

2 Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien)

Helfried Moosbrugger & Augustin Kelava

- 2.1 Objektivität – 8**
 - 2.1.1 Durchführungsobjektivität – 9
 - 2.1.2 Auswertungsobjektivität – 9
 - 2.1.3 Interpretationsobjektivität – 10

- 2.2 Reliabilität – 11**
 - 2.2.1 Retest-Reliabilität – 12
 - 2.2.2 Parallelttest-Reliabilität – 12
 - 2.2.3 Testhalbierungs-Reliabilität – 12
 - 2.2.4 Innere Konsistenz – 13

- 2.3 Validität – 13**
 - 2.3.1 Inhaltsvalidität – 15
 - 2.3.2 Augenscheinvalidität – 15
 - 2.3.3 Konstruktvalidität – 16
 - 2.3.4 Kriteriumsvalidität – 18

- 2.4 Skalierung – 18**

- 2.5 Normierung (Eichung) – 19**

- 2.6 Testökonomie – 20**

- 2.7 Nützlichkeit – 21**

- 2.8 Zumutbarkeit – 22**

- 2.9 Unverfälschbarkeit – 23**

- 2.10 Fairness – 23**

- Literatur – 25**

- Wenn man mit der Frage konfrontiert wird, worin der eigentliche Unterschied zwischen einem unwissenschaftlichen »Test« (etwa einer Fragensammlung) und einem wissenschaftlich fundierten, psychologischen Test besteht, so ist die Antwort darin zu sehen, dass sich ein psychologischer Test dadurch unterscheidet, dass er hinsichtlich der Erfüllung der sog. Testgütekriterien empirisch überprüft wurde.

Testgütekriterien

Die Testgütekriterien stellen ein Instrument der Qualitätsbeurteilung psychologischer Tests dar. Das Testmanual eines vorliegenden Tests sollte in geeigneter Weise darüber informieren, welche Testgütekriterien in welcher Weise erfüllt sind. Als Gütekriterien haben sich in den vergangenen Jahren eine Reihe von Aspekten etabliert (Testkuratorium, 1986), die nicht zuletzt auch die Basis der DIN 33430 zur berufsbezogenen Eignungsbeurteilung bilden (DIN 2002, vgl. Westhoff, Hellfritsch, Hornke, Kubinger, Lang, Moosbrugger, Püschel & Reimann, 2004). Üblicherweise werden folgende zehn Kriterien unterschieden (vgl. hierzu auch Kubinger, 2003):

1. Objektivität
2. Reliabilität
3. Validität
4. Skalierung
5. Normierung (Eichung)
6. Testökonomie
7. Nützlichkeit
8. Zumutbarkeit
9. Unverfälschbarkeit
10. Fairness

2.1 Objektivität

Die Objektivität eines Tests ist ein wesentliches Gütekriterium, das die Vergleichbarkeit von Testleistungen verschiedener Testpersonen sicherstellt.

Es wird wie folgt definiert:

Definition

Ein Test ist dann objektiv, wenn er dasjenige Merkmal, das er misst, unabhängig von Testleiter, Testauswerter und von der Ergebnisinterpretation misst.

Drei Aspekte der Objektivität

Objektivität bedeutet, dass den Testdurchführenden kein Verhaltensspielraum bei der Durchführung, Auswertung und Interpretation eingeräumt wird. Völlige Objektivität wäre also dann gegeben, wenn sowohl jeder beliebige Testleiter, der einen bestimmten Test mit einer bestimmten Testperson durchführt, als auch jeder beliebige Testauswerter die Testleistung der Testperson genau gleich auswertet und interpretiert.

Sinnvollerweise wird das Gütekriterium der Objektivität in drei Aspekte differenziert (z.B. Lienert & Raatz, 1998), nämlich in die *Durchführungs-, Auswertungs- und Interpretationsobjektivität*:

2.1.1 Durchführungsjektivität

Durchführungsjektivität liegt vor, wenn das Testergebnis nicht davon abhängt, welcher Testleiter den Test mit der Testperson durchführt. Die Wahrscheinlichkeit einer hohen Durchführungsjektivität wird größer, wenn der Test standardisiert ist, d.h. wenn die Durchführungsbedingungen nicht von Untersuchung zu Untersuchung variieren, sondern von den Testautoren bzw. Herausgebern eines Tests festgelegt sind. Zu diesem Zweck werden im Testmanual genaue Anweisungen gegeben. Sie erstrecken sich auf das Testmaterial, etwaige Zeitbegrenzungen und die Instruktion (das ist jener Teil, in dem den Testpersonen - mündlich oder schriftlich - erklärt wird, was sie im Test zu tun haben, einschließlich der Bearbeitung etwaiger Probebeispiele). Es muss auch angegeben werden, ob und wie etwaige Fragen der Testpersonen zum Test behandelt werden sollen. Normalerweise verweist man auf die Instruktion, weshalb dort alles Wesentliche enthalten sein sollte.

Die Standardisierung eines Tests ist dann optimal, wenn die Testperson in der Testsituation die einzige Variationsquelle darstellt, alle anderen Bedingungen hingegen konstant oder kontrolliert sind, so dass sie nicht als Störvariablen wirken können. Die »Testleistung« soll also nur von der Merkmalsausprägung des Individuums abhängen. Es sind in bedeutsamem Maße Variablen bekannt (z.B. Versuchsleitereffekte in Form von »verbal conditioning« in Einzelversuchen), die als Bestandteil der Testsituation die Testleistung in unkontrollierter Weise beeinflussen (vgl. z.B. Rosenthal & Rosnow, 1969); sie können die interne Validität gefährden und zu Artefakten führen (vgl. Sarris & Reiß, 2005). Aus diesem Grunde wird oftmals so weit wie möglich auf eine über die Instruktion hinausgehende Interaktion zwischen Testleiter und Testperson verzichtet; nicht zuletzt deshalb ist eine computerbasierte Testdurchführung der Durchführungsjektivität förderlich. ■ Beispiel 2.1 veranschaulicht die Auswirkungen unterschiedlicher Instruktionen bei einem Leistungstest.

Kontrollierte Durchführungsbedingungen

Standardisierung

Beispiel 2.1

»Auswirkung der Instruktion bei einem Leistungstest«

Als Beispiel möge ein weit verbreiteter Konzentrationstest, das *Frankfurter Aufmerksamkeitsinventar* (FAIR; Moosbrugger & Oehlschlägel, 1996), dienen. Wenn man sich vorstellt, der Testleiter würde sagen, dass es darauf ankommt, in einem Fall **»möglichst ohne Fehler, aber so schnell Sie können«** zu arbeiten und in einem anderen Fall nur **»so schnell Sie können«** zu arbeiten, wird offensichtlich, dass das Testergebnis bedeutsam von der Instruktion beeinflusst werden kann.

2.1.2 Auswertungsobjektivität

Auswertungsobjektivität ist dann gegeben, wenn bei vorliegendem Testprotokoll (Antworten der Testpersonen auf die Testitems) das Testergebnis nicht von der Worten des Testauswerters abhängt. Bei Tests mit Multiple-Choice-

Übereinstimmung verschiedener Testaus- werter

Aufgaben (Mehrfachwahlaufgaben) ist Auswertungsobjektivität im Allgemeinen problemlos zu erreichen. Wenn hingegen ein offenes Antwortformat verwendet wird, müssen detaillierte Auswertungsregeln vorliegen, deren einheitliche Anwendung empirisch überprüft werden muss (■ Beispiel 2.2).

Das Ausmaß der Auswertungsobjektivität lässt sich messbar angeben im Grad der Übereinstimmung, die von verschiedenen Testauswertern bei der Auswertung einer bestimmten Testleistung erreicht wird. Ein Test ist umso auswertungsobjektiver, je einheitlicher die Auswertungsregeln von verschiedenen Testauswertern angewendet werden. Eine statistische Kennzahl der Auswerterübereinstimmung kann z.B. in Form des »Konkordanzkoeffizienten W « nach Kendall (1962) berechnet werden. (Für weitere Übereinstimmungsmaße sei im Überblick auf Wirtz & Caspar (2002) verwiesen.)

Beispiel 2.2

»Auswertungsobjektivität bei einem Intelligenztest«

Es ergeben sich beispielsweise Schwierigkeiten bei der Auswertung einer Intelligenztest-Aufgabe zum *Finden von Gemeinsamkeiten*, wenn für eine eher »schwache« Antwort nur ein Punkt, für eine »gute« Antwort hingegen zwei Punkte gegeben werden sollen. Nennt eine Testperson als Gemeinsames für das Begriffspaar »Apfelsine – Banane« beispielsweise »Nahrungsmittel«, eine andere hingegen »Früchte«, so muss der Test klare Anweisungen im Manual dafür enthalten, welche Antwort höher bewertet werden soll als die andere, um Auswertungsobjektivität zu gewährleisten.

Im Falle des HAWIE-R (Tewes, 1991) sind klare Anweisungen im Manual enthalten.

2.1.3 Interpretationsobjektivität

Regeln für die Test- interpretation

Die Standardisierung eines Tests umfasst über die Durchführungs- und Auswertungsvorschriften hinausgehend klare Regeln für die Testinterpretation. Interpretationsobjektivität liegt dann vor, wenn verschiedene Testanwender bei Testpersonen mit demselben Testwert zu denselben Schlussfolgerungen kommen. Hier kann der Testautor im Testmanual Hilfestellungen geben, indem er durch ausführliche Angaben von Ergebnissen aus der sog. Eichstichprobe (Normentabellen) den Vergleich der Testperson mit relevanten Bezugsgruppen ermöglicht (vgl. Goldhammer & Hartig, 2007, ► Kap. 8 in diesem Band).

Zusammenfassend kann man sagen, dass das Gütekriterium Objektivität dann erfüllt ist, wenn das Testverfahren, bestehend aus Testunterlagen, Testdarbietung, Testauswertung und Testinterpretation so genau festgelegt ist, dass der Test unabhängig von Ort, Zeit und Testleiter und Auswerter durchgeführt werden könnte und für eine bestimmte Testperson bzgl. des untersuchten Merkmals dennoch dasselbe Ergebnis zeigen würde.

2.2 Reliabilität

Das Gütekriterium der Reliabilität betrifft die Messgenauigkeit des Tests und ist wie folgt definiert:

Definition

Ein Test ist dann reliabel (zuverlässig), wenn er das Merkmal, das er misst, exakt, d.h. ohne Messfehler, misst.

Messgenauigkeit des Tests

Das Ausmaß der Reliabilität eines Tests wird über den sog. Reliabilitätskoeffizienten erfasst, der einen Wert zwischen Null und Eins annehmen kann ($0 \leq \text{Rel.} \leq 1$) (vgl. Schermelleh-Engel & Werner, 2007, ► Kap. 6 in diesem Band). Ein Reliabilitätskoeffizient von Eins bezeichnet das Freisein von Messfehlern. Eine völlige Reliabilität würde sich bei einer Wiederholung der Testung an derselben Testperson unter gleichen Bedingungen und ohne Merkmalsveränderung darin äußern, dass der Test zweimal zu dem gleichen Ergebnis führt. Ein Reliabilitätskoeffizient von Null hingegen zeigt an, dass das Testergebnis ausschließlich durch Messfehler zustande gekommen ist. Der Reliabilitätskoeffizient eines guten Tests sollte 0.7 nicht unterschreiten.

Formal ist die Reliabilität definiert als der Anteil der wahren Varianz an der Gesamtvarianz der Testwerte (vgl. Moosbrugger, 2007a, ► Kap. 5 in diesem Band). Die wahre Varianz bemisst dabei die Merkmalsstreuung der »wahren« Testwerte. Der verbleibende Anteil an der Gesamtvarianz der beobachteten Testwerte kommt aufgrund des Messfehlers zustande und repräsentiert damit die »Unreliabilität« oder Messfehlerbehaftetheit eines Messinstrumentes.

■ Beispiel 2.3 hebt die Bedeutung eines reliablen Messinstrumentes hervor.

Beispiel 2.3

»Die Auswirkung von Messfehlern«

Als Beispiel für ein reliables Messinstrument soll in Analogie der Meterstab betrachtet werden. Mit diesem Messinstrument lassen sich Längen sehr genau bestimmen, z.B. die Körpergröße einer Person.

Nun stelle man sich vor, ein »Maßband« sei nicht aus einem längenbeständigen Material, sondern aus einem Gummiband beschaffen. Es ist offensichtlich, dass ein solches Maßband etwa bei einem Schneider zu äußerst unzufriedenen Kunden führen würde, die etwa über zu lange Hosen oder zu weite Blusen klagen müssten, wenn das Maßband bei der Messung zufällig gedehnt worden wäre.

In Übertragung z.B. auf die Intelligenzdiagnostik zur Identifizierung von Hochbegabungen ($\text{IQ} > 130$) resultieren bei mangelnder Reliabilität viele Fehlurteile, weil die Intelligenz je nach Größe und Vorzeichen des Messfehlers häufig über- oder unterschätzt würde.

Um das Ausmaß der Reliabilität zu bestimmen, wurden im Rahmen der Klassischen Testtheorie mehrere Verfahren entwickelt. So unterscheidet man vier Vorgehensweisen (vgl. Moosbrugger & Rauch, 2004):

1. Retest-Reliabilität
2. Paralleltest-Reliabilität
3. Testhalbierungs-Reliabilität
4. Innere Konsistenz

2.2.1 Retest-Reliabilität

Um die Reliabilität nach dem Retest-Verfahren zu bestimmen, wird ein und derselbe Test (unter der idealen Annahme, dass sich das zu messende Merkmal selbst nicht verändert hat) zu zwei verschiedenen Zeitpunkten vorgelegt. Die Reliabilität wird dann als Korrelation zwischen den beiden Testergebnissen ermittelt.

Bei der Retest-Reliabilität ist zu beachten, dass die ermittelte Korrelation in Abhängigkeit vom Zeitintervall zwischen beiden Testungen variieren kann. Je nach Zeitabstand ist nämlich eine Vielzahl von Einflüssen auf die Messungen denkbar, die sich reliabilitätsverändernd auswirken können, insbesondere Übungs- und Erinnerungseffekte oder ein sich tatsächlich veränderndes Persönlichkeitsmerkmal. Veränderungen der gemessenen Testwerte über die zwei Situationen hinweg können als »Spezifität« mittels der sog. Latent-State-Trait-Modelle (Steyer, 1987) explizit identifiziert und berücksichtigt werden (vgl. Kelava & Schermelleh-Engel, 2007, ► Kap. 15 in diesem Band).

2.2.2 Paralleltest-Reliabilität

Etliche reliabilitätsverändernde Einflüsse (z.B. Übungs- und Erinnerungseffekte, aber auch Merkmalsveränderungen) können eliminiert bzw. kontrolliert werden, wenn die Reliabilität nach dem Paralleltest-Verfahren bestimmt wird. Dieses Verfahren wird oftmals als »Königsweg« der Reliabilitätsbestimmung bezeichnet. Hierfür wird die Korrelation zwischen den beobachteten Testwerten in zwei »parallelen Testformen« berechnet, die aus inhaltlich möglichst ähnlichen Items (sog. »Itemzwillingen«) bestehen.

Parallel sind zwei Testformen dann, wenn sie trotz nicht identischer Itemstichproben zu gleichen Mittelwerten und Varianzen der Testwerte führen.

2.2.3 Testhalbierungs-Reliabilität

Oftmals ist es nicht möglich, einen Test zu wiederholen oder parallele Testformen herzustellen (sei es, dass die Testpersonen zu einem zweiten Termin nicht zur Verfügung stehen, dass die Verzerrungen durch eine Wiederholung zu hoch wären, oder dass ein Itempool nicht groß genug ist, um zwei parallele Testformen herzustellen). In solchen Fällen ist es angebracht, den Test in zwei möglichst parallele Testhälften zu teilen und die »Testhalbierungs-Reliabilität« (Split-Half-Reliabilität) als Korrelation der beiden Testhälften zu bestimmen. Gewöhnlich wird allerdings ein Korrekturfaktor berücksichtigt, um die verminderte Split-Half-Reliabilität wieder auf die ursprüngliche Test-

2.3 · Validität

länge hochzurechnen. Die Korrektur führt zu einer Aufwertung der Reliabilität (z.B. Spearman-Brown-Formel; Gleichung 6.5; vgl. Schermelleh-Engel & Werner, 2007, ► Kap. 6 in diesem Band).

2.2.4 Innere Konsistenz

Die Konsistenzanalyse stellt eine Verallgemeinerung der Testhalbierungsmethode in der Weise dar, dass jedes Item eines Tests als eigenständiger Testteil betrachtet wird. Je stärker die Testteile untereinander positiv korrelieren, desto höher ist die interne Konsistenz des Verfahrens (Cronbach- α -Koeffizient der Reliabilität; Cronbach, 1951; vgl. Moosbrugger & Hartig, 2003, S. 412).

Auf die systematische Herleitung der Reliabilitätsmaße aus der Klassischen Testtheorie und auf ihre Berechnung wird von Schermelleh-Engel und Werner (2007; ► Kap. 6 in diesem Band) näher eingegangen.

► Hinweis

Während bei Tests, die nach der Klassischen Testtheorie (KTT; vgl. Moosbrugger, 2007a, ► Kap. 5 in diesem Band) konstruiert wurden, der Reliabilitätskoeffizient eine pauschale Genauigkeitsbeurteilung der Testwerte ermöglicht (s. Konfidenzintervalle; vgl. Moosbrugger, 2007a, ► Kap. 5 in diesem Band), ist bei Tests, die nach der Item-Response-Theorie (IRT; vgl. Moosbrugger, 2007b, ► Kap. 10 in diesem Band) konstruiert worden sind, darüber hinaus eine speziellere, testwertabhängige Genauigkeitsbeurteilung der Testwerte mit Hilfe der »Informationsfunktion« der verwendeten Testitems möglich.

2.3 Validität

Das Gütekriterium der Validität befasst sich mit der Übereinstimmung zwischen dem Merkmal, das man messen will, und dem tatsächlich gemessenen Merkmal. Die Validität (Gültigkeit) wird wie folgt definiert:

Definition

Ein Test gilt dann als valide (»gültig«), wenn er das Merkmal, das er messen soll, auch wirklich misst und nicht irgendein anderes.

Bei der Validität (vgl. Hartig, Frey & Jude, 2007, ► Kap. 7 in diesem Band) handelt es sich hinsichtlich der Testpraxis um das wichtigste Gütekriterium überhaupt. Die Gütekriterien Objektivität und Reliabilität ermöglichen eine hohe Messgenauigkeit, liefern aber nur die günstigen Voraussetzungen für das Erreichen einer hohen Validität, da ein Test, der eine niedrige Reliabilität aufweist, keine hohe Validität haben kann.

Liegt eine hohe Validität vor, so erlauben die Ergebnisse eines Tests die Generalisierung des in der Testsituation beobachteten Verhaltens auf das zu messende Verhalten außerhalb der Testsituation. Formal könnte man daher die Validität eines Tests als Korrelation der Testwerte in der Testsituation mit einem

Generalisierung auf beobachtbares Verhalten außerhalb der Testsituation

korrespondierenden Verhalten außerhalb der Testsituation (Kriterium) definieren. Bei Vorliegen eines bestimmten zu messenden Kriteriums ist diese Form der Validität leicht angebar. Anwendungspraktisch wird man die Validität eines Tests nicht nur mit einer einzigen Korrelation ausdrücken können. Vielmehr ist in Abhängigkeit seiner Anwendungsbereiche (Kriterien) eine Fülle verschiedener Validitäten möglich, die in ihrer Gesamtheit darüber Aufschluss geben, inwieweit ein Test das zu Messende misst und nicht etwas anderes.

Bei der Beurteilung der Validität können verschiedene Aspekte herangezogen werden (■ Beispiel 2.4):

Beispiel 2.4

»Validitätsaspekte der Schulreife«

Wenn man beispielsweise die Validität eines Tests für das zu messende Kriterium »Schulreife« beurteilen will, wäre als erster Aspekt zu prüfen, ob Operationalisierungen von Schulreife in Testaufgaben (Items) umgesetzt wurden. Dabei wären insbesondere Items gefragt, die das interessierende Merkmal *inhaltlich repräsentativ* abbilden (*Inhaltsvalidität*). Im Falle der Schulreife wären insbesondere Testaufgaben z.B. für die Fähigkeit mit Zahlenmengen umzugehen, für das Sprachverständnis und für die sprachliche Ausdrucksfähigkeit zu konstruieren.

Die konstruierten Items würden vor allem dann eine hohe Akzeptanz erfahren, wenn sie Verhaltens- und Erlebensweisen überprüfen, die auch dem Laien als für das Merkmal relevant erscheinen. Dies ist dann der Fall, wenn diese Items eine hohe sog. *Augenscheinvalidität* haben. Jedem Laien ist intuitiv einsichtig, dass Schulreife sich auch dadurch kennzeichnet, dass Kinder mit kleinen Zahlenmengen umgehen können müssen etc. Insofern kann man vom bloßen Augenschein her jenen Items, die solche Fähigkeiten erfassen, Validität zusprechen.

Nun ist es aber so, dass das Merkmal »Schulreife« aus verschiedenen Merkmalen besteht; neben kognitiven Fähigkeiten sind auch soziale Kompetenzen sowie motivationale Variablen von Bedeutung. Die Beschaffenheit der verschiedenen Merkmale und die Homogenität der zur Erfassung der einzelnen Merkmale konstruierten Items sowie die Abgrenzung zu anderen Merkmalen werden im Rahmen der sog. *Konstruktvalidität* empirisch untersucht.

Letztlich ist es nicht immer möglich, das Zielmerkmal als Ganzes oder wenigstens Stichproben daraus in einem Test zusammenzustellen. Möchte man die diagnostische Aussagekraft eines Tests (also z.B. eines Schulreifetests) für konkrete Anwendungen beurteilen, so kann man den Zusammenhang zwischen Kriterium (tatsächliche Schulreife z.B. in Form des Lehrerurteils) und Testwert (Schulreifetest) als die Korrelation zwischen beiden berechnen und diese als Maß der *Kriteriumsvalidität* betrachten. Die Kriteriumsvalidität beschreibt, wie gut sich der Test zur Erfassung des zu messenden Kriteriums eignet.

Um ein differenziertes Bild der Gültigkeit eines Tests zu erhalten, untersucht man also sinnvollerweise folgende Validitätsaspekte:

2.3 · Validität

1. Inhaltsvalidität
2. Augenscheinvalidität
3. Konstruktvalidität
4. Kriteriumsvalidität

2.3.1 Inhaltsvalidität

Definition

Unter Inhaltsvalidität versteht man, inwieweit ein Test oder ein Testitem das zu messende Merkmal repräsentativ erfasst.

Man geht dabei von einem Repräsentationsschluss aus, d.h., dass die Testitems eine repräsentative Stichprobe aus dem Itemuniversum darstellen, mit dem das interessierende Merkmal erfasst werden kann. Die Inhaltsvalidität wird in der Regel nicht numerisch anhand eines Maßes bzw. Kennwertes bestimmt, sondern aufgrund »logischer und fachlicher Überlegungen« (vgl. Cronbach & Meehl, 1955; Michel & Conrad, 1982). Dabei spielt die Beurteilung der inhaltlichen Validität durch die Autorität von Experten eine maßgebende Rolle.

Am einfachsten ist die Frage nach der Inhaltsvalidität eines Tests dann zu klären, wenn die einzelnen Items einen unmittelbaren Ausschnitt aus dem Verhaltensbereich darstellen, über den eine Aussage getroffen werden soll (wenn z.B. Rechtschreibkenntnisse anhand eines Diktates überprüft werden oder die Eignung eines Autofahrers anhand einer Fahrprobe ermittelt wird).

Repräsentationsschluss

2.3.2 Augenscheinvalidität

Mit inhaltlicher Validität leicht zu verwechseln (vgl. Tent & Stelzl, 1993) ist die Augenscheinvalidität, da oftmals inhaltlich validen Tests zugleich auch Augenscheinvalidität zugesprochen wird.

Definition

Augenscheinvalidität gibt an, inwieweit der Validitätsanspruch eines Tests, vom bloßen Augenschein her einem Laien gerechtfertigt erscheint.

Vor dem Hintergrund der Mitteilbarkeit der Ergebnisse und der Akzeptanz von Seiten der Testpersonen kommt der Augenscheinvalidität eines Tests eine ganz erhebliche Bedeutung zu. Nicht zuletzt auch wegen der Bekanntheit der Intelligenzforschung haben z.B. Intelligenz-Tests eine hohe Augenscheinvalidität, da Laien aufgrund von Inhalt und Gestaltung des Tests es für glaubwürdig halten, dass damit Intelligenz gemessen werden kann. Aus der wissenschaftlichen Perspektive ist die Augenscheinvalidität allerdings nicht immer zufriedenstellend, denn die Validität eines Tests muss auch empirisch durch Kennwerte belegt werden.

Akzeptanz eines Tests

2.3.3 Konstruktvalidität

Unter dem Aspekt der Konstruktvalidität beschäftigt man sich mit der theoretischen Fundierung des von einem Test tatsächlich gemessenen Merkmals.

Definition

Ein Test weist Konstruktvalidität auf, wenn der Schluss vom Verhalten der Testperson innerhalb der Testsituation auf zugrunde liegende psychologische Persönlichkeitsmerkmale (»Konstrukte«, »latente Variablen«, »Traits«) wie Fähigkeiten, Dispositionen, Charakterzüge, Einstellungen aufgezeigt wurde. Die Enge dieser Beziehung wird aufgrund von testtheoretischen Annahmen und Modellen überprüft.

Gemeint ist, ob z.B. von den Testaufgaben eines »Intelligenztests« wirklich auf die Ausprägung einer latenten Persönlichkeitsvariablen »Intelligenz« geschlossen werden kann oder ob die Aufgaben eigentlich ein anderes Konstrukt (etwa »Gewissenhaftigkeit« anstelle des Konstruktes »Intelligenz«) messen.

Bei der Beurteilung der Konstruktvalidität sind prinzipiell struktursuchende und strukturprüfende Ansätze zu unterscheiden.

Der erste Ansatz basiert auf einer **struktursuchenden deskriptiven Vorgehensweise**:

- Zur Gewinnung von Hypothesen über die ein- bzw. mehrdimensionale Merkmalsstruktur der Testitems werden sog. Exploratorische Faktorenanalysen (EFA) zum Einsatz gebracht (vgl. Moosbrugger & Schermelleh-Engel, 2007, ► Kap. 13 in diesem Band).
- Innerhalb der einzelnen Merkmale geben die Faktorladungen analog der Trennschärfekoeffizienten einer Itemanalyse Auskunft über die Homogenität der Testitems (vgl. Kelava & Moosbrugger, 2007, ► Kap. 4 in diesem Band).
- Die solchermaßen gewonnenen Merkmalsdimensionen erlauben eine erste deskriptive Einordnung in ein bestehendes theoretisches Gefüge theoretischer Konstrukte. Dabei kann z.B. die Bildung eines »nomologischen Netzwerkes« nützlich sein (vgl. Hartig, Frey & Jude, 2007, ► Kap. 7 in diesem Band).

Bei der Bildung eines »nomologischen Netzwerkes« steht die Betrachtung theoriekonformer Zusammenhänge zu anderen Tests im Vordergrund. Dazu formuliert man a priori theoriegeleitete Erwartungen über den Zusammenhang des vorliegenden Tests bzw. des/der von ihm erfassten Merkmals/-e mit konstruktverwandten und konstruktfernen bereits bestehenden Tests. Danach wird der vorliegende Test empirisch mit den anderen Tests hinsichtlich Ähnlichkeit bzw. Unähnlichkeit verglichen, wobei zwischen konvergenter Validität und diskriminanter/divergenter Validität unterschieden wird (vgl. Schermelleh-Engel & Schweizer, 2007, ► Kap. 14 in diesem Band):

Struktursuchendes Vorgehen

Konvergente Validität

Um zu zeigen, dass ein Test das zu messende Merkmal misst und nicht irgendein anderes, kann die Übereinstimmung mit Ergebnissen aus Tests für gleiche oder ähnliche Merkmale ermittelt werden. So soll z.B. die Korrelation eines neuartigen Intelligenztests mit einem etablierten Test, wie etwa dem HAWIE-R (Tewes, 1991), zu einer hohen Korrelation führen, um zu zeigen, dass auch der neue Test das Konstrukt »Intelligenz« misst.

Diskriminante bzw. divergente Validität

Um zu zeigen, dass ein Test das zu messende Merkmal misst und nicht eigentlich ein anderes, muss er von Tests für andere Merkmale abgrenzbar sein. So soll ein Konzentrationsleistungstest ein diskriminierbares eigenständiges Konstrukt, nämlich »Konzentration«, erfassen und nicht das Gleiche wie andere Tests für andere Konstrukte. Wünschenswert sind deshalb niedrige korrelative Zusammenhänge zwischen Konzentrationstests und Tests für andere Variablen. Zum Nachweis der diskriminanten Validität ist es nicht hinreichend, dass der zu validierende Test nur mit irgendwelchen offensichtlich konstruktfernen Tests verglichen wird, sondern dass er auch zu relativ konstruktnahen Tests in Beziehung gesetzt wird. So wäre z.B. eine niedrige Korrelation zwischen Konzentration und Intelligenz wünschenswert (so z.B. FAKT-II, Moosbrugger & Goldhammer, 2007).

Der zweite Ansatz erlaubt es anhand einer **strukturprüfenden Vorgehensweise** inferenzstatistische Schlüsse bzgl. der Konstruktvalidität zu ziehen. Dies ist allerdings nur auf der Basis von Testmodellen mit latenten Variablen möglich (insbesondere anhand von IRT-Modellen und latenten Strukturgleichungsmodellen), welche eine explizite und inferenzstatistisch überprüfbare Beziehung zwischen zuvor genau definierten, latenten Merkmalen (bspw. Intelligenz) und den manifesten Itemvariablen (bspw. Testitems) herstellen:

- Die in exploratorischen Faktorenanalysen gefundene Struktur kann an neuen Datensätzen mit Konfirmatorischen Faktorenanalysen (CFA) überprüft werden (Jöreskog & Sörbom, 1996; vgl. auch Moosbrugger & Schermelleh-Engel, 2007, ► Kap. 13 in diesem Band).
- Die einzelnen Dimensionen können mit Hilfe von IRT-Modellen konfirmatorisch bezüglich der Homogenität der Testitems eines Tests inferenzstatistisch überprüft werden (vgl. Moosbrugger, 2007b, ► Kap. 10 in diesem Band).
- Eine weitere konfirmatorische Vorgehensweise der Konstruktvalidierung ermöglichen Multitrait-Multimethod-Analysen im Rahmen latenter Strukturgleichungsmodelle (vgl. Eid, 2000; Schermelleh-Engel & Schweizer, 2007, ► Kap. 14 in diesem Band). Dabei wird der Zusammenhang zwischen verschiedenen Merkmalen (traits) unter Herauspriorisierung der Methodeneinflüsse strukturprüfend untersucht.

Auf die beiden Ansätze zur Überprüfung der Konstruktvalidität soll in späteren Abschnitten ausführlich eingegangen werden.

Strukturprüfendes Vorgehen

Praktische Anwendbarkeit eines Tests für die Vorhersage

Zeitliche Verfügbarkeit des Kriteriums

2.3.4 Kriteriumsvalidität

Die Kriteriumsvalidität bezieht sich auf die praktische Anwendbarkeit eines Tests für die Vorhersage von Verhalten und Erleben.

Definition

Ein Test weist Kriteriumsvalidität auf, wenn vom Verhalten der Testperson innerhalb der Testsituation erfolgreich auf ein »Kriterium«, nämlich auf ein Verhalten außerhalb der Testsituation, geschlossen werden kann. Die Enge dieser Beziehung ist das Ausmaß an Kriteriumsvalidität (Korrelationschluss).

Kriteriumsvalidität liegt z.B. bei einem »Schulreifetest« vor allem dann vor, wenn jene Kinder, die im Test leistungsfähig sind, sich auch in der Schule als leistungsfähig erweisen und umgekehrt, wenn jene Kinder, die im Test leistungsschwach sind, sich auch in der Schule als leistungsschwach erweisen. Die Überprüfung der Kriteriumsvalidität ist im Prinzip an keine bestimmten testtheoretischen Annahmen gebunden und erfolgt i. d. R. durch Bestimmung der Korrelation zwischen der Testvariablen und der Kriteriumsvariablen.

Abhängig von der zeitlichen Verfügbarkeit des Außenkriteriums, nämlich ob es bereits in der Gegenwart oder erst in der Zukunft vorliegt, spricht man von *Übereinstimmungsvalidität* (sog. konkurrender Validität) oder von *Vorhersagevalidität* (prognostischer Validität). Im ersten Fall ist also der Zusammenhang eines Testwertes mit einem Kriterium von Interesse, das zeitgleich »existiert«, im zweiten Fall steht die Prognose einer »zukünftigen« Ausprägung eines Merkmals im Vordergrund.

2.4 Skalierung

Das Gütekriterium der Skalierung betrifft bei Leistungstests vor allem die Forderung, dass eine leistungsfähigere Testperson einen besseren Testwert als eine weniger leistungsfähige erhalten muss, d.h., dass sich also die Relation der Leistungsfähigkeit auch in den Testwerten widerspiegelt. Die Forderung der Skalierung bezieht sich sowohl auf interindividuelle Differenzen als auch auf intraindividuelle Differenzen und in analoger Form auch auf Persönlichkeitstests.

Definition

Ein Test erfüllt das Gütekriterium der Skalierung, wenn die laut Verrechnungsregel resultierenden Testwerte die empirischen Merkmalsrelationen adäquat abbilden.

Skalenniveau

Die Umsetzbarkeit dieses Gütekriteriums hängt insbesondere vom Skalenniveau des Messinstrumentes ab. In der Regel reicht eine Messung des Merkmals auf Nominalskalenniveau nicht aus, um die größer/kleiner Relation

2.5 · Normierung (Eichung)

zwischen den Testpersonen zu beschreiben. Damit eine leistungsfähigere Testperson einen besseren Testwert als eine leistungsschwächere erhält, muss zumindest eine Messung auf Ordinalskalenniveau erfolgen. Eine Messung auf Intervallskalenniveau erlaubt darüber hinaus eine Beurteilung der Größe inter- und intraindividuelle Differenzen. Verhältnisse zwischen Testleistungen können nur auf Rationalskalenniveau bestimmt werden; dieses wird in der Psychologie nur selten erreicht.

Während man sich im Rahmen der »Klassischen Testtheorie« (vgl. Moosbrugger, 2007a, ► Kap. 5 in diesem Band) damit zufrieden geben muss, z.B. die Anzahl der gelösten Aufgaben zu einem Testwert zu verrechnen, ist im Rahmen der »Item-Response-Theorie« das Gütekriterium der Skalierung empirisch überprüfbar, indem untersucht wird, ob das Verhalten aller Testpersonen einem ganz bestimmten mathematischen Modell folgt (vgl. Moosbrugger, 2007b, ► Kap. 10 in diesem Band).

2.5 Normierung (Eichung)

Der Zweck der Normierung eines Verfahrens besteht darin, möglichst aussagekräftige »Vergleichswerte« von solchen Personen zu erhalten, die der Testperson hinsichtlich relevanter Merkmale (z.B. Alter, Geschlecht, Schulbildung) ähnlich sind (»Eichstichprobe«).

Definition

Unter der Normierung (Eichung) eines Tests versteht man das Erstellen eines Bezugssystems, mit dessen Hilfe die Ergebnisse einer Testperson im Vergleich zu den Merkmalsausprägungen anderer Personen eindeutig eingeordnet und interpretiert werden können.

Man dokumentiert die Ergebnisse der Testeichung in Form sog. »Normtabellen«, wobei die Eichstichprobe aus einer möglichst großen und repräsentativen Stichprobe bestehen soll. Die Testergebnisse der untersuchten Person werden dann bei der normorientierten Beurteilung in Relation zu den Testergebnissen von Personen aus der Eichstichprobe interpretiert (vgl. Goldhammer & Hartig, 2007, ► Kap. 8 in diesem Band).

Bei der Relativierung eines Testergebnisses an der Eichstichprobe ist es am anschaulichsten, wenn der Prozentsatz derjenigen Personen bestimmt wird, die im Test besser bzw. schlechter abschneiden als die Referenztestleistung in der Eichstichprobe. Aus diesem Grunde wird als Normwert auch der *Prozentrang* der Testwerte in der Eichstichprobe verwendet. Er kumuliert die in der Eichstichprobe erzielten prozentualen Häufigkeiten der Testwerte bis einschließlich zu jenem Testwert, den die gerade interessierende Testperson erzielte.

Weitere Normierungstechniken, die zur Relativierung eines Testergebnisses herangezogen werden, beziehen sich in der Regel auf den Abstand des individuellen Testwertes x_i vom Mittelwert in der entsprechenden Eichstichprobe \bar{x} und drücken die resultierende Differenz in Einheiten der Standardabweichung SD der Verteilung aus (Standardwerte: $z_i = \frac{x_i - \bar{x}}{SD}$). Bekannt sind

Eichstichprobe

Prozentrangnormen

Standardnormen

Geltungsbereich der Normtabelle

und verwendet werden darüber hinaus folgende Normwerte, die auf den Standardwerten aufbauen, z.B. *IQ-Werte*, *T-Werte*, *Centil-Werte*, *Stanine-Werte*, *Standardschulnoten*. Auf diese Normwerte wird von Goldhammer und Hartig (2007; ► Kap. 8 in diesem Band) näher eingegangen.

Bei der Interpretation der Normwerte ist zu berücksichtigen, ob das Merkmal in der Population normalverteilt ist. Andernfalls sind lediglich Prozentrangwerte zur Interpretation heranziehbar, da diese nicht verteilungsgebunden sind. Nichtnormalverteilte Merkmale können durch eine »Flächentransformation« normalisiert werden (vgl. Lienert & Raatz, 1998; McCall, 1939; s. Kelava & Moosbrugger, 2007, ► Abschn. 4.8 in diesem Band).

Bei einer Normierung ist darüber hinaus der Geltungsbereich der Normtabellen eines Tests klar zu definieren. D.h., die für die Normierung erhobene Vergleichsstichprobe (Eichstichprobe) muss repräsentativ für die Grundgesamtheit von Personen sein, für die der Test prinzipiell anwendbar sein soll.

Um eine angemessene Vergleichbarkeit der Personen zu ermöglichen, muss gewährleistet sein, dass die Normtabellen nicht veraltet sind. So sieht bspw. die DIN 33430 (Westhoff et al., 2005) bei Verfahren bzw. Tests zur berufsbezogenen Eignungsbeurteilung vor, dass spätestens nach 8 Jahren die Gültigkeit der Eichwerte zu überprüfen ist und ggf. eine Neunormierung vorgenommen werden muss.

Wesentliche Gründe für die Notwendigkeit von Neunormierungen können z.B. Lerneffekte in der Population (insbesondere in Form einer Bekanntheit des Testmaterials) oder auch im Durchschnitt tatsächlich veränderte Merkmale in der Population sein, wie das nachfolgende ■ Beispiel 2.5 darstellen soll, das eine Verringerung der Testleistung in der Population beschreibt:

Beispiel 2.5

»Normenverschiebung im AID vs. AID2« (entnommen aus Kubinger & Jäger, 2003)

In Bezug auf den AID aus dem Jahre 1985 und den AID 2 aus dem Jahre 2000 zeigte sich eine Normenverschiebung im Untertest »Unmittelbares Reproduzieren-numerisch« (Kubinger, 2001): Die Anzahl der in einer Folge richtig reproduzierten Zahlen (z.B.: 8-1-9-6-2-5) lag im Jahr 2000 im Vergleich zu früher, vor ca. 15 Jahren, über das Alter hinweg fast durchwegs um 1 niedriger. Waren es 1985 bei den 7- bis 8- bzw. 9- bis 10-Jährigen noch 5 bzw. 6 Zahlen, die durchschnittlich in einer Folge reproduziert werden konnten, so waren es im Jahr 2000 nunmehr 4 bzw. 5 Zahlen. Ein Nichtberücksichtigen dieses Umstandes würde bedeuten, dass Kinder in ihrer Leistungsfähigkeit im Vergleich zur altersgemäßen Durchschnittsleistung wesentlich unterschätzt würden.

2.6 Testökonomie

Wirtschaftlichkeit eines Tests

Die Ökonomie bezieht sich auf die Wirtschaftlichkeit eines Tests und wird durch die Kosten bestimmt, die bei einer Testung entstehen. I. d. R. stimmen die Interessen von Testpersonen, Auftraggebern und Testleitern in dem

Wunsch überein, keinen überhöhten Aufwand zu betreiben. Dennoch lassen sich oftmals die Kosten nicht beliebig minimieren, ohne dass andere Gütekriterien (etwa Objektivität und Reliabilität) darunter leiden.

Definition

Ein Test erfüllt das Gütekriterium der Ökonomie, wenn er, gemessen am diagnostischen Erkenntnisgewinn, relativ wenig Ressourcen wie Zeit, Geld oder andere Formen beansprucht.

Im Wesentlichen beeinflussen zwei Faktoren die Ökonomie bzw. die Kosten eines Tests, nämlich der finanzielle Aufwand für das Testmaterial und der zeitliche Aufwand für die Testdurchführung.

Der bei einer Testung entstehende *finanzielle Aufwand* kann sich vor allem aus dem Verbrauch des Testmaterials ergeben oder aus der Beschaffung des Tests selbst. Zudem kann bei computergestützten Tests die Beschaffung aufwändiger Computerhardware und -software einen wesentlichen Kostenfaktor darstellen. Nicht zu vergessen sind anfallende Lizenzgebühren für Testautoren und Verlage, die mit den Beschaffungskosten des Materials einhergehen.

Das zweite Merkmal der Ökonomie, nämlich der *zeitliche Aufwand*, bildet oftmals einen gewichtigeren Faktor als die Testkosten alleine. Die Testzeit umfasst nicht nur die Nettozeit der Bearbeitung des Tests, durch die sowohl den Testpersonen als auch dem Testleiter Kosten entstehen, sondern auch die Zeit der Vorbereitung, der Auswertung und Ergebnismeldung.

Sinnvollerweise kann man also sagen, dass der Erkenntnisgewinn aus dem Einsatz eines Tests größer sein muss als die entstehenden Kosten. Die Ökonomie in diesem Sinne ist oft nur im Vergleich mit ähnlichen Tests bestimmbar. Vor allem Tests, die am Computer vorgegeben werden können, erfüllen dieses Kriterium vergleichsweise leichter. Einen wichtigen Beitrag zur ökonomischeren Erkenntnisgewinnung kann auch durch das Adaptive Testen (vgl. Frey, 2007, ► Kap. 11 in diesem Band) geleistet werden, bei dem nur jene Aufgaben von der Testperson zu bearbeiten sind, die für sie den größten Informationsgewinn mit sich bringen.

Eine höhere Wirtschaftlichkeit darf natürlich nicht zu Lasten der anderen Gütekriterien im Vordergrund stehen. So ist eine geringere Ökonomie eines Tests bei einer konkreten Fragestellung insbesondere dann in Kauf zu nehmen, wenn z.B. aus Validitätsgründen der Einsatz gerade dieses Tests sachlich gerechtfertigt ist, weil nur mit ihm die konkrete Fragestellung fachgerecht beantwortbar ist.

Finanzieller und zeitlicher Aufwand

2.7 Nützlichkeit

Definition

Ein Test ist dann nützlich, wenn für das von ihm gemessene Merkmal praktische Relevanz besteht und die auf seiner Grundlage getroffenen Entscheidungen (Maßnahmen) mehr Nutzen als Schaden erwarten lassen.

Praktische Relevanz eines Tests

2

Für einen Test besteht dann praktische Relevanz, wenn er erstens ein Merkmal misst, das im Sinne der Kriteriumsvalidität nützliche Anwendungsmöglichkeiten aufweist, und zweitens dieses Merkmal nicht auch mit einem anderen Test erfasst werden könnte, der alle übrigen Gütekriterien mindestens genauso gut erfüllt. Das Kriterium der Nützlichkeit wird am nachfolgenden

■ Beispiel 2.6 veranschaulicht.

Beispiel 2.6

»Nützlichkeit des Tests für medizinische Studiengänge (TMS)«

Die Konstruktion eines Tests zur Studieneignungsprüfung für ein medizinisches Studium (TMS, Institut für Test- und Begabungsforschung, 1988) erfüllte seinerzeit das Kriterium der Nützlichkeit. Da ein Bedarf der korrekten Selektion und Platzierung der potentiellen Medizinstudenten angesichts der Kosten, die mit einem Studium eines medizinischen Faches verbunden sind, bestand, konstruierte man in den 1970er Jahren einen Test, der das komplexe Merkmal »Studieneignung für medizinische Studiengänge« erfassen und eine Vorhersage bezüglich des Erfolgs der ärztlichen Vorprüfung ermöglichen sollte (Trost, 1994). Da es zu diesem Zeitpunkt keinen anderen Test gab, der dies in ähnlicher Form in deutscher Sprache zu leisten vermochte. Der Nutzen des TMS wurde anhand aufwändiger Begleituntersuchungen laufend überprüft. (Der TMS wurde 1996 aus wirtschaftlichen Gründen wieder abgeschafft.)

2.8 Zumutbarkeit

Definition

Ein Test erfüllt das Kriterium der Zumutbarkeit, wenn er absolut und relativ zu dem aus seiner Anwendung resultierenden Nutzen die zu testende Person in zeitlicher, psychischer sowie körperlicher Hinsicht nicht über Gebühr belastet.

Zeitliche, physische und psychische Beanspruchung der Testperson

Psychologische Tests müssen so gestaltet werden, dass die Testpersonen bezüglich des Zeitaufwandes sowie des physischen und psychischen Aufwandes geschont werden. Die Zumutbarkeit eines Tests betrifft dabei ausschließlich die Testpersonen und nicht den Testleiter. Die Frage nach der Beanspruchung des Testleiters ist hingegen eine Frage der Testökonomie.

Im konkreten Fall ist eine verbindliche Unterscheidung zwischen zu- und unzumutbar oft schwierig, da es jeweils um eine kritische Bewertung dessen geht, was unter »Nutzen« zu verstehen ist. Dabei spielen gesellschaftliche Normen der Zumutbarkeit eine wesentliche Rolle. Beispielsweise gilt es als durchaus akzeptabel, einem Anwärter auf den Beruf des Piloten für die Auswahl einen sehr anspruchsvollen und beanspruchenden »Test« zuzumuten. Allerdings würde bei der Auswahl einer Sekretärin ein ähnlich beanspruchendes Verfahren auf weniger Verständnis stoßen.