizienz interessiert sind. Erste Vorschläge stammen vom Autor, der auf die Ausnutzung des Messwertbereiches hinweist (Ausschöpfungsquotient; Müller, 2000). Ein weiterer Kennwert versucht eine Maßzahl für das Verhältnis von Messgenauigkeit und Messaufwand anzugeben (Messeffizienz).

LITERATUR

Albers, W. & Albers, G. (1983): On the prominence structure of the decimal system. In R.W. Scholz (Hrsg.), Decision making under uncertainty. Amsterdam: North-Holland.

Angleitner, A. (1997): Testrezension zu Minnesota Multiphasic Personality Inventory (MMPI). Zeitschrift für Differentielle und Diagnostische Psychologie, 18, 4-10.

Archer, R. P., Maruish, M., Imhof, E. A. & Piotrowski, C. (1991): Psychological test usage with adolescent clients: 1990 survey findings. Professional Psychology: Research and Practice, 22, 247-252.

Behr, E. (1996): Ergonomie im Produkt-Entstehungsgang – aus der Sicht des öffentlichen Auftraggebers. Psychologische Beiträge, 38, 100-108.

Bennet, J. (1984): Managing to meet usability requirements: establishing and meeting software development goals. In J. Bennet, D. Case, J. Scandelin & M. Smith (Eds.), Visual display terminals: usability issues and health concerns. Engelwood Cliffs, NJ: Prentice Hall.

Brickenkamp, R. (1997): Handbuch psychologischer und pädagogischer Tests. Göttingen: Hogrefe.

Bubb, H. & Bubb, P. (1996): Möglichkeiten und Grenzen der Umsetzung ergonomischer Erkenntnisse. Psychologische Beiträge, 38, 140-163.

Czida, W., Herda, S. & Itzfeld, W. D. (1978): Factors of user-perceived quality of interactive systems. Bericht Nr. 40. Bonn: Gesellschaft für Mathematik und Datenverarbeitung, Institut für Software-Technologie.

Dutke, S. (1994): Mentale Modelle: Konstrukte des Wissens und Verstehens. Göttingen: Verlag für Angewandte Psychologie.

Eason, K. D. (1984): Towards the experimental study of usability. Behavior and Information Technology, 3, 133-143.

Engels, R. (1997): Replik zur Rezension des MMPI. Zeitschrift für Differentielle und Diagnostische Psychologie, 18, 10-15.

Frauenhoffer, D., Ross, M. J., Gfeller, J., Searight, H. R. & Piotrowski, C. (1998): Psychological test usage among licensed mental health practitioners: A multidisciplinary survey. Journal of psychological practice, 4, 28-33.

Fricke, R. (1972): Testgütekriterien bei lehrzielorientierten Tests. Zeitschrift für erziehungswissenschaftliche Forschung, 6, 150-175.

Galer, I. A. (1987): Applied ergonomics handbook. London: Butterworths.

Glöckner-Rist, A. & Schmidt, P. (Hrsg.) (1999): ZUMA-Informationssystem. Elektronisches Handbuch sozialwissenschaftlicher Erhebungsinstrumente, Version 4.00. Mannheim: Zentrum für Umfragen, Meinungen und Analysen.

Grubitzsch, S. & Rexilius, G. (1978): Testtheorie - Testpraxis. Hamburg: Rowohlt.

Gulliksen, H. (1950; repr. 1987): Theory of mental tests. Lawrence Erlbaum Associates, Inc.

Häcker, H., Leutner, D. & Amelang, M. (Hrsg) (1998): Standards für pädagogisches und psychologisches Testen. Diagnostica, Suppl.1.

Henss, R. (1989): Zur Vergleichbarkeit von Ratingskalen mit unterschiedlicher Kategorienzahl. Psychologische Beiträge, 31, 264-284.

ISO 9241 Part 10. Ergonomic requirements for office work with visual display terminals (VDTs), Part 10: Dialog Principles, First Committee Draft, September, 1991.

Jäger, R. (1974): Bemerkungen zu: Die Standardskala der kritischen Differenz. Diagnostica, 20,165-168.

Jäger, R. (1976): f_G – ein statistischer Test zur Bestimmung der Differenzierungsfähigkeit psychologischer Skalen. Psychologische Beiträge, 18, 214-223.

Junga, M. (1979): Oh, diese Testhefte! Westermanns Pädagogische Beiträge, 31, 184-185.

Kelley, T. L. (1921): The reliability of test scores. Journal of Educational Research, 3,370-379.

Kelley, T. L. (1942): The reliability coefficient. Psychometrika, 7, 75-83.

Klauer, K. J. (1987): Kriteriumsorientierte Tests. Göttingen: Hogrefe.

Kluwe, R. H. (2001): Zur Lage der Psychologie: Perspektiven der Fortentwicklung einer erfolgreichen Wissenschaft. Psychologische Rundschau, 52, 1-10.

Kubinger, K. (1997): Editorial zum Themenheft „Testrezensionen: 25 einschlägige Verfahren". Zeitschrift für Differentielle und Diagnostische Psychologie, 18, 1-3.

Kuder, G. F. & Richardson, M. W. (1937): The theory of the estimation of test reliability. Psychometrika, 2, 151-160.

Lehfeld, H. & Erzigkeit, H. (1993) In H. J. Möller und A. Rohde (Hrsg.). Psychische Krankheit im Altern. Berlin: Springer.

Lehrl, S. & Kinzel, W. (1973): Die Standardskala der kritischen Differenz. Diagnostica, 19, 75-88.

Mayhew, D. J. (1999): The usability engineering lifecycle. San Francisco: Morgan Kaufmann Publishers.

Müller, J. M. (2000): Neue Leistungs- und Effizienzkennwerte für psychologische Testverfahren: Breite, Differenziertheit, Personenunterscheidungsvermögen, Effizienz und Ausschöpfungsquotient. Poster auf dem Kongress der DGPS in Jena.

Müller-Böhling, D. (1991): Anforderungen an Tests zur Messung der Arbeitszufriedenheit für die Anwendung in der betrieblichen Praxis. In L. Fischer. Arbeitszufriedenheit. Stuttgart: Verlag für Angewandte Psychologie.

Murrell, K. F. H. (repr. 1971): Ergonomics – Man in his Working Environment. London: Chapman and Hall.

Nielsen, J. & Mack, R. L. (1994): Usability inspection methods. New York: Wiley.

Piotrowski, C. & Keller, J. W. (1992): Psychological testing in applied settings: A literature review from 1982-1992. Journal of Training & Practice in Professional Psychology, 6, 74-82.

Piotrowski, C., Belter, R. W. & Keller, J. W. (1998): The Impact of „Managed Care" on the Practice of Psychological Testing: Preliminary Findings. Journal of Personality Assessment, 70, 441-447.

Polson, P. G. & Lewis, C. H. (1990): Theory-based design for easily learned interfaces. Human-Computer Interaction, 5, 191-220.

Rost, J. (1996): Lehrbuch der Testtheorie, Testkonstruktion. Bern: Huber.

Rost, J. (2000): Haben ordinale Raschmodelle variierende Trennschärfen? Eine Antwort auf die Wiener Repliken. Psychologische Rundschau, 51, 36-37.

Rückert, J. (1993): Psychometrische Grundlagen der Diagnostik. Göttingen: Hogrefe.

Schmale, H. (1996): Psychologische Aspekte der Ergonomie. Psychologische Beiträge, 38, 52-57.

Schinka, J. A. & Borum, R. (1994): Readability of Normal Personality Inventories. Journal of Personality Assessment, 62, 95-101.

Schmidtke, H. (Hrsg.) (1993): Ergonomie. München: Hanser.

Schorr, A. (1995): Stand und Perspektiven diagnostischer Verfahren in der Praxis. Ergebnisse einer repräsentativen Befragung westdeutscher Psychologen. Diagnostica, 41, 3-20.

Schulz, C., Schuler, H. & Stehle, W. (1985): Die Verwendung eignungsdiagnostischer Methoden in deutschen Unternehmen. In H. Schuler & W. Stehle (Hrsg.):

Organisationspsychologie und Unternehmenspraxis. Perspektiven der Kooperation. Stuttgart: Verlag für Angewandte Psychologie.

Spada, H. (1997): Lage und Entwicklung der Psychologie in Deutschland, Österreich und der Schweiz. Psychologische Rundschau http://www.hogrefe.de/PsychologischeRundschau/artikel1.html am 2.3.12001.

Steck, P. (1991): Bemerkungen zu L. Tents Beitrag „Psychodiagnostische Verfahren und die minima scientifica". Diagnostica, 37, 89-92.

Steck, P. (1997): Aus der Arbeit des Testkuratoriums. Psychologische Testverfahren in der Praxis. Diagnostica, 43, 267-284.

Steyer, R. & Eid, M. (1993): Messen und Testen. Berlin: Springer.

Stoll, R. (1978): Testgebrauch in der schweizerischen Beratungspraxis. In U. Pulver, A. Sang & F. W. Schmid (Hrsg.), Ist Psychodiagnostik verantwortbar? Wissenschaftler und Praktiker diskutieren Anspruch, Möglichkeiten und Grenzen psychologischer Erfassungsmittel (S. 321-339). Bern: Huber.

Tent, L. (1991): Psychodiagnostische Verfahren und die minima scientifica. Diagnostica, 37, 83-88.

Testzentrale (2000): Testkatalog. Göttingen: Hogrefe.

Testkuratorium der Föderation für die Testbeurteilung (1986): Beschreibung der einzelnen Kriterien für die Testbeurteilung. Diagnostica, 32, 358-360.

Wade, T. C. & Baker, T. B. (1977): Opinions and use of psychological test. American Psychologist, 32, 874-882.

Wandmacher, J. (1993): Software Ergonomie. Berlin: de Gruyter.

Willumeit, H., Gediga, G. & Hamborg, K.-C. (1995): Validation of the ISOMetrics usability inventory. Forschungsberichte der Universität Osnabrück, Nr. 105.

World Health Organisation (1991): Internationale Klassifikation psychischer Störungen, ICD-10, Kapitel V (F). Klinisch-diagnostische Leitlinien. Bern: Huber.

Wright, B. & Masters, G. (1982): Rating Scale Analysis. Rasch Measurement. Chicago: MESA Press.

Anhang A

Die verlangte Messgenauigkeit würden folgenden Reliabilitäten entsprechen: Test A, r=.50, Test B, r=.92 Test C, r=.98.

Das Personenunterscheidungsvermögen eines diagnostischen Tests unter Berücksichtigung der Messwertverteilung

ABSTRACT

A new criteria is developed to display the performance of a test in seperating testscores. This coefficient is named Personenunterscheidungsvermögen (PUV) and helps the testuser to decide whether a test is an approbiate choice for a specific diagnostic question. The PUV-coefficient does not only depend on the reliability of a test, because the shape of the testscoredistribution, who differs sometimes from the expected normaldistribution, shows an effect. Ergonomy guidelines are taken into consideration while developing the coefficient to strengthen the practicability in understanding and using the coefficient. The PUV-coefficient supplement approaches to improve the quality standard in applying psychometric assessment. Monte-Carlo-Studies offer data from different types of nonnormal distributed testscores what allows to evaluate the effect on the PUV. The result reveals, that a deviation from the normaldistribution does not automatically reduce the PUV, but ceiling-floor-effects shows a decrease in the PUV-performance of a test. SAS-Macros are attached and offer the possibility to evaluate the PUV for a test.

Keywords: ergonomy, testscore distribution, quality control, psychometric.

ZUSAMMENFASSUNG

Dieser Beitrag stellt einen neuen Testkennwert vor, der dem Testanwender in der Praxis bezüglich einer konkreten diagnostischen Fragestellung (wie groß ist die Wahrscheinlichkeit, mit Hilfe der Testergebnisse zwei Personen voneinander zu unterscheiden?) eine direkt interpretierbare Aussage macht. Das ‚Personenunterscheidungsvermögen' (PUV) zeigt sich dabei von der Reliabilität und von der Messwertverteilung abhängig, ein Sachverhalt der bislang – trotz seiner praktischen Bedeutung – nicht durch einen Testkennwert abgebildet wurde. Bei der Entwicklung des Kennwertes wurden ergonomische Leitlinien berücksichtigt, um eine höhere Anwenderfreundlichkeit zu gewährleisten. Diese Leitlinien zur leichteren Handhabung eines Kennwertes resultierten aus einer kritischen Betrachtung des bekannten – jedoch in der Praxis nicht hinreichend berücksichtigten – Reliabilitätskoeffizienten. Der PUV-Koeffizient stellt damit einen Beitrag zur Qualitätssicherung innerhalb der psychologischen Diagnostik dar. Monte-Carlo-Studien simulieren verschiedene nicht-normalverteilte Messwertverteilungen (Gleich-, U- und J-Verteilungen), deren PUV-Wert anschließend evaluiert wird. Die Ergebnisse zeigen, dass eine Abweichung von der Normalverteilung nicht generell ungünstig für den Nachweis von Personenunterschieden ist. Das verwendete SAS-Makro zur Berechnung des Kennwertes ist angefügt und ermöglicht einem Testentwickler, das Personenunterscheidungsvermögen selbstständig zu bestimmen.

Schlagwörter: Testpraxis, Qualitätssicherung, Gütekriterien, Ergonomie, Messwertverteilung, Simulationsstudie.

TESTBESCHREIBUNG IM SPANNUNGSFELD VON WISSENSCHAFT UND PRAXIS[1]

Ein Testanwender orientiert sich seit nun mehr fast achtzig Jahren (nimmt man Kellys Arbeit in den 20er Jahren zur Reliabilität als Beginn) bei der Wahl eines psychologischen Testverfahrens an den Gütekriterien (Reliabilität, Validität, Objektivität; Testkuratorium, 1986), da diese den Bedarf nach wissenschaftlich fundierten Kennwerten erfüllen. Die Forderung einer Qualitätssicherung (QS; vgl. DIN ISO9000/1/4-1/2; vgl. die Diskussion innerhalb der Fachgruppe Differentielle Psychologie und Persönlichkeitspsychologie) in der psychologischen Diagnostik stellt inzwischen neue Herausforderungen an die Testbeschreibung, was bedeutet, dass nicht nur eine *wissenschaftlich* fundierte Beschreibung der Messqualität verlangt wird, sondern auch *praxisrelevante* Anforderungen an eine Testbeschreibung gestellt werden. Der Handlungsbedarf, die Gütekriterien für die Testpraxis zu modifizieren oder zu ergänzen, drängt sich ebenfalls aus dem Verhalten der Testanwender auf (Schorr, 1995 und Steck, 1997 im deutschen Sprachraum; Piotrowski, Belter & Keller, 1998; Archer, Maruish, Imhof & Piotrowski, 1991; Frauenhoffer, Ross, Gfeller Searright & Piotrowski, 1998; Piotrowski & Keller, 1992; Wade & Baker, 1977 im anglo-amerikanischen Sprachraum), da sich zeigt hat, dass die verfügbaren Gütekriterien nur mangelhaft in die Testauswahl mit einfließen (Kubinger, 1997).

[1] Dieser Beitrag bezieht sich auf eine Posterpräsentation auf dem 42. Kongress der DGPS in Jena.

MOTIVATION ZUR ENTWICKLUNG NEUER TESTKENNWERTE

Die Standards für psychologisches und pädagogisches Testen (Häcker, Leutner & Amelang, 1998) wie auch die Bemühungen um eine DIN-Norm (Hornke, 2000) versuchen als Reaktion auf diesen Misstand, mit den vorhandenen Kennwerten einen normativen Umgang einzuführen. Diese Aktivitäten schließen jedoch nicht aus, dass neben den Standards und Normen nicht auch neue, verbraucherfreundliche Kennwerte entwickelt werden sollten, um eine mögliche Ursache für die Vernachlässigung der klassischen Gütekriterien zu überwinden. Eine Analyse des Reliabilitätskoeffizienten hat gezeigt, dass wichtige Testinformationen für die Testpraxis ungeeignet kommuniziert werden (Müller, 2000b, 2000c). Als Kritikpunkte wurden zum Beispiel die nicht lineare Beziehung zum Messfehler herausgestellt, die nicht intervallskalierte Metrik des Koeffizienten sowie die Notwendigkeit der Anwendung von Formeln, um die benötigte Information zu erhalten (Müller, 2000b). Aus dieser Problemanalyse heraus werden im Weiteren allgemeine ergonomische Leitlinien definiert, die eine höhere Anwenderfreundlichkeit für neue Testkennwerte gewährleisten sollen. Der Bedarf nach neuen Kennwerten ergibt sich auch dann, wenn praxisrelevante Informationen über einen Test (wie die Messwertverteilung) nicht einbezogen wurden, obwohl sie einen Effekt (siehe unten) auf die Einsatzmöglichkeiten eines Tests haben. Die Forderung nach neuen, anwenderfreundlichen Kennwerten ergibt sich damit aus theoretischen wie praktischen Gründen. Im Weiteren soll die Fragestellung im Vordergrund stehen, Personenunterschiede über dessen Testwerte nachzuweisen. Damit soll der neue Kennwert für die folgende diagnostische Fragestellung eine

unmittelbare Antwort geben: *‚Wie groß ist die Wahrscheinlichkeit, mit Hilfe eines Tests zwischen zwei zufällig ausgewählten Testpersonen einen signifikanten Unterschied festzustellen?‘*

ERGONOMISCHE LEITLINIEN

Die Antwort auf diese Frage über einen Kennwert muss ergonomischen Leitlinien folgen. Die Forderung nach einer ‚Ergonomie‘ (Murrell, 1969; Schmidtke, 1993) ergibt sich prinzipiell immer dann, wenn eine Mensch-Technik-Schnittstelle vorliegt; und eine Testanwender-Test-Situation stellt eine solche (vgl. Müller, 2000b) dar. Die vom Autor aufgestellten ergonomischen Leitlinien lauten wie folgt:

I. *Vollständige Information*: Ein Testkennwert muss alle zur Beantwortung einer diagnostischen Fragestellung praktisch bedeutsamen Faktoren berücksichtigen.

II. *Praxisbezug*: Ein Testkennwert muss bezüglich der diagnostischen Fragestellungen – ohne weitere Umrechnungen – direkt aussagekräftig sein.

III. *Universalität*: Ein Kennwert soll unabhängig von spezifischen Testeigenschaften und Konstruktionswegen (insbesondere des testtheoretischen Modells) einheitlich und vergleichbar (= standardisiert) berechnet werden können.

IV. *Skalierung des Kennwertes*: Der Kennwert sollte auf aussagekräftigen Maßeinheiten basieren und intervallskaliert sein.

V. *Voraussetzungen beim Testanwender*: Die Interpretation des Kennwertes sollte möglichst wenig Hintergrundwissen beim Testanwender erfordern.

BERÜCKSICHTIGUNG DER VERTEILUNGSFORM VON MESSWERTEN

Der erste Punkt der Leitlinien (vollständige Information) führt auf die Berücksichtigung der Verteilungsform der Messwerte. Der Einfluss verschiedener Messwertverteilungen soll an zwei Tests veranschaulicht werden. Beide Tests sollen bezüglich aller sonstigen Bedingungen vergleichbar sein, also bzgl. ihrer Reliabilität (beide sind perfekt reliabel), der Menge an Testausgängen (bei beiden gleich drei: Ausgang A, B und C) sowie der Anzahl getesteter Personen (bei beiden gleich sechs); sie unterscheiden sich allein in der Verteilung ihrer Messwerte:

- Verteilungsform I (zwei Personen in A, zwei in B und zwei in C; vgl. Abb. 1) und
- Verteilungsform II (eine Person in A, vier in B und eine in C; vgl. Abb. 1).

Verteilung I **Verteilung II**

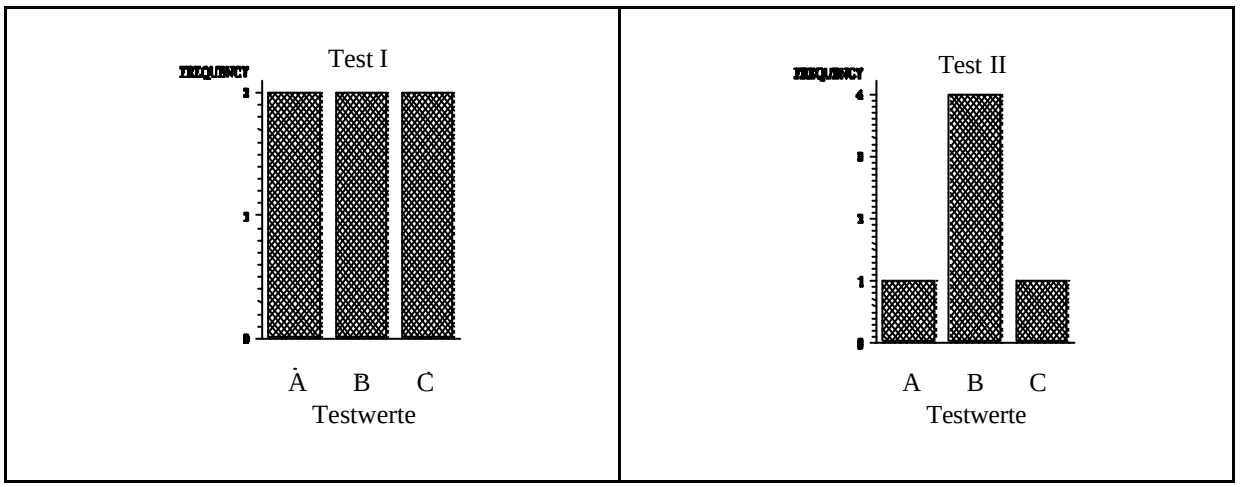


Abbildung 1. Unterschiedliche Verteilungsformen zweier Tests.

Es wird nun innerhalb jeder Verteilung ein *vollständiger Paarvergleich* aller Personen durchgeführt (insgesamt 15 Paarvergleiche; vgl. Formel 2) und entschieden, ob der

Unterschied zwischen beiden Personen signifikant (Messwert i ≠ Messwert j) ausfällt oder nicht. Zählt man die Anzahl der signifikanten Paarvergleiche zusammen, dann ergeben sich für die Verteilungsform I insgesamt 12 und für Verteilungsform II nur 9 Unterscheidungen. Die beiden Tests differieren offensichtlich aufgrund ihrer unterschiedlichen Messwertverteilung in ihrem Personenunterscheidungsvermögen.

Die Berücksichtigung der Messwertverteilung macht insbesondere dann Sinn, wenn die Messwertverteilung von Tests tatsächlich von der Normalverteilung abweicht. Diese Abweichungen sind in der Praxis tatsächlich zu beobachten, zum Beispiel im Falle des CFT von Weiß und Osterland (1979; vgl. Abb. 2; ebenso die deutsche Version des Short Form-36 Health Survey für den Einsatz in der Rehabilitation; vgl. Zwingmann, Metzger & Jäckel; 1998).

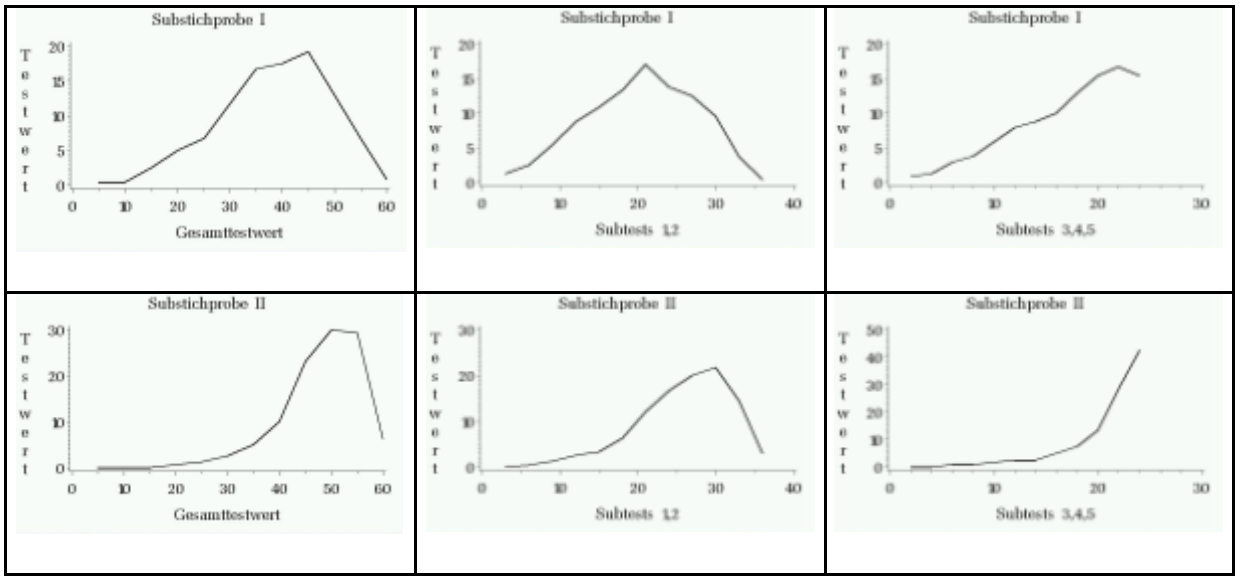


Abbildung 2. Verteilung der Rohwerte des CFT von Weiß und Osterland (1979).

Derartige Verteilungsanomalien können nach Lienert und Raatz (1994) drei Ursachen haben: Heterogene Analysestichproben (vgl. die Stichproben I und II in Abb. 2),

mangelhafte Testkonstruktion und drittens eine Verteilungsanomalie im Persönlichkeitsmerkmal (vgl. die Subtests 1,2 und 3,4,5 innerhalb der Substichproben I und II). Für einen Testanwender ist eine solche Ursachenanalyse jedoch hinfällig, da sie aus seiner Sicht nicht mehr zu korrigieren ist. Vielmehr interessiert den Testanwender die Einschränkung seiner diagnostischen Möglichkeiten. Dieser Effekt auf das Personenunterscheidungsvermögen soll über den Kennwert quantifiziert werden.

ALLGEMEINE BESTIMMUNG EINES PUV-WERTES

Um auf einen allgemein anwendbaren Kennwert zu gelangen, kann man - einen Algorithmus zur Unterscheidung zweier Testergebnisse vorausgesetzt (vgl. Formel 3 mit *k* der kritischen Differenz) - entsprechend Formel 1 die durchschnittliche Unterscheidungswahrscheinlichkeit bestimmen. Das *tU* steht hierbei für die Anzahl aller Paarvergleiche, während das *sU* für die Anzahl signifikanter Personenunterscheidungen steht. Der PUV-Koeffizient gibt demnach die durchschnittlich zu erwartende Unterscheidungsfähigkeit des Tests an.

Formel 1

$$PUV = \frac{sU}{tU} * 100\%$$

wobei

Formel 2

$$tU = \frac{n * (n-1)}{2}$$

und

Formel 3

$$sU = \sum_{i,j}^{n} s_{i,j} \begin{cases} s_{i,j} = 1, wenn\ \ x_i - x_j \leq k \\ s_{i,j} = 0, wenn\ \ x_i - x_j \rhd k \end{cases}$$

ist.

Im konkreten Fall muss der Abstand k unter der Festlegung einer Irrtumswahrscheinlichkeit testtheoretisch begründet werden. Die Bestimmung von k bringt grundlegende Probleme der Testdiagnostik zum Vorschein, weshalb sich ein eigener Abschnitt damit beschäftigen wird.

Zuvor soll die Irrtumswahrscheinlichkeit auf 5% festgelegt werden, damit der PUV-Koeffizient bezüglich dieses Faktors vergleichbar bleibt. Diese Festlegung dient allein der Standardisierung, und es bleibt einem Testanwender unbenommen, für eine spezifische Fragestellung eine abweichende Irrtumswahrscheinlichkeit festzulegen.

ENTSCHEIDUNGSALGORITHMEN IN DEN JEWEILIGEN TESTTHEORIEN

Die Entscheidungsalgorithmen gehen in den verschiedenen Testtheorien von unterschiedlichen Annahmen aus: In der Klassischen Testtheorie (KTT) folgt aus dessen Axiomen eine globale – über die Probanden hinweg vergleichbare – Reliabilität und damit ein einheitlicher Messfehler für alle Personen. Für die Schätzung des Messfehlers innerhalb der Item-Response-Theorien (IRT) werden hingegen der Personenparameter q_v und die Anzahl k der Items sowie deren

Informationsgehalt mit einbezogen. Der Standardfehler wird somit als Funktion des Personen- und des Itemparameters innerhalb des Raschmodells nach Formel 4 (nach Rost, 1996, S. 322) berechnet.

Formel 4

$$Var(E_q) = \frac{1}{\sum_{i=1}^{k} p_{vi}(1 - p_{vi})}$$

Die einerseits präzise, aber andererseits umständliche Lösung führte in der Praxis dazu, dass selbst für raschskalierte Tests (z. B. bei Kubinger & Wurst, 1988) auf die einfachere Formel der KTT zurückgegriffen wird (siehe Formel 5). Die kritische Differenz ist demnach eine nach theoretischen Ansprüchen nur suboptimale Lösung, zumal die Schätzung der kritischen Differenz durch die Abweichung einer Messwertverteilung von der Normalverteilung verzerrt ist.

Formel 5

$$(x_2 - x_1)_{0.05} = 1{,}96 * s_x * (2(1 - r_{tt}))^{1/2}$$

Ein verteilungsfreies Verfahren (oder eine nicht-lineare Flächentransformation) würde zwar den statistischen Ansprüchen genügen, jedoch gerade die in der Praxis Einfluss nehmende Verteilungsanomalie wird hierdurch in ihrem Effekt ausgeblendet. Gerade dieser Effekt soll jedoch durch den Kennwert sichtbar werden. Die Berechnung der kritischen Differenz in IRT ist ebenfalls nicht unproblematisch, da selbst raschskalierte Tests neueren und strengeren Modellprüfungen nicht genügen, was Ponocny und Ponocny-Seliger (2000) anhand von T-Rasch bei der Überprüfung der lokalen stochastischen Unabhängigkeit zeigen konnte. Das Für und

Wider von Modellannahmen ist deshalb nicht nur theoretisch zu diskutieren (denn je mehr Parameter, desto besser der Fit), sondern die praktischen Konsequenzen müssen mit berücksichtigt werden – im Sinne einer hinreichenden Genauigkeit bei vertretbarem Aufwand. Die Messwerte werden deshalb vor dem Hintergrund der Messfehlertheorie interpretiert und die kritische Differenz nach Formel 5 angewendet.

SIMULATIONSSTUDIE

Nachdem geklärt wurde, wie das PUV an konkreten Daten ermittelt wird, können nun folgende Fragen durch eine Simulationsstudie beantworten werden:

1) Welchen Einfluss hat die Form der Messwertverteilung auf das PUV?
2) Welcher Zusammenhang ergibt sich zwischen der Reliabilität und dem PUV je nach Verteilungsform?
3) Wie genau ist die Schätzung des PUV in Abhängigkeit von der Stichprobengröße?

Die letzte Frage bezieht sich auf die statistischen Eigenschaften wie die Erwartungstreue und der Schätzfehler des PUV, um neben den theoretischen Aspekten auch die praktische Interpretation eines PUV-Wertes zu erleichtern.

Um die Fragen beantworten zu können, variiert der Versuchsplan für die Simulationsstudie drei Faktoren: Die Verteilungsform der Messwerte, die Reliabilität und die Stichprobengröße, wobei für jede Kombination der Bedingungen der PUV-

Wert berechnet wird. Auf jeden dieser drei Faktoren wird im Folgenden im Detail eingegangen.

Die Generierung von Zufallsdaten erfolgt aus verschiedenen typischen Abweichungen (vgl. Lienert & Raatz, 1994) von: der Normalverteilung (A), die Gleichverteilung (B), die U-Verteilung (C), die leicht (D), die mittlere (E) und die stark linksschiefe Verteilung (F; diese entspricht einer J-Verteilung nach Lienert und Raatz und wird auch als Boden-Decken-Effekt interpretiert). Die Deskription der Schiefen und Exzesse aller Verteilungen findet sich in Tabelle 1 (die Werte wurden auf der Basis einer einmaligen, zufälligen Ziehung von 5000 Einzelwerten gebildet). Die graphische Darstellung der Verteilungsformen befindet sich in Abbildung 3.

Tabelle 1

Schiefe und Exzess der sechs Verteilungen

Verteilungsform	Schiefe	Exzess
U-Verteilung	0.00	-1.88
Gleichverteilung	0.03	-1.18
Normalverteilung	0.02	-0.04
Leicht schiefe Verteilung	0.60	0.50
Mittel schiefe Verteilung	2.11	11.42
Stark schiefe Verteilung	3.63	24.24

Zur Generierung der Zufallsdaten wurden SAS-Prozeduren verwendet (rannor; ranuni; ranpoi) und mit Hilfe des SAS-Makro %PVU (Anhang A) der PUV-Koeffizient berechnet. Die zunehmende Schiefe in der Verteilung wurde durch eine Quadrierung der Werte aus einer Normalverteilung erreicht (‚Leicht schiefe Verteilung' mit x^2, ‚Mittel schiefe Verteilung' mit x^4 und die ‚Stark schiefe Verteilung'

mit x^6). Die U-Verteilung wurde durch eine gespiegelte Poission-Verteilung angenähert (Programm in Anhang B).

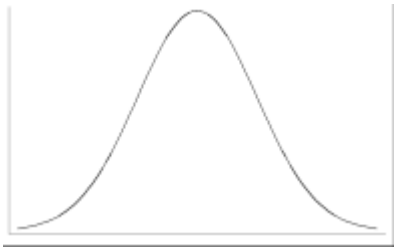
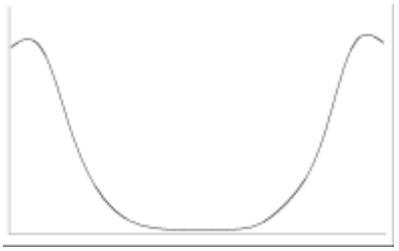
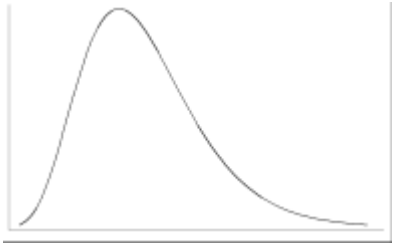
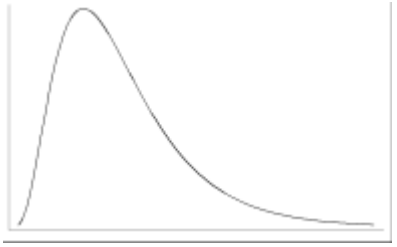
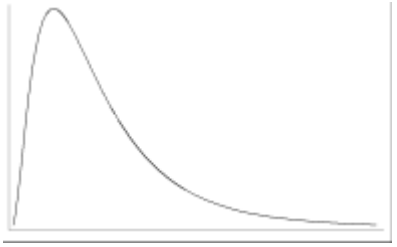
Normalverteilung	Gleichverteilung	U-Verteilung
		
A	B	C
Leicht schiefe Verteilung	Mittel schiefe Verteilung	Stark schiefe Verteilung
		
D	E	F

Abbildung 3. Normalverteilung, Gleichverteilung, U-Verteilung und leicht, mittel, stark-rechtsschiefe Verteilungen.

Neben der Verteilungsform ist auch der Einfluss der *Probandenanzahl* (= Stichprobengröße) zu prüfen. Eine systematische Variation des Stichprobenumfangs sollte Hinweise erbringen, inwieweit der Kennwert auf einen festen, unverzerrten Wert konvergiert (Erwartungstreue; Bronstein & Semendjajew, 1997, S. 682). Zur Identifikation eines Bias werden wiederholt (jeweils 30-mal) unabhängige Stichproben aus den sechs theoretischen Verteilungen gezogen, wobei der Umfang

der Stichproben schrittweise erhöht wird (N = 10, 25, 50, 100, 250, 500, 1000, 2500, 5000). Mit Hilfe dieser Analyse wird eine Schätzung der Genauigkeit eines PUV-Wertes über Konfidenzintervalle möglich.

Da die *Reliabilität als Moderatorvariable* den PUV-Wert je nach Verteilung verschieden verändern kann, erfolgen alle Berechnungen mehrfach mit abgestuften Reliabilitäten (die Stichproben wurden hierzu jeweils erneut generiert). Insgesamt wurden 13 Reliabilitätsstufen gebildet: r = 0.1, .10, .20, .30, .40, .50, .60, .70, .80, .85, .90, .95, .99.

ERGEBNISSE DER SIMULATIONSSTUDIE

Der Einfluss der Verteilungsformen auf das PUV ist in Tabelle 2 dargestellt. Hierzu wurden die einzelnen PUV-Werte zunächst über die unterschiedlichen Reliabilitätsbedingungen aggregiert.

Tabelle 2

Durchschnittliche PUV-Wert* der sechs Verteilungsformen

Verteilungsform	PUV-Wert
Gleichverteilung	31.92 %
U-Verteilung	31.48 %
Normalverteilung	30.44 %
Leicht schiefe Verteilung	29.96 %
Mittel schiefe Verteilung	26.33 %
Extrem schiefe Verteilung	21.29 %

* Die PUV-Werte wurden über 390 Einzelwerte gebildet (13 Reliabilitätsabstufungen mal 30 Wiederholungen bei einem Stichprobengröße von N = 5000).

Die Werte in Tabelle 2 zeigen, dass die Gleichverteilung noch vor der U- und der Normalverteilung am besten abschneidet. Die zunehmend schieferen Verteilungen zeigen niedrigere PUV-Werte. Die durchgehend recht niedrig erscheinenden Mittelwerte in Tabelle 2 sind auf die zusammenfassende Bewertung von Reliabilitätsstufen mit geringer Ausprägung zurückzuführen. Eine differenziertere Aufschlüsselung der Zusammenhänge von Reliabilität und PUV je Verteilungsform (vgl. 2. Fragestellung) ist in Abbildung 4 dargestellt (Ergebnisse nur für Stichprobengrößen mit N= 5000).

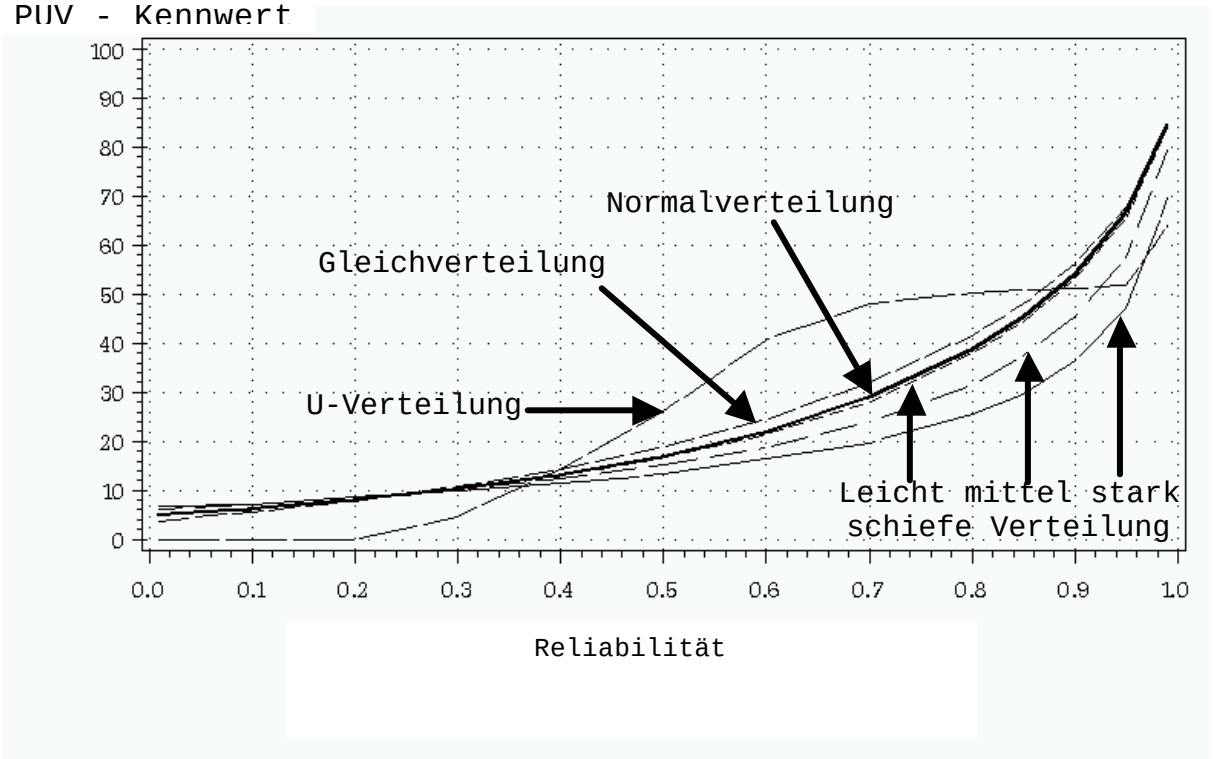


Abbildung 4. Einfluss der Verteilungsform und der Reliabilität auf das Personenunterscheidungsvermögen (PUV).

Der Verlauf für die Normalverteilung zeigt die für die Reliabilität typische, nicht-lineare Beziehung zum Messfehler und folgerichtig auch zum Effekt der Separierbarkeit von Testpersonen (vgl. Müller, 2000b). Erst für relativ hohe

Reliabilitäten steigen die Kurven stark an (mit Ausnahme der U-Verteilung). Abbildung 4 vermittelt den Eindruck, dass für psychologische Testverfahren kaum mehr als ein 80%iges Unterscheidungsvermögen zu erwarten ist – selbst bei hervorragenden Reliabilitäten von über r = .99. Je stärker die Verteilung von der Normalverteilung im Sinne einer größer werdenden Schiefe abweicht, desto stärker sinkt die Wahrscheinlichkeit, Personen als unterschiedlich zu erkennen, während umgekehrt bei Verteilungen mit geringerem Exzess eine höhere Wahrscheinlichkeit zu beobachten ist.

NUTZUNG DER ERGEBNISSE FÜR DIE PRAXIS

Um den Informationsgewinn von Abbildung 4 auszuschöpfen, sollen nur vertikale und horizontale Vergleiche den Einfluss der Verteilungsform auf das PUV herausstellen. Ein vertikaler Vergleich (Reliabilität= .80) der Normalverteilung mit einer stark schiefen Verteilung zeigt, dass bei einem sogenannten Deckeneffekt die Wahrscheinlichkeit um ca. 15% absinkt (im Vergleich zu U-Verteilung sogar um ca. 25%) und damit eine Effektgröße erzielt, die durchaus praktische Bedeutsamkeit erreicht. Horizontal (40% PUV) betrachtet erreicht ein normalverteilter Test dieses Unterscheidungsvermögen mit r = .80. Eine U-Verteilung erreicht dieses PUV schon mit einem r von .60.

Abbildung 4 weist somit einen deutlichen und praxisrelevanten Effekt der Verteilungsform auf das PUV nach. Die U-Verteilung zeigt – im Vergleich zu den anderen - einen überraschenden Verlauf, was dadurch erklärt werden kann, dass

durch die Häufung der Personen an beiden Enden der Skala für einen Großteil der Personenvergleiche eine fast maximale Distanz vorliegt. Es genügt deshalb eine relativ niedrige Reliabilität, um eine Unterscheidung zweier Probanden anhand ihrer Testwerte wahrscheinlich zu machen.

STATISTISCHE EIGENSCHAFTEN DES PUV-KOEFFIZIENTEN

Die *Erwartungstreue* einer PUV-Schätzung wird über die *systematische Veränderung* der *Stichprobengröße* überprüft. Hierzu wurde der Verlauf des Mittelwertes und der Verlauf des Standardfehlers als Funktion des Stichprobenumfangs in Abbildung 5a,b dargestellt.

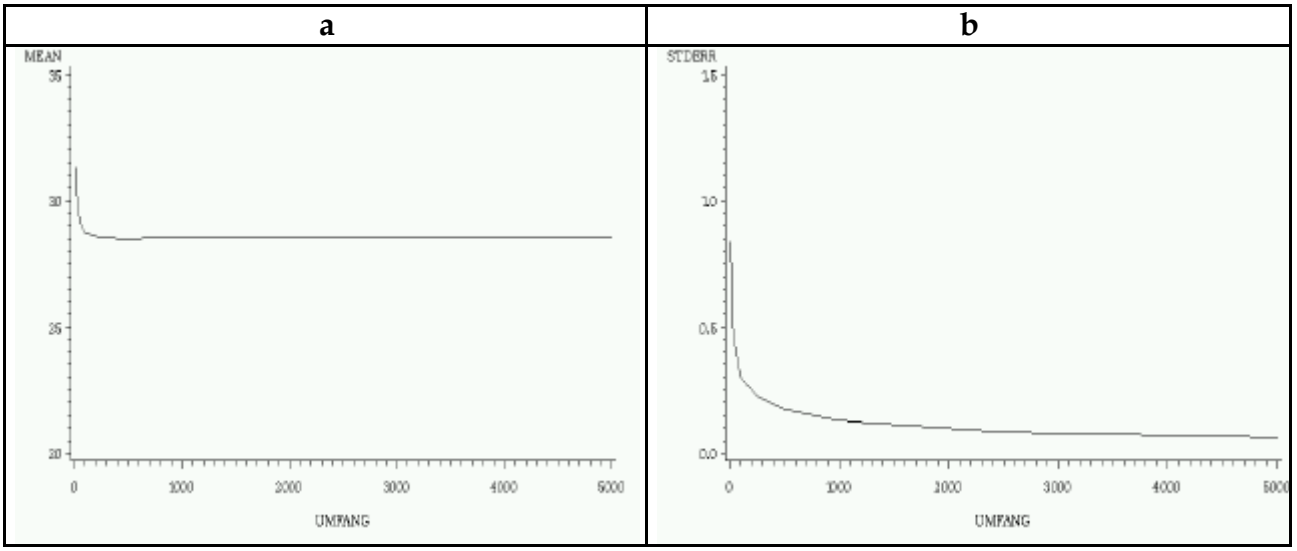


Abbildung 5a,b. Mittelwert (a; MEAN) und Standardfehler (b; STDERR) als Funktion des Stichprobenumfangs aggregiert über verschiedene Verteilungsformen und Reliabilitäten.

Anmerkung: Die Verläufe für die verschiedenen Verteilungsformen weichen nur unwesentlich von den hier gezeigten aggregierten Zusammenhängen ab, weshalb auf eine differenziertere Darstellung verzichtet wurde.

Abbildung 5a zeigt, dass der Mittelwert ab einer Stichprobengröße von ca. 25 Personen kaum wesentlich von den Schätzungen mit 5000 Personen abweicht, d.h., das mit zunehmender Stichprobengröße keine Verzerrung vorliegt, was als Beleg für die Konsistenz des Schätzers interpretiert wird. Lediglich die Schätzung mit sehr kleinen Sichtproben von N = 10 zeigt eine systematisch Überschätzung um ca. 2% (dieser Effekt zeigt sich ebenfalls bei getrennten Analyse je Verteilungsform).

Der Standardfehler (Abb. 5b) sinkt mit zunehmender Stichprobengröße, was auf die Konvergenz der Schätzung hinweist. Die Höhe des Standardfehlers fällt gering aus und bewegt sich in der Regel unter einem PUV-Prozentpunkt, d.h. die Werte konvergieren ausgesprochen schnell auf ein bestimmtes Niveau. Die Simulationsstudie liefert somit keine Hinweise auf einen Bias des PUV-Kennwertes, weshalb dem PUV-Schätzer sehr günstige statistische Eigenschaften zugeschrieben werden können.

Das dritte Ziel der Simulationsstudie lag in der Bereitstellung von Informationen über die Schätzgenauigkeit. Diese benötigt ein Testanwender, um zwei Tests auf einen Unterschied in ihrem PUV-Wert beurteilen zu können. Der Standardfehler bleibt in ca. 94% über alle Bedingungen hinweg unter 1%, in 98,3% unter 1,5% und in weniger als 0.01% über 2%. Dies bedeutet, dass in der Regel ein numerischer Unterschied von 2 PUV-Prozent-Punkten genügt, um auf einen statistisch bedeutsamen Unterschied zu schließen (die praktisch bedeutsame Differenz wird in der Regel deutlich darüber liegen) – und dies schon bei Stichprobenumfängen von nur 25 Probanden. Dieses überraschend gute Ergebnis bei nur wenigen Probanden

erklärt sich über die schnell steigende Anzahl an Paarvergleichen für die Ermittlung des PUV-Wertes (für 25 Probanden ergeben sich 300 Paarvergleiche; vgl. Formel 2). Der PUV-Wert zeigt sich somit als sehr effizient in der Datenausnutzung.

BERECHNUNG AN REALEN DATEN

Im Folgenden sollen für reale Daten der PVU-Wert bestimmt werden[2]. Als Datengrundlage dient die Subskala ‚Resignation' des ‚Stressverarbeitungsfragebogens für Kinder und Jugendliche' (SVF-KJ; N= 1123) und die Subskala ‚Unsicherheit' der Symptomcheckliste SCL-90-R (Derogatis, 1977; Franke, 1995). Die Verteilungsform und die Kennwerte (Reliabilität und das PVU) sind in Abbildung 6 dargestellt, wobei aus den Verteilungen ersichtlich ist, dass der Subtest ‚Unsicherheit' einen Boden-Effekt zeigt.

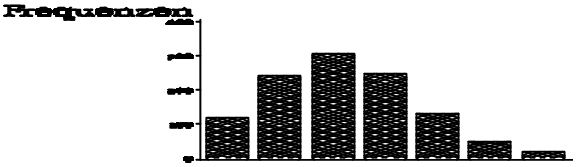
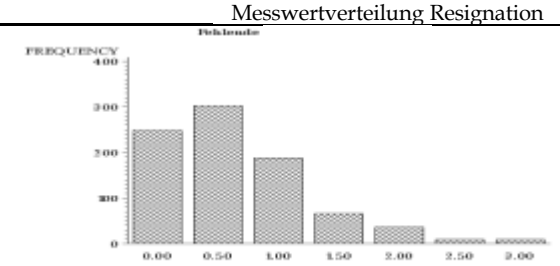
Skalenbezeichnung	Reliabilität	Verteilungsform	PUV in %
Resignation	0.81	 Messwertverteilung Resignation	41.6
Unsicherheit	0.81	 Messwertverteilung Unsicherheit	30,6

Abbildung 6. Zwei Verteilungsformen realer Daten und deren PUV-Koeffizienten am Beispiel des SVF-KJ von Hampel, Petermann und Dickow (1999).

[2] Frau PD Dr. Hampel und Frau Dr. Franke sei für die Überlassung der Daten gedankt.

Diese in Abbildung 6 dargestellte Abweichung der Messwertverteilung von einer Normalverteilung wirkt sich offenbar deutlich auf den PUV-Wert aus. Die ‚Resignationsskala' kann in 41,6% aller Paarvergleiche einen Unterschied finden, während die Subskala ‚Unsicherheit' bei *gleicher* Reliabilität nur in 30,6% aller Paarvergleiche einen Unterschied feststellen kann. Diese Ergebnisse decken sich mit denen aus Abbildung 4. Für die Normalverteilung und eine Reliabilität von .81 würde man einen Wert von ca. 40% erwarten, während für eine mittel schiefe Verteilung nur ein PUV-Wert von ca. 30% erwartet wird. Der PUV-Wert zeigt demnach auch in einer realen Testsituation nachvollziehbare Ergebnisse, die für die Testanwendung über den Einsatz eines Tests mit einfliesen können.

Für den Fall, dass einem Testanwender nur die Verteilungsform und nicht die Rohdaten eines Tests bekannt sind, kann anhand von Abbildung 4 zumindest eine grobe Schätzung des Personenunterscheidungsvermögens erfolgen.

DISKUSSION

Zu Beginn des Beitrags wurde die Notwendigkeit der Ergänzung bestehender Testkennwerte betont und die Tatsache, dass praxisorientierte Kennwerte auch ergonomische Aspekte mit berücksichtigen müssen. Der Kennwert wurde zur Beantwortung der Frage entwickelt, mit welcher Wahrscheinlichkeit zwei Testpersonen anhand ihrer Testwerte unterschieden werden können. Diese Frage beantwortet der Kennwert *direkt*, ohne dass weitere Umrechnungen erforderlich

wären. Er ist zudem *unabhängig* von einer speziellen Testtheorie definiert, womit die angestrebte Vergleichbarkeit erfüllt wird. Der PUV-Koeffizient berücksichtigt hierbei auch explizit die *Verteilung* der Testergebnisse und bezieht damit die notwendigen Randbedingungen ebenfalls mit ein. Der Kennwert zeigt sich in der statistischen Analyse zudem als ein effizienter und erwartungstreuer Schätzer.

Eine inhaltliche Erkenntnis aus den Simulationsstudien ergibt sich aus der Beobachtung, dass die Normalverteilung nicht allen anderen Verteilungen grundsätzlich überlegen ist, wie der Vergleich zu Gleich- oder U-Verteilungen zeigt. Die Einschränkung der diagnostischen Möglichkeiten bei Boden- oder Deckeneffekten konnte über den PUV-Wert quantifiziert werden.

Inwieweit bei der Hervorhebung die Rolle der Verteilung von Messwerten Rückschlüsse auf die Testkonstruktion impliziert wird, ist eine weiterführende Fragestellung, die an dieser Stelle offen bleibt. Sicher ist, dass die Abbildung des Effektes der Verteilungsform auf die Wahrscheinlichkeit, Personen zu unterscheiden, durch den PUV-Koeffizient überwunden werden kann und damit eine wichtige Lücke geschlossen wird.

AUSBLICK: ZUKÜNFTIGE TESTKENNWERTE

Der hier vorgestellte PUV-Kennwert sowie andere Kennwerte des Autors (‚Differenziertheit' und ‚Ausschöpfung', Müller, 2000b; ‚Ausdehnung', Müller, 2000a) und zukünftig geplante praxisorientierte Kennwerte (‚Messeffizienz', Müller, 2000a) könnten zu einer neuen Gliederung von Testkennwerten führen. Die bekannten klassischen Gütekriterien mit ihrem primär wissenschaftlichen Anspruch, die wesentlichen testtheoretischen Informationen präzise darzustellen, werden als Kennwerte der ersten Generation bezeichnet, wohingegen der neue PUV-Koeffizient zur zweiten Generation von Testkennwerten gehört. Die Kennwerte der zweiten Generation sind durch ihren Praxisbezug gekennzeichnet und verwenden die Informationen aus den Kennwerten der ersten Generation bzw. bereiten sie für die Praxis auf. Ebenfalls zur zweiten Generation der Testkennwerte gehört die Differenziertheit und der Ausschöpfungsquotient von Müller (2000b). Zu den Kennwerten der ersten Generation gehören über die Gütekriterien hinaus noch weitere Basisinformationen, wie z. B. die Anzahl von Items, die Testzeit, Ausdehnung (vgl. Müller, 2000a) u.a.

Es erscheint für jegliche Neuentwicklung innerhalb der Testtheorie – wie sie derzeit aktuell in den Item-Response-Theorien erfolgt – wünschenswert, dass sich eine zweite Entwicklungsphase anschließt und die wissenschaftlichen Abhandlungen über das Messmodell für die Beantwortung von praxisrelevanten Fragen aufbereitet werden, wobei Vereinfachungen zugunsten einer leichteren Interpretation verkraftbar erscheinen, solange dies die Testauswahl effektiv verbessert.

LITERATUR

Archer, R. P., Maruish, M., Imhof, E. A. & Piotrowski, C. (1991). Psychological test usage with adolescent clients: 1990 survey findings. Professional Psychology: Research and Practice, 22, 247-252.

Bronstein, I. N. & Semendjajew, K. A. (1997). Taschenbuch der Mathematik. Thun und Frankfurt: Harri Deutsch.

Derogatis, L. R. (1977). SCL-90-R, administration, scoring & procedures manual for the (Revised) version. Johns Hopkins University School of Medicine: Eigendruck.

DIN ISO9000. Qualitätsmanagement und Qualitätssicherungsnormen. Leitfaden zur Auswahl und Anwendung. Berlin 1990.

DIN ISO9001. Qualitätsmanagementsysteme – Modell zur Qualitätssicherung/Modelldarlegung in Design, Entwicklung, Produktion, Montage und Wartung. Berlin, 1994.

DIN ISO9004-1. Qualitätsmanagement und Elemente eines Qualitätssicherungssystems. Leitfaden. Berlin 1994.

DIN ISO9004-2. Qualitätsmanagement und Elemente eines Qualitätssicherungssystems. Leitfaden für Dienstleistung. Berlin 1992.

Franke, G. (1995). SCL-90-R. Die Symptom-Checkliste von Derogatis - Deutsche Version. Göttingen: Beltz.

Frauenhoffer, D., Ross, M. J., Gfeller, J., Searight, H. R. & Piotrowski, C. (1998). Psychological test usage among licensed mental health practitioners: A multidisciplinary survey. Journal of psychological practice, 4, 28-33.

Häcker, H., Leutner, D. & Amelang, M. (Hrsg.) (1998). Standards für pädagogisches und psychologisches Testen. Diagnostica, Suppl.1.

Hampel, P., Petermann, F. & Dickow, B. (1999). Stressverarbeitungsfragebogen nach Janke und Erdmann angepaßt für Kinder und Jugendliche (SVF-KJ). Göttingen: Hogrefe.

Hornke, L. (2000). DIN-Normenentwurf psychologische Eignungsdiagnostik. Podiumsdiskussion auf dem 42. Kongress der DGPS in Jena, vom 24. bis 28. September, 2000.

Kubinger, K. & Wurst, E. (1988). Adaptives Intelligenz Diagnostikum. Weinheim: Beltz.

Kubinger, K. (1997). Editorial zum Themenheft „Testrezensionen: 25 einschlägige Verfahren". Zeitschrift für Differentielle und Diagnostische Psychologie, 18, 1-3.

Lienert, G. & Raatz, U. (1994). Testaufbau und Testanalyse. Weinheim: Psychologische Verlags Union.

Müller, J. M. (2000a). Unterschiedliche Variationen in psychologischen Eigenschaften - eine Interpretation der Erstreckung einer Raschskalierung (eingereicht bei der Zeitschrift für Differentielle und Diagnostische Psychologie).

Müller, J. M. (2000b). Die Differenziertheit eines Tests: Ein nach ergonomischen Leitlinien entwickelter Testkennwert zur Darstellung der Messmöglichkeiten für die Praxis. (eingereicht bei der Diagnostica).

Müller, J. M. (2000c). Neue Leistungs- und Effizienzkennwerte für psychologische Testverfahren: Breite, Differenziertheit, Personenunterscheidungsvermögen, Effizienz und Ausschöpfungsquotient. Poster auf dem 42. Kongress der DGPS in Jena vom 24. bis 28. September, 2000.

Müller, J.M. (1998). Erörterungen zur Informativität und dessen Implikationen für die Konstruktion von psychologischen Meßinstrumenten. Zeitschrift für Differentielle und Diagnostische Psychologie, 19, 41.

Murrell, K. F. H. (1969; repr. 1971). Ergonomics – Man in his Working Environment. London: Chapman and Hall.

Piotrowski, C. & Keller, J. W. (1992). Psychological testing in applied settings: A literature review from 1982-1992. Journal of Training & Practice in Professional Psychology, 6, 74-82.

Piotrowski, C., Belter, R. W. & Keller, J. W. (1998). The Impact of „Managed Care“ on the Practice of Psychological Testing: Preliminary Findings. Journal of Personality Assessment. 70, 441-447.

Ponocny, I. & Ponocny-Seliger, E. (2000). Die Testanalyse mittels exakter Rasch-Modelltests und des Windows-Programms T-Rasch 1.0 – eine erste Bilanz. Vortrag auf dem 42. Kongress der DGPS in Jena vom 24. bis 28. September, 2000.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen, Danmarks Paedodogiske Institut. Chicago: University of Chicago Press.

Rost, J. (1999). Was ist aus dem Rasch-Modell geworden. Psychologische Rundschau, 50, 140-156.

Rost, J. (2000). Haben ordinale Rasch-Modelle variierende Trennschärfen? Eine Antwort auf die Wiener Repliken. Psychologische Rundschau, 51, 36-37.

Schmidtke, H. (Hrsg.) (1993): Ergonomie. München: Hanser.

Schorr, A. (1995). Stand und Perspektiven diagnostischer Verfahren in der Praxis. Ergebnisse einer repräsentativen Befragung westdeutscher Psychologen. Diagnostica, 41, 3-20.

Steck, P. (1997). Aus der Arbeit des Testkuratoriums. Psychologische Testverfahren in der Praxis. Diagnostica, 43, 267-284.

Testkuratorium der Föderation für die Testbeurteilung (1986). Beschreibung der einzelnen Kriterien für die Testbeurteilung, Diagnostica, 32, 358-360.

Wade, T. C. & Baker, T. B. (1977). Opinions and use of psychological test. American Psychologist, 32, 874-882.

Weiß, R. & Osterland, J. (1979). Grundintelligenztest CFT1 Skala1. Braunschweig: Westermann.

Zwingmann, C., Metzger, D. & Jäckel, W.H. (1998). Short Form-36 Health Survey (SF-

36): Psychometrische Analysen der deutschen Version bei Rehabilitanden mit chronischen Rückenschmerzen. Diagnostica, 44, 209-219.

ANHANG A

```
/* Datei-Name:  PUV_TEST.sas                                        */
/* Autor:       Jörg Michael Müller 0421-218-9131                   */
/*              email: joergm@uni-bremen.de                         */
/* Gruppe:      Universität Bremen                                  */
/* Datum:       29092000                                            */
/* Inhalt:      Berechnet das Personenunterscheidungsvermögen       */
/*              eines Tests. Datei (sasfile) und Testwert (Score)*/
/*              plus die Reliabilität (reli) müssen angegeben       */
/*              Reliabilität anstelle .80 ganze Zahlen (also 80) */
/*              werden. (Aufruf des Makro siehe die letzen 2 Zeilen*/
*****************************************************************;

options nosource nonotes nodate nomlogic nomprint nomerror nosymbolgen
nomrecall;

%macro means (datain, score);
%global s frq;
proc means data=&datain noprint;
        var &score;
output out=o std=s n=frq;
data _null_; set o;
call symput("frq",left(frq));
call symput("s",s);
run;
%mend means;

%macro TPU(datain, score);
%global TPU;
%means(&datain, &score)
data tpu; set &datain;
        %let TPU=%eval((&frq * (&frq -1))/2);
        %put &TPU;
run;
%mend TPU;

%macro M_EPU(datain, reli, score);
%global EPU;
data d2; set &datain(keep=&score);
reli=&reli/100;
        krit= 1.96 * (&s * (sqrt (2 * (1-reli) )));
data d3; set d2;
call symput ('krit', krit);
run;
proc transpose data=&datain(keep=&score) out=ot prefix=v;  run;
data d4; set ot;
array a{*} v1-v&frq;
array b{*} v1-v&frq;
EPU=0;
```

```
krit=abs(&krit);
do i= 1 to &frq;
do j= i+1 to &frq;
diff=abs(a{i} - b{j});
if diff > krit then EPU=EPU+1;
end;
end;
data fertig; set d4;
TPU=&TPU;
PUV=(EPU/TPU)*100;
reli=&reli/100;
Personen=&frq;
proc print; var personen reli tpu epu puv;
title1 "Der Testscore &score erreicht";
title2 "ein Personenunterscheidungsvermögen von ... (siehe PUV)"; run;
title1;
title2;
run;
%mend M_EPU;

*********************************;
* Aufruf der Marcros;
* SAS-Datensatz (sasfile) in work.sasfile,
* die den Summenscore (score) enthält, sowie die
* Reliabilität (Reli);

%tpu(sasfile, score)
%m_epu(sasfile, reli, score)
```

ANHANG B

```
%macro random(zeilen, var);
data macro.simdata;
do n=1 to &zeilen./2;
x = ranpoi(0, 1);
output;
end;
do n=0 to %eval(&zeilen -2)/2;
x =12- ranpoi(0, 1);
output;
end;
run;
%mend random;
```

Variationsbreite psychologischer Eigenschaften: Definition und Messung über die Raschskalierung

High and low variations in psychological dimensions: definition and measurement with a Rasch scale

ABSTRACT

Psychometric assessment do not concern about different extensions of distributions from different measurements because variability cannot be compared without an universal scaling. Common standardization (i.e. stanine) puts all distribution to an equivalent range. This article asummes differences in the variability of a psychological dimension and shows how to display these differences. For this purpose a new kind of standardization based on the differences of probabilities is shown, which can be derived from the application of a Rasch scale. This application to measure the *spreedness* of a psychological dimension leads to a new interpretation of a Rasch scale. The spreadness is a characteristic of the trait (specifically hypothesis), and indifferent to the methodology approach (indifference hypothesis). This hypotheses are confirmed by proofing the extension of the Rasch scales of different tests in different modi. The consequences and possibilities for personality psychology and diagnostic are discussed.

Keywords: Scaling, Psychometrics, Item-Response-Theory, Psychological Assessment.

ZUSAMMENFASSUNG

Die Differentielle Psychologie geht bislang davon aus, dass die Variation von Personen in verschiedenen psychologischen Eigenschaften vergleichbar groß ist. Diese Annahme wird hinterfragt und einer empirischen Überprüfung zugänglich gemacht. Zunächst wird der Begriff der *Ausdehnung* einer psychologischen Eigenschaft eingeführt und anhand eines Beispiels erläutert. Innerhalb des Beispiels wird zudem eine Möglichkeit zur Messung der Ausdehnung aufgezeigt. Die Maßeinheit definiert sich über die Differenz von Lösungswahrscheinlichkeiten und bildet damit eine testübergreifende Vergleichsbasis. Es zeigt sich, dass bei Gültigkeit des Raschmodells sich diese Differenz in den Lösungswahrscheinlichkeiten bestimmen lässt und die Raschskala selbst als Maß der Ausdehnung interpretiert werden kann. Die sich aus der Einführung der Ausdehnung einer Eigenschaft eröffnenden psychodiagnostischen Fragestellungen und Möglichkeiten werden aufgezeigt.

Schlagwörter: psychometrische Skalierung, Rasch-Skalierung, IRT, psychologische Diagnostik.

EINFÜHRUNG

Die Annahme, dass sich Personen in einer psychologischen Eigenschaft[1] unterscheiden, ist grundlegend für die Differentielle Psychologie und psychologische Diagnostik, die die Messung dieser Unterschiede zum Gegenstand haben (Stern, 1900). Das Ziel einer Diagnostik liegt hierbei in der Skalierung von Eigenschaftsausprägungen für den Vergleich von Personen (Vergleiche innerhalb einer Messwertverteilung als Norm). Es blieb bislang offen, inwieweit Größenunterschiede in den Messwertverteilungen (Range der Messwerte) verschiedener Eigenschaften vorliegen, z. B. ob Intelligenz (nach Kubinger & Wurst, 1988) eine variationsreichere Eigenschaft als die Einstellung zur Strafrechtsreform (nach Wakenhut, 1974) ist[2]. Es wäre für die Differentielle Psychologie eine grundsätzliche Erkenntnis, ob sich Eigenschaften in der Dispersion von Ausprägungen unterscheiden. Die maximale Variation von Personen in einer Eigenschaft wird als die *Ausdehnung* einer Eigenschaft bezeichnet. Der Begriff, wie auch ein Ansatz zur Messung der Ausdehnung einer Eigenschaft wird am Beispiel *Spielstärke im Schach* veranschaulicht. Das Schachbeispiel dient dabei zur Verdeutlichung des Sachverhaltes, dass Unterschiede in der Ausdehnung von Eigenschaften grundsätzlich zu erwarten sind. Gleichzeitig stellt das Beispiel eine Methode vor, mit dessen Hilfe die Ausdehnung gemessen werden kann.

[1] Es sollen im Weiteren nur eindimensionale Eigenschaften betrachtet werden.

[2] Als Hilfskognition kann sich der Leser eine psychologische Dimension auch als ‚Faden' vorstellen, auf welchem die Personen ‚aufgereiht' sind. Entsprechend dieser Analogie interessiert die ‚Länge' des Fadens.

DAS SCHACHBEISPIEL

Zunächst spielen zwei Personen (Spieler A und Spieler B) mehrere Schachpartien, wobei nur *Siege* oder *Niederlagen* beachtet werden (*Patt* wird zur Vereinfachung des Beispiels ausgeschlossen). Es zeigt sich nach mehreren Partien, dass einer der beiden Spieler eine durchschnittlich höhere Gewinnwahrscheinlichkeit besitzt. Liegt die Gewinnwahrscheinlichkeit von Spieler A mit Spieler B bei 2:1, so heißt dies, dass Spieler A im Durchschnitt zwei von drei Partien gewinnt. Eine Expertengruppe definiert nun willkürlich diesen Unterschied in den Gewinnchancen als Maßeinheit des *besseren Schachspielers*[3]. Lässt man nun in gleicher Weise die *gesamte* schachspielende Population mehrfach gegeneinander antreten, so dass sich die Gewinnwahrscheinlichkeiten schätzen ließen, würden sich *Stufen* von *Spieler X besser als Spieler Y* bilden (vgl. Abb. 1). Bei jeder *Stufe* wäre zu verlangen, dass der *bessere* Spieler die *schlechteren* Spieler in mindestens 2 von 3 Partien schlägt.

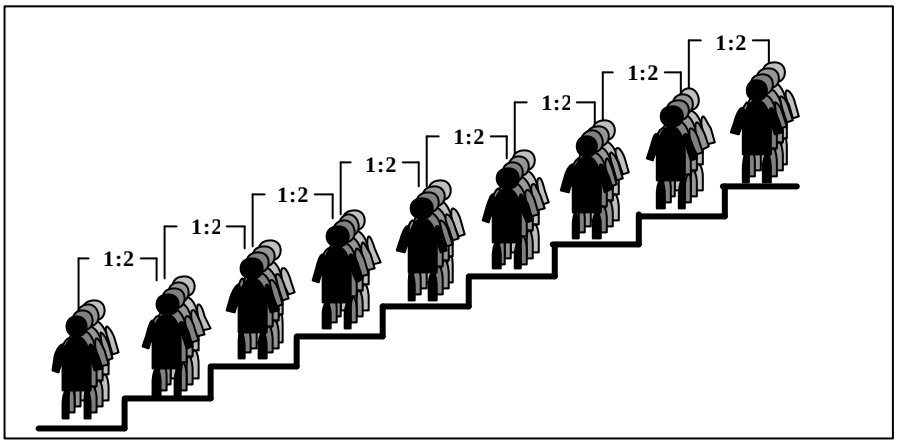


Abbildung 1. Skalierung der Fähigkeit ‚Spielstärke im Schach' über Wahrscheinlichkeitsdifferenzen

[3] Die statistische Definition für ‚besserer Schachspieler' ist hiervon abzugrenzen.

Anmerkung: Auf jeder höheren Stufe befinden sich Personen, die Personen auf einer darunter liegenden Stufe in mindestens 2 von 3 Fällen in einer Schachpartie schlagen.

Die Schachsituation kann als eine einfache Form der Skalierung einer Eigenschaft angesehen werden. Drei Aspekte dieser *Skalierung* sollen im Weiteren interessieren: Der erste Punkt betrifft die Veranschaulichung einer Methode, wie die Ausdehnung einer Eigenschaft operationalisiert werden kann. Anscheinend lässt sich die gesamte Variation der Eigenschaftsausprägungen einer Population in Distanzen von Wahrscheinlichkeitsdifferenzen unterbringen. Zweitens ist aus dem Beispiel zu erkennen, dass diese Anzahl an Distanzen demnach endlich ist. Zuletzt ist zu erwarten, dass sich bei verschiedenen Eigenschaften auch unterschiedliche Ausdehnungen zeigen. Aufgrund des Beispiels lässt sich nun die *Ausdehnung* einer Eigenschaft folgendermaßen definieren (siehe Definition 1):

Definition 1: Definition der *Ausdehnung* einer Eigenschaft

Die *Ausdehnung* einer Eigenschaft bezeichnet die Spannweite der Unterschiede zwischen der kleinsten und größten Eigenschaftsausprägung. Die Eigenschaftsausdehnung wird auf der Basis von Wahrscheinlichkeitsdifferenzen erfasst. Die Ausdehnung ist invariant gegenüber der Messmethode und spezifisch bezüglich des Messinhaltes.

EIN VERSUCH DER MESSUNG DER AUSDEHNUNG ÜBER EINE Z-TRANSFORMATION

Die anhand des Schachbeispiels dargestellte Ausdehnung einer Eigenschaft muss auch in der Messwertverteilung von Testergebnissen existieren. Dieser Abschnitt versucht nun mit einer z-Transformation der Rohwerte diese Unterschiede in der Variation der Personen zu messen, obwohl sich zeigen wird, dass dieser Versuch nicht zum Ziel führen wird. Da der Ansatz über eine z-Transformation jedoch immer wieder in Diskussionen auftaucht, soll der Grund des Scheiterns in diesem Abschnitt dargestellt werden, bevor ein anderer Ansatz verfolgt wird.

Verschiedene Personen erzeugen aufgrund ihrer unterschiedlichen Eigenschaftsausprägung eine Streuung in den Messwerten. Die Operationalisierung der Ausdehnung führt hier auf die Frage: „Wie *breit* ist diese Verteilung bzw. die Streuung?“ Zunächst soll die Verteilung der Messwerte für einen Test *A* bzgl. einer Eigenschaft *U* betrachtet werden (vgl. Abbildung 2a). Die x-Achse repräsentiert die Rohpunktwerte aus Test *A*, während der Pfeil unter der Verteilung die (noch nicht messbare) Ausdehnung der Eigenschaft *U* repräsentiert. Abbildung 2b zeigt eine Standardnormalverteilung und Abbildung 2c zeigt eine zweite Verteilung von Messwerten für einen Test *B* bzgl. einer Eigenschaft *V*, wobei die Pfeile, die die Abbildungen 2a, b, c verbinden, eine z-Transformation symbolisieren sollen.

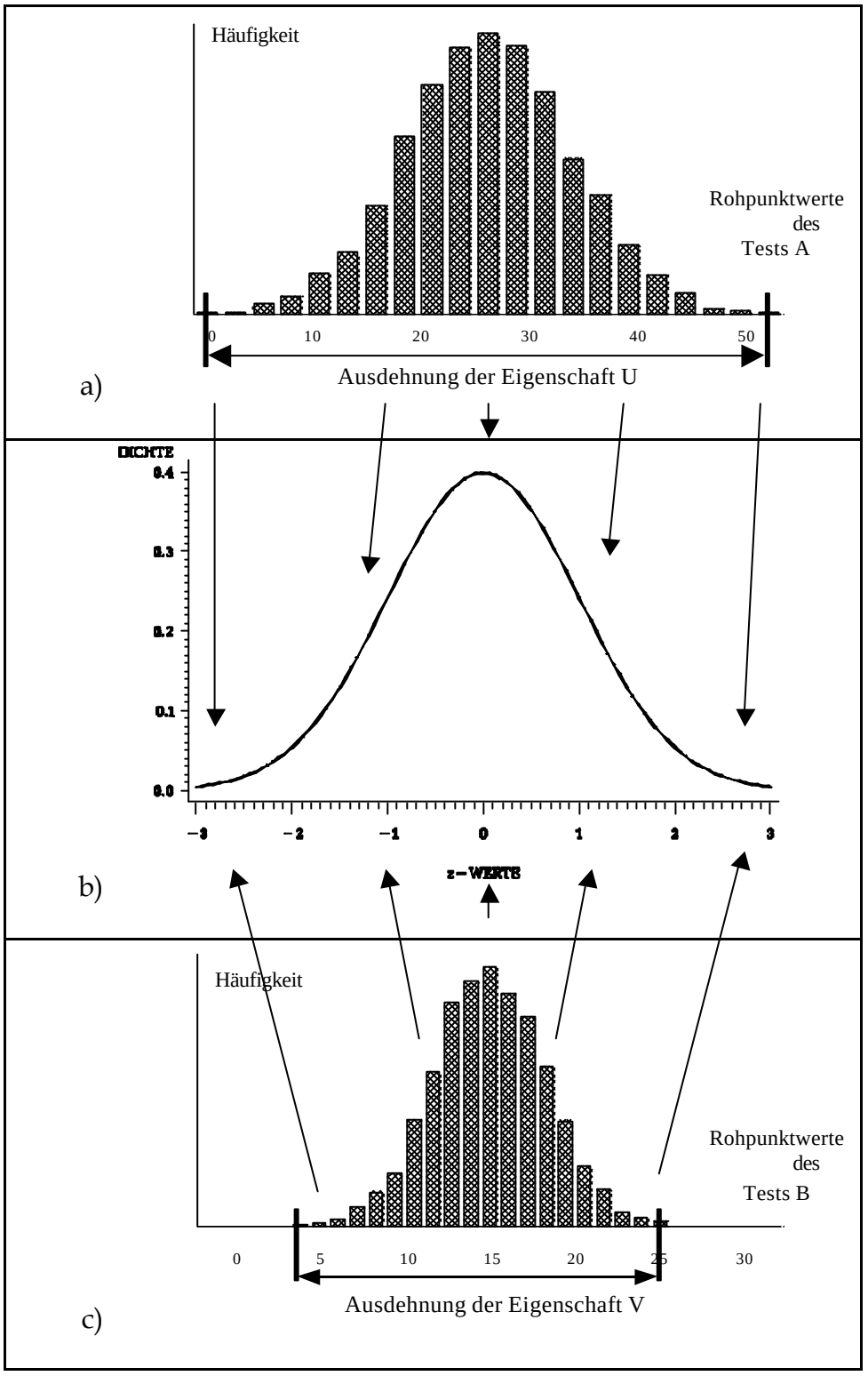


Abbildung 2a, b, c. Zwei Messwertverteilungen (a, c) und eine Standardnormalverteilung (b)

Die Rohwerte in Abbildung 2a und 2c scheiden als Vergleichsgrundlage der Streuungen aus, da sie in unterschiedlichen Einheiten gemessen wurden. Um die Messwerte vergleichbar zu machen, werden sie gewöhnlich z-transformiert. Das Ziel einer z-Transformation besteht darin, die Position einer Person in verschiedenen Messwertverteilungen vergleichbar zu machen. Dieses Ziel unterscheidet sich jedoch von der Forderung, die Unterschiede in der Variation der Messwerte zu erhalten und darzustellen. Die z-Standardisierung führt dazu, dass die Standardabweichung *jeder* empirischen Verteilung gleich eins wird und die ursprünglich existierenden Unterschiede in den Streuungen auf die *gleiche Größe* angeglichen werden, d. h. bei jeder Messwertverteilung resultiert *immer* ein Range von 6 (zweimal +/- 3 z-Wert-Einheiten, bei einer Begrenzung von 99% der Fläche einer Verteilung).

Andere Transformationen (T-Werte, Stanine, CEEB, Prozentränge; weitere in Aiken, 1997, S. 79) standardisieren verschiedene Messwertverteilungen ebenfalls auf einen *konstanten* Wert. Für das Ziel der Darstellung unterschiedlicher Streuung (=Ausdehnung) ist es jedoch kontraproduktiv, da die Unterschiede in den Messwertverteilungen herausgenommen werden. Aus diesem Grund können die genannten Skalierungen die Ausdehnung nicht abbilden. Im Weiteren wird gezeigt, dass eine Raschskalierung die bestehenden Variationsunterschiede erhält. Ihre Erstreckung wird dabei zum Maß der Ausdehnung.

Im Folgenden soll das Raschmodell auf Analogie zum Schachbeispiel hin untersucht werden, um die im Schachbeispiel definierte Maßeinheit (Differenzen in den Gewinn- bzw. Antwortwahrscheinlichkeiten) auch im Raschmodell darzustellen.

DIE RASCHSKALIERUNG ALS WAHRSCHEINLICHKEITSBASIERTE MESSUNG DER AUSDEHNUNG EINER PSYCHOLOGISCHEN EIGENSCHAFT

Um eine Analogie zu den Antworten einer Person auf Testaufgaben herstellen zu können, müssen die im Schachbeispiel grundlegenden direkten Paarvergleiche zur *Skalierung* der Eigenschaft *Spielstärke im Schach* neu interpretiert werden. Betrachtet man den Gegenspieler B als (Test-)Aufgabe i für Spieler A, so erreicht Spieler A bei Aufgabe i (= Spieler B) eine Lösungswahrscheinlichkeit von ca. 66% (2 von 3). Formal würde dies folgendermaßen ausgedrückt werden: $p_{(A,B)} = f(q_A, q_B)$, d. h. die Gewinnwahrscheinlichkeit ist eine Funktion der Spielstärke der Person A und der Aufgabenschwierigkeit (Spielstärke von Gegenspieler B). Würde ein Spieler gegen sich selbst spielen, so erwartet man eine Gewinnwahrscheinlichkeit von 50%. Die Besonderheit im Schachspiel besteht also darin, dass Spieler sowohl einen Probanden als auch eine Testaufgabe darstellen können, wobei beide den gleichen Parameter erhalten (Person=Schwierigkeit). Bei einem üblichen Testverfahren sind diese Parameter voneinander unabhängig.

Die Gültigkeit eines Raschmodells (Rasch, 1960) vorausgesetzt, ermöglicht dieses bei bekannten Personen- und Itemparameter eine Bestimmung der Lösungswahrscheinlichkeiten. Formel 1 zeigt die Bestimmung der Lösungswahrscheinlichkeiten $p(x_{Ai}=1)$ für eine Person A bzgl. eines Testitems *i*.

Formel 1

$$p(x_{Ai}) = \frac{\exp(x_{Ai}(q_A - s_i))}{1 + \exp(q_A - s_i)}$$

Bestimmt man für zwei Personen deren Lösungswahrscheinlichkeit bzgl. eines Vergleichsitems, dann lässt sich hierüber die benötigte Wahrscheinlichkeitsdifferenz bilden. Auf dieser Grundlage ließe sich dann die Anzahl der Wahrscheinlichkeitsdifferenzen wie im Schachbeispiel bestimmen.

Welche Parameter lagen im Schachbeispiel vor? Spieler B wird, um einen Anker in der Skalierung zu haben, auf Null gesetzt, d. h. Spieler A hatte eine Aufgabe mit der Schwierigkeit Null zu lösen. Damit liegen zunächst zwei Parameter fest, $\theta_B = 0$ und $\sigma_i = 0$. Der Personenparameter von Spieler A wird auf eins gesetzt ($\theta_A = 1$; im Schachbeispiel befand er sich ebenfalls eine Stufe über Person B). Damit ergibt sich innerhalb des Raschmodells eine abgeleitete Wahrscheinlichkeitsdifferenz als Maßeinheit statt der bislang willkürlich festgesetzten. Die Maßeinheit ergibt sich aus der Differenz der Lösungswahrscheinlichkeit für beide Spieler und kann über die Itemfunktion veranschaulicht werden (Abbildung 3).

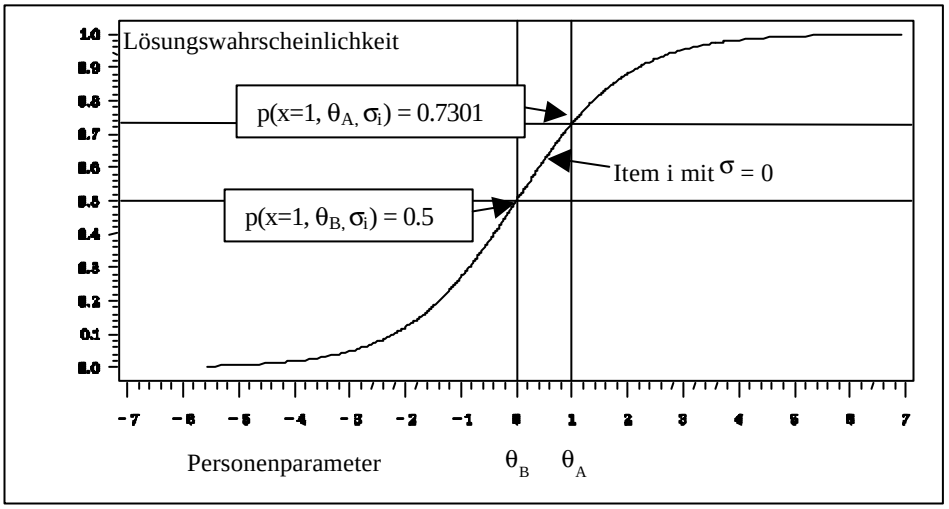


Abbildung 3. Die Itemfunktion des Raschmodells

Für zwei Personen A und B werden zwei Hilfslinien in Abbildung 3 eingezeichnet, anhand derer man die geschätzte Lösungswahrscheinlichkeit bzgl. des Items σ_i bestimmen kann. Als Lösungswahrscheinlichkeiten erhält man $p(x=1, \theta_A, \sigma_i) = 0.7301$ für Person A und $p(x=1, \theta_B, \sigma_i) = 0.5$ für Person B. Die Differenz der Lösungswahrscheinlichkeiten von Person A und Person B bzgl. des Items *i* beträgt $\Delta p=0.2301$. Anstelle der willkürlichen Maßeinheit (zwei von drei Partien) tritt nun der Logit der Raschskalierung. Die ‚Stufenbildung' innerhalb des Raschmodells wird in Abbildung 4 veranschaulicht.

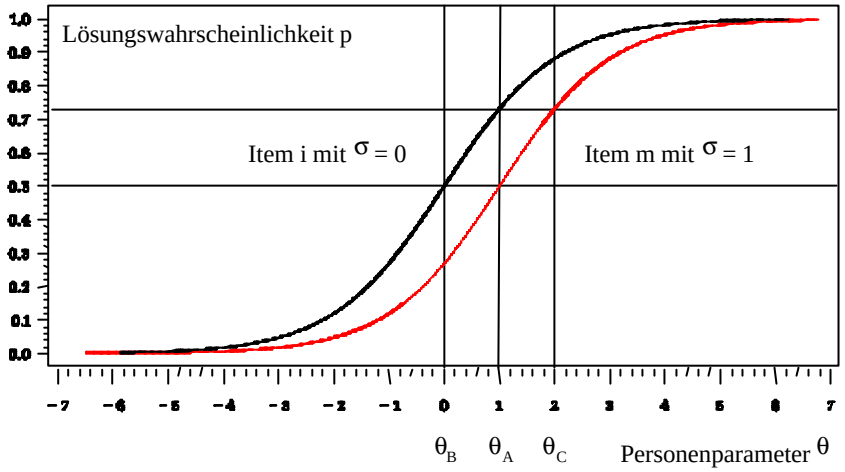


Abbildung 4. Stufenbildung innerhalb des Raschmodells

Tritt ein weiterer Spieler C mit einer höheren Fähigkeit auf, der wiederum Spieler A in 2 von 3 Partien schlagen kann, so verschiebt sich das Niveau um eine Stufe nach oben. Im Schachbeispiel existiert allerdings noch keine testtheoretisch fundierte Metrik, um die quantitative Distanz von Spieler B zu Spieler C mit zwei anzugeben. Diese Metrik liegt jedoch bei der Gültigkeit eines Raschmodells vor, den die Personenparameter liegen nach Fischer (1988) mindestens auf dem Niveau einer Intervallskala.

Die Maßeinheiten des Raschmodells können demnach als Wahrscheinlichkeitsdifferenz interpretiert werden. Diese besondere Art der Definition einer Maßeinheit im Raschmodell bietet – im Unterschied zu den zahlreichen Flächentransformationen – eine feste Vergleichsgrundlage der Ausdehnung in inhaltlich *verschiedenen* Eigenschaften.

Die Ausdehnung A einer Merkmalsmessung ist damit identisch mit der Erstreckung einer Raschskalierung, bzw. der Differenz aus dem höchsten Personenparameter q_{max} und dem kleinsten Personenparameter q_{min} (Formel 2).

Formel 2

$$A = q_{\max} - q_{\min}$$

Aus einem Raschmodell lässt sich somit ein Maß ableiten, das zur Bestimmung der Ausdehnung einer psychologischen Eigenschaft verwendet werden kann.

INVARIANZ GEGENÜBER DER MESSMETHODE - SPEZIFITÄT BZGL. DES INHALTES

Mit der Operationalisierung der Ausdehnung einer Eigenschaft können nun Hypothesen getestet werden, die sich aus der Definition ableiten lassen. Die Ausdehnung ist der Definition nach invariant gegenüber der Messmethode (die Repräsentativität der Personenstichprobe vorausgesetzt), d. h. die Ausdehnung ist

unabhängig von der Itemstichprobe (Anzahl und Auswahl der Items) und – da die Messgenauigkeit u. a. eine Funktion der Items ist (vgl. Informationsfunktion in Rost, 1996) – auch von der Messgenauigkeit.

Wenn die Ausdehnung von diesen Merkmalen der Messmethode unabhängig ist, so sollten folgende Hypothesen nicht widerlegt werden:

1. Aspekt der Invarianzhypothese: die Itemanzahl (Kurzform gleich der Langform) hat keinen Einfluss auf die Ausdehnung (H1).

2. Aspekt der Invarianzhypothese: die Itemauswahl hat keinen Einfluss auf die Ausdehnung (Parallelform gleich der Standardvorgabe; H2).

3. Aspekte der Invarianzhypothese: die Reliabilität hat keinen Einfluss auf die Ausdehnung (H3).

4. Spezifitätshypothese: unterschiedliche Subtests zeigen unterschiedliche Ausdehnungen (H4).

Diese Hypothesen werden anhand der Angaben im Manual des AID von Kubinger und Wurst (1988) überprüft (siehe Tabelle 1).

Zur Überprüfung der Invarianzhypothese gegenüber der Erfassungsmethode (H1, H2 als Einzelvergleiche im Faktor Methode: Standardvorgabe, Kurzform, Parallelform) wurde jeweils eine einfaktorielle Varianzanalyse gerechnet, deren Ergebnis hypothesenkonform nicht signifikant ausfiel ($F_{(2;22)}$=1,18; p=0.326)[4]. Die

[4] Die leicht niedrigeren Werte für die Kurzform lassen sich aus mathematischen Problemen mit der Schätzung von extremen Parametern erklären. Dies erschwert den Vergleich verschiedener Raschskalierungen mit extrapolierten und nicht extrapolierten Werten. (Dieser Hinweis stammt von Prof. Gittler; persönliche Mitteilung vom 15.2.2000).

Ausdehnung ist zudem von den messtechnischen Eigenschaften des Tests unabhängig (H3), wie die Reliabilität in der letzten Spalte in Tabelle 1 zeigt; diese korreliert hypothesenkonform nicht signifikant mit der Ausdehnung nur zu r=0.162 (p=0.678). Die erwartete Spezifität der Ausdehnung bzgl. der Eigenschaft (H4) konnte ebenso bestätigt werden ($F_{(8;16)}$ =19,53; p<0.0001).

Tabelle 1

Ausdehnung (Erstreckung der Raschskalierung)
verschiedener Eigenschaften mit verschiedenen Messmethoden
im AID von Kubinger und Wurst (1988)

Untertest-Nr.	Untertest	Standardmäßige Vorgabe oder ‚Überlangform'	Kurzform	Parallelform	r **
1	Alltagswissen	21,1	19,2	21,3	.95
2	Realitätssicherheit	13,3	13,2	13,1	.70
3	Angewandtes Rechnen	21,7	17,7	20,5	.95
4	Soziale und sachliche Folgerichtigkeit	14,6	14,7	14,8	.91
5	Unmittelbares Reproduzieren	*	*	*	
6	Synonyme Finden	16,9	13,9	16,0	.94
7	Kodieren und Assoziieren	*	*	*	
8	Antizipieren und Kombinieren	22,4	*	*	.81
9	Funktionen Abstrahieren	14,7	12,9	14,5	.93
10	Analysieren und Synthetisieren	14,2	14,2	14,7	.95
11	Soziales Erfassen und sachliches Reflektieren	15,5	13,5	15,5	.94

Anmerkung: Die Werte in den Spalten drei bis fünf ergeben sich aus der Differenz des höchsten und niedrigsten Personenparameters der Raschskalierung. Die Werte entstammen den Normierungstabellen von Kubinger und Wurst, 1988, S. 163-179.

* Keine Angaben im Manual.

** Split-half-Reliabilität mit einer Korrektur nach Spearman-Brown, aus Kubinger und Wurst, 1988, S. 20.

Die Werte in Tabelle 1 zeigen damit erhebliche Unterschiede zwischen den Eigenschaften (Spezifität), jedoch nur geringe zwischen den unterschiedlichen Erfassungsmodi (Invarianz). Auch andere Ergebnisse deuten in diese Richtung, wie bei Fischer (1974, S.336; Tachistoskoptest: Testhälfte A mit einer Ausdehnung A=8,3 und Testhälfte B mit einer Ausdehnung A=8,7).

Da die Ausdehnung einer Eigenschaft von der Messmethode unabhängig ist, sollte sie – einmal ermittelt – ein stabiles Kennzeichen der Eigenschaft bleiben, unter der Voraussetzung, dass die Eigenschaftsausprägungen selbst zeitinvariant sind. Anhand verschiedener raschskalierter Tests wird exemplarisch die Ausdehnung einiger Eigenschaften in Tabelle 2 zusammengestellt.

Tabelle 2

Exemplarische Eigenschaftsausdehnungen

Eigenschaft	Autor	Jahr	Testab-kürzung	Testbezeichnung	Ausdehn-ung
Verbaler Intelligenztest	Metzler & Schmidt	1992	MWT	Mehrfach-Wortschatz-Test	11,4[1,3,6]
Averbale Intelligenz	Forman & Pieswanger	1979	WMT	Wiener Matrizen Test	8,2[1,3,5]
Einstellung zur Sexualmoral	Wakenhut	1974	-	Sexualmoral	8,1[1,3]
Feldabhängigkeit	Hergovich	1999	-	Gestaltwahrnehmungstest	7,3[1,3]
Einstellung zur Strafrechtsreform	Wakenhut	1974	-	Strafrechtsreform	7,2[1,3]
Beschwerdeliste	Fahrenberg	1975	FBL-K	Freiburger Beschwerdeliste	6,4[1,3,8]
Räumliches Vorstellungs-vermögen	Gittler	1990	3 DW	Dreidimensionaler Würfeltest	5,9[1,3]
Umgang mit Zahlen bei Kindern	Rasch	1960	BPP-N	Zahlenreihentest	3,5
Role-Conflict	Anderson	1983	-	Role-Conflict	2,8[1,3,7]

Anmerkung: Nicht in allen Fällen war eindeutig zu entscheiden, ob die Extremwerte extrapolierte Werte darstellen.

[1] entnommen aus einer Tabelle

[2] entnommen aus der Graphik

[3] nicht extrapolierte Extremwerte

[4] extrapolierte Extremwerte
[5] Daten aus Fischer & Pendl (1980)
[6] Daten aus Metzler & Schmidt (1992)
[7] Daten von Stouffer & Toby (1951)
[8] Daten aus Piel, Hautzinger & Scherbarth-Roschmann (1991)

Nachdem einige Ausdehnungen verschiedener Eigenschaften aufgeführt wurden, lässt sich auch die eingangs gestellte Frage, ob denn Intelligenz oder die Einstellung zur Strafrechtsform Unterschiede bzgl. ihrer Ausdehnung aufzeigen, beantworten: Intelligenz ist eine Eigenschaft, in der sehr viel größere Unterschiede zwischen den Menschen bestehen als bzgl. der Einstellung zur Strafrechtsreform.

DISKUSSION

Dieser Beitrag stellt einen bislang unbeachteten Aspekt der Raschskalierung heraus: die Erstreckung bzw. Distanz vom größten zum kleinsten Personenparameter. Die Raschskalierung der Personenparameter wurde bislang aufgrund ihrer hohen Korrelation ($r > .90$; Rost, 1996) mit den Rohpunktwerten wenig beachtet und wegen des vermeintlich geringen Informationsgehaltes des öfteren nicht gesondert angegeben (z. B. bei Hehl und Wirsching, 1983; WMT, Forman & Piswanger, 1979). Mit der Reinterpretation einer Raschskalierung wird der Liste vorteilhafter Eigenschaften einer Item-Response-Theorie (Rost, 1999) und deren Anwendung in einer Testkonstruktion ein zusätzliches Argument angefügt. Mit der Erfassung der Ausdehnung einer Eigenschaft wird dem Diagnostiker eine neue Information über seinen Messgegenstand an die Hand gegeben, nämlich inwieweit sich Personen in der zu messenden Dimension überhaupt (im quantitativen Sinne) unterscheiden.

Dies sagt bislang (noch) nichts darüber aus, inwieweit diese Unterschiede auch von prädiktiver Bedeutung sind. Allerdings könnte die Ausdehnung – falls sie denn mit einer prädiktiven Validität einher geht – diesen Sachverhalt erklären helfen. Die Ausdehnung kann zudem für die Überprüfung der Konstruktvalidität verwendet werden. Geben zwei Tests vor, das gleiche Konstrukt zu messen, so müsste auch die gleiche Ausdehnung in beiden Tests vorliegen. Falls dennoch bedeutsame Unterschiede auftreten, würde dies gegen eine konvergente Validität sprechen. Die Ausdehnung ergänzt damit andere Erkenntnisse über das psychologische Konstrukt, wie sie u. a. aus der Überprüfung der dimensionalen Struktur folgen. Fasst man eine psychologische Eigenschaft als Vektor auf, so wird neben der Richtung (relativ zu anderen Dimensionen) nun auch die Länge des Vektors bestimmbar.

Vorteilhaft ist die ökonomische Erfassung der Ausdehnung einer Eigenschaft; allerdings sind gerade die Randbereiche einer Raschskalierung besonders unreliabel (Kubinger und Wurst, 1988) – worunter die Reliabilität des Ausdehnungskoeffizienten insgesamt leidet. Dies spricht nicht grundsätzlich gegen die Interpretation der Erstreckung einer Raschskalierung. Es sollte deshalb vor einer Interpretation des Maßes auf zweierlei geachtet werden: erstens müssen genügend schwierige und leichte Items im Itempool enthalten sein, und zweitens müssen sich genügend Personen mit extremen Eigenschaftsausprägungen in der Konstruktionsstichprobe befinden. Sind beide Punkte erfüllt, so resultiert ein stabiler Ausdehnungskoeffizienten (vgl. die Verminderung der Ausdehnung bei Gittler, 1990, S.56).

AUSBLICK

Die Beschreibung der statistischen Eigenschaften der Ausdehnung (insbesondere dessen Standardfehler) wurde in diesem Beitrag nicht behandelt, die Schätzung von Konfidenzintervallen für Personenparameter ist jedoch prinzipiell gelöst (Rost, 1996). Zukünftig wäre eine Bestimmung der zahlreichen psychologischen Konstrukte denkbar, wodurch sich neue Fragestellungen für die Differentielle Psychologie und die psychologische Diagnostik ergeben:

- Warum sind einige Eigenschaften variationsreicher als andere?
- Welchen Einfluss hat Übung auf die Ausdehnung?
- Lassen sich Übungsgewinne in Maßeinheiten der Raschskalierung ausdrücken?
- Welchen Einfluss hat die Vertrautheit der Testpersonen mit den Testaufgaben eines Leistungstests? Zum Beispiel sind abstrakte Würfelaufgaben (Messung der Raumvorstellung mit Gittlers 3-DW, 1990) einer Testperson eher weniger vertraut als verbale Aufgabentypen im Sinne des MWT (Lehrl, Merz, Erzigkeit & Galster, 1974)
- Ist die Ausdehnung pro Lebensalter ein Indikator für die Entwicklungsgeschwindigkeit bei Kindern und Jugendlichen in einer Dimension?
- Welchen Einfluss hat die Konstruktbreite auf die Ausdehnung? Würfelaufgaben sind ein Beispiel für eine eng umgrenzte Aufgabenstellung im Gegensatz zu dem Aufgabenpool, welcher zur Bestimmung der crystallized general intelligence (Horn & Cattell, 1966) verwendet wird.

Bislang ist die Erfassung der Ausdehnung noch an eine Raschskalierung gebunden und damit auf eindimensionale Konstrukte mit dichotomen Antwortformat beschränkt. Inwieweit sich dieses Prinzip der Wahrscheinlichkeitsdifferenzen auch auf andere IRT Modelle übertragen lässt, bleibt offen.

LITERATUR

Aiken, L. R. (1997): Psychological Testing and Assessment. Boston: Allyn & Bacon.

Anderson, E. B. (1983): Analysing Data using the Rasch-Model. In: S. B. Anderson, J. & S. Helmick: On Educational Testing. San Francisco: Jossey-Bass.

Hehl, F.-J. & Wirsching, M. (1983): Psychosomatischer Einstellungs-Fragebogen (PEF). Göttingen: Hogrefe.

Hergovich, A. (1999): Vorstellung und Validierung des Gestaltwahrnehmungstests zur Messung der Feldabhängigkeit. Diagnostica, 45(1), 20-34.

Horn, J. L. & Cattell, R. B. (1966): Refinement and test of the theory of fluid and crystallized ability intelligence. Journal of Educational Psychology, 57, 253-270.

Fahrenberg, J. (1975): Die Freiburger Beschwerdeliste (FBL-K). Göttingen: Hogrefe.

Fischer, G. H. (1974): Einführung in die Theorie psychologischer Tests. Bern: Huber.

Fischer, G. H. (1988): Spezifische Objektivität: eine wissenschaftstheoretische Grundlage des Rasch-Modells. In: K. D. Kubinger: Moderne Testtheorie – Ein Abriß samt neuester Beiträge. Weinheim: PVU.

Fischer, G. H. & Pendl, P. (1980): Individualized Testing on the Basis of the Dichotomous Rasch Model. In: L. J. T. van der Kamp, W. F. Langerak, & D. N. M. de Gruijter: Psychometrics for Educational Debates. Chichester: Wiley.

Forman, A. K. & Piswanger, K. (1979): Wiener Matrizen-Test. Weinheim: Beltz.

Gittler, G. (1990): Dreidimensionaler Würfeltest 3DW. Weinheim: Beltz.

Kubinger, K. & Wurst, E. (1988): Adaptives Intelligenz Diagnostikum. Weinheim: Beltz.

Lehrl, S., Merz, J., Erzigkeit, H. & Galster, V. (1974): Der WMT-A – ein wiederholbarer Intelligenz-Kurztest, der weitgehend unabhängig von seelisch-geistigen Störungen ist. Der Nervenarzt, 45, 364-369.

Metzler, P. & Schmidt, K.-H. (1992): Rasch-Skalierung des Mehrfachwahl-Wortschatztests (MWT). Diagnostica, 38(1), 31-51.

Piel, E., Hautzinger, M. & Scherbarth-Roschmann, P. (1991): Analyse der Freiburger Beschwerdeliste (FBL-K) mit Hilfe des stochastischen Testmodells von Rasch. Diagnostica, 37(3), 226-235.

Rasch, G. (1960): Probabilistic models for some intelligence and attainment tests. Copenhagen, Danmarks Paedodogiske Institut (Chicago: University of Chicago Press).

Rost, J. (1996): Lehrbuch der Testtheorie, Testkonstruktion. Bern: Huber.

Rost, J. (1999): Was ist aus dem Rasch-Modell geworden? Psychologische Rundschau, 50(3), 140-156.

Stern, W. (1900): Über Psychologie der individuellen Differenzen. Leipzig: Barth.

Stouffer, S. A. & Touby, J. (1951): Role Conflict and Personality. American Journal of Sociology, 56, 395-406.

Wakenhut, R. (1974): Messung gesellschaftlich-politischer Einstellungen mithilfe der Rasch-Skalierung. Bern: Huber.