

Zentrale Aspekte der Testkonstruktion

I n h a l t

1. Die Aufgabenanalyse

Wenn Sie hier anklicken, erhalten Sie Informationen über die Berechnung des Schwierigkeitsindizes und der Trennschärfe. Außerdem erfahren Sie etwas über die Möglichkeiten der Aufgabenselektion und über die Notwendigkeit der Verteilungsanalyse.

2. Bestimmung der Reliabilität

Um die formale Exaktheit eines Testes zu prüfen, bedient man sich des Prinzips der parallelen Messung. Es werden 4 Möglichkeiten gezeigt, die Zuverlässigkeit eines Tests nachzuweisen.

3. Bestimmung der Validität

Bei der Validität wird nicht, wie bei der Reliabilität, nach der formalen Exaktheit eines Tests gefragt, sondern danach, ob der Test inhaltlich dazu berechtigt ist ein bestimmtes Merkmal zu erfassen. Sie erfahren alles über die verschiedenen Formen der Validität.

1. Die Aufgabenanalyse

a. Schwierigkeitsindex

Definition

Der Schwierigkeitsindex gibt an, wie groß der Anteil der Probanden an allen Probanden ist, die ein Item "richtig" beantwortet haben. Ein Item wird als schwierig bezeichnet, wenn es nur von wenigen Probanden bzgl. eines bestimmten Merkmals "richtig" beantwortet wird.

Hinweis:

Eigentlich kann man nicht einfach sagen "richtig" beantworten, weil der Index auch bei Items berechnet wird, deren Antwortmöglichkeiten jenseits von richtig oder falsch liegen. Deshalb werden solche Items, die in Richtung des zu untersuchenden Merkmals beantwortet werden, auch als Items entsprechend der Schlüsselantwort bezeichnet.

Ziel

Der Schwierigkeitsindex soll entsprechend der differentialpsychologischen Perspektive die Frage beantworten, ob ein Test mit seinen Items Probanden mit hoher Merkmalsausprägung, von Probanden mit geringer Merkmalsausprägung zu trennen vermag. Zu einer solchen Unterscheidung sind zwei Gruppen von Items unbrauchbar: erstens solche Items, die von allen Probanden gelöst werden, zweitens solche Items, die von keinem Probanden gelöst werden. Der Schwierigkeitsindex soll Items identifizieren, die brauchbarer sind als diese zwei Gruppen.

Berechnung

Es wird zwischen Schwierigkeitsindizes für zweistufige Antworten und solchen für mehrstufige Antworten unterschieden. Zu berücksichtigen sind außerdem die Besonderheiten von Schnelligkeits- und reinen Leistungstests. Klicken Sie bitte auf die nebenstehenden Begriffe, um weitere Informationen zu erhalten.

Schnelligkeitstest

Für unterschiedliche Tests gibt es verschiedene erwünschte Schwierigkeitsindizes. Bei Schnelligkeitstests sollte die Schwierigkeit der Items nicht zu hoch sein, da nur wenig Zeit zur Lösung zur Verfügung steht. Die Differenzierung zwischen der Merkmalsausprägungen der Versuchspersonen wird nicht dadurch ermittelt, dass ein Item gelöst wurde oder nicht, sondern wie viele Items im Test gelöst wurden.

Beispielhafter Test: D2

Leistungstests

Bei reinen Leistungstests werden die Items nach aufsteigender Schwierigkeit angeordnet. Da es bei dieser Art von Tests keine Zeitbegrenzung gibt, kann das Leistungsniveau der Versuchsperson dadurch festgestellt werden, bis zu welcher Aufgabe im Test sie vordringt. Hier ist also die Zahl der gelösten Items entscheidend.

Beispielhafter Test: MWTB2

b. Trennschärfeanalyse

Definition

Der Trennschärfekoeffizient bringt zum Ausdruck, ob das einzelne Item wie der Gesamttest dazu in der Lage ist, "gute" von "schlechten" Probanden zu unterscheiden. Operational definiert sich somit die Trennschärfe als die Korrelation des Items mit dem Gesamttestwert.

Bedeutung des Koeffizienten

Ein hoher Trennschärfekoeffizient besagt, dass das entsprechende Item "gute" von "schlechten" Probanden deutlich unterscheidet, indem "gute" Probanden das Item meist richtig und "schlechte" Probanden das Item meist falsch beantworten.

Ein Trennschärfekoeffizient um 0 zeigt, dass das Item sowohl von "guten" als auch von "schlechten" Probanden gleichermaßen beantwortet wird. Solche Aufgaben sind nicht brauchbar.

Ein negativer Trennschärfekoeffizient bringt zum Ausdruck, dass das Item von den "guten" Probanden falsch und von den "schlechten" Probanden richtig beantwortet wird. Auf solche Items sollte ebenfalls verzichtet werden, weil es durchaus Schwierigkeiten bereiten könnte, diese Items vor den Probanden zu rechtfertigen.

Berechnung

Die Berechnung der Trennschärfe hängt vom Skalenniveau der Daten ab. Da es sich beim Trennschärfekoeffizienten um ein Korrelationsmaß handelt, kann die Trennschärfe bei dichotomen Itemwerten (bspw. ja-nein Antworten) als punktbiseriale Korrelation berechnet werden. Bei mehrfach gestuften Itemwerten (die mindestens intervallskaliert sind und linear mit dem Gesamtestwert in Beziehung stehen), wird die Produkt-Moment-Korrelation berechnet.

Hinweis:

Klicken Sie bitte auf den Begriff SPSS, um exemplarisch die Berechnung der Trennschärfekoeffizienten mittels Produkt-Moment-Korrelation zu verfolgen.

c. Aufgabenselektion

Rationale Selektion

Mit der Methode der rational-simultanen Selektion werden Trennschärfe und Itemschwierigkeit in Beziehung gesetzt. Hierbei zeigt sich, dass Items mittlerer Schwierigkeit die größte Trennschärfe besitzen und Items mit kleinem oder großem Schwierigkeitsgrad weniger trennscharf sind (parabolischer Verlauf des Punkteschwarms). Über die grafische Darstellung dieser Beziehung, lässt sich anhand bestimmter Regeln entscheiden, welche Items für die Testendform beibehalten werden und welche als ungeeignet wegfallen sollen:

1. Beibehalten werden Items mittlerer Schwierigkeit, deren Trennschärfe am höchsten ist.
2. Unbedingt auszuschneiden sind Items mit fehlender oder negativer Trennschärfe.
3. Items mit geringer Trennschärfe werden nach ihrer Schwierigkeit ausgewählt. Hier werden eher Items mit einer geringen bzw. hohen Schwierigkeit zurückbehalten, um so auch in den Extrembereichen des untersuchten Merkmals noch eine gute Differenzierung zu erzielen.

Hinweis:

Klicken Sie auf den Begriff SPSS, um das Verfahren der rationalen Selektion mit SPSS zu verfolgen.

d. Verteilungsanalyse

Ziel

Im Rahmen der Aufgabenanalyse interessiert es außerdem, ob der Test eine hinreichende Streuung der Punktwerte besitzt und ob die Testpunktwerte annähernd normal verteilt sind. Die Normalverteilung ist zwar keine notwendige Voraussetzung für einen guten Test, sollte aber im Hinblick auf eine spätere Eichung angestrebt werden.

Hinweis:

Klicken Sie bitte auf den Begriff SPSS, um die Berechnung des Kolmogorov-Smirnov Tests zur Überprüfung der Normalverteilungsanalyse zu starten.



2. Bestimmung der Reliabilität

Definition allgemein

Bei der Messung der Reliabilität geht es um die formale Exaktheit der Merkmalerfassung. Es ist der Grad der Genauigkeit, mit der der Test misst, was er faktisch misst, ohne dass darauf Rücksicht genommen wird, ob es auch das ist, was er messen soll (hiernach fragt die Validität).

Definition konkret

Von der Reliabilität eines Tests hängt es ab, ob eine Wiederholung desselben Messvorganges stets die gleichen Resultate erbringt. Es kann aber anhand des ermittelten Reliabilitätskoeffizienten nicht gesagt werden, ob das eingesetzte Testverfahren auch wirklich das zu untersuchende Merkmal erfasst. Ein Test kann somit hoch reliabel (er misst zu allen Testzeitpunkten gleich bzw. ähnlich), aber nur gering valide sein (er misst gar nicht das zu untersuchende Merkmal).

Zeitpunkt

Ideal ist der Zeitpunkt nach Abschluss der Aufgabenanalyse mit der Fertigstellung der Testendform. Möchte man eine Vorschätzung der Reliabilität anstellen, so kann dies schon auf Grundlage der Berechnung der inneren Konsistenz der Analysedaten getan werden.

Stichprobe

Auch für die Reliabilitätskontrolle gilt, dass die Stichprobe repräsentativ für den späteren Geltungsbereich des Testes sein soll. Allerdings sind die Anforderungen nicht so hoch wie bei der Eichstichprobe (bzgl. der Variabilität der Rohwerte). Darum können auch sogenannte anfallende Stichproben - Stichproben, die leicht zugänglich sind; wie etwa Studenten - herangezogen werden. Die Stichprobe sollte in der Regel 200 - 500 Probanden umfassen. Wenn es die Verhältnisse erlauben, kann auch die Analysestichprobe als Kontrollstichprobe herangezogen werden. Voraussetzung dafür ist allerdings, dass sich die Testvorform nicht wesentlich von der Testendform unterscheidet.

Methoden

Grundsätzlich lässt sich der Reliabilitätskoeffizient durch das Konzept der parallelen Messung bestimmen. Eine parallele Messung ist eine Messwiederholung mit demselben Test oder eine Messung mit einem sog. Paralleltest. Die Korrelation zwischen diesen beobachteten Messwerten ergibt die Reliabilität eines Tests. Parallele Messungen liegen demnach vor, wenn man

- a) einen Test bei derselben Stichprobe wiederholt,
- b) einen Paralleltest einsetzt,
- c) einen Test in zwei äquivalente Hälften aufteilt und diese als Paralleltests betrachtet,
- d) jede einzelne Aufgabe eines Tests als "Paralleltest" ansieht.

Hieraus ergeben sich die vier klassischen Verfahren zur Bestimmung eines Reliabilitätskoeffizienten (weitere Informationen erhalten Sie durch einen Klick auf das Feld der "4 Verfahren...").

Testwiederholungsmethode

Allgemeines Vorgehen

Ein Test wird bei einer Stichprobe von Probanden und nach einem gewissen Zeitabstand wiederholt. Die Rohwertepaare aus dem Test und der Testwiederholung werden miteinander korreliert.

Zur Bestimmung des Zeitabstands zwischen den beiden Messzeitpunkten, kann keine eindeutige Aussage gemacht werden. Einmal wird ein kurzes Intervall gefordert, damit das untersuchte Persönlichkeitsmerkmal möglichst unverändert bleibt. Ein anderes mal wird ein langes Intervall gefordert, damit eventuelle Erinnerungsspuren verblassen. Es muss ganz von der Eigenart des Tests abhängig gemacht werden, ob ein kurzes oder langes Zeitintervall für die Testwiederholung bestimmt wird.

Hinweis:

Die Testwiederholungsmethode wird auch als Retest-Methode bezeichnet.

Besonderheit

Da durch psychologische Messung zumeist das untersuchte Merkmal durch den Vorgang der Messung in bestimmter Weise verändert wird (Übungsfortschritt, Einsicht), ist es eigentlich nur bei psychophysischen Messungen/Tests sinnvoll die Methode der Testwiederholung zur Reliabilitätsbestimmung einzusetzen. Allerdings sollten bei der größten Zahl psychologischer Untersuchungen (vor allem Intelligenz- und Leistungstests) Testwiederholungen vermieden werden. In der Praxis werden meist nur reine Schnelligkeitstests über die Retest-Methode auf Reliabilität bestimmt. Tests im Sinne des Niveauekonzepts sollten durch andere Methoden untersucht werden.

Paralleltestmethode

Allgemeines Vorgehen

Zwei Parallelformen werden einer Stichprobe sofort oder in einem gewissen zeitlichen Abstand nacheinander in Zufallsfolge dargeboten. Dabei wird die eine Zufallshälfte der Stichprobe mit Testform A, die andere Hälfte mit der Testform B beginnen. Bei der Wiederholung wird dann entsprechend eines Überkreuzungsplans gewechselt. Zwischen den beiden Testdurchführungen soll höchstens ein Abstand von wenigen Tagen liegen. Nach Möglichkeit ist der Test sofort zu wiederholen. Die erhobenen Rohwertepaare werden miteinander korreliert und ergeben die Reliabilität.

Besonderheit

Die Paralleltest-Methode gilt als die beste Methode, um die Testreliabilität zu bestimmen. Allerdings erweist es sich als äußerst schwierig zwei äquivalente Testformen aufzubauen (bspw. durch den Einmaligkeitscharakter einer Testaufgabe).

Testhalbierungsmethode

Allgemeines Vorgehen

Der Test wird bei einer Stichprobe einmal durchgeführt. Dann wird der Test in zwei äquivalente Aufgabengruppen aufgeteilt. Die Rohwerte beider Hälften werden miteinander korreliert.

Hinweis:

Die Testhalbierungsmethode ist auch als Split-Half-Methode bekannt.

Besonderheit

Es bestehen mehrere Möglichkeiten, den Test in zwei gleichwertige Testhälften zu zerlegen.

a. odd-even-Methode

Nach Darbietung des Tests werden die Aufgaben aus geradzahligem und ungeradzahligem Reihungsnummer in zwei Hälften aufgeteilt.

Ist besonders bei Niveautests zu empfehlen, bei denen die Items nach steigendem Schwierigkeitsgrad angeordnet werden. Auf diese Weise bekommt man in beide Testhälften gleiche Schwierigkeiten.

b. Analysedaten

Noch besser ist es, außer der Schwierigkeit auch die Trennschärfe zu berücksichtigen. Man sucht dazu aufgrund der Analysedaten Aufgabenpaare von annähernd gleicher Schwierigkeit und Trennschärfe zusammen. Von den Paaren teilt man dann zufällig je eine der Aufgaben einer der beiden Testhälften zu (bei Niveautests).

c. Zufall

Wenn die Items gleich schwer sind, kann die Halbierung auch durch Zufall erfolgen.

d. Testzeit

Hier wird die für den ganzen Test vorgesehene Zeit halbiert. Nach Ablauf der ersten Hälfte der Testzeit werden die Probanden aufgefordert, an der eben von ihnen bearbeiteten Aufgabe ein Kennzeichen anzubringen und arbeiten dann an der nächstfolgenden Testaufgabe ohne Unterbrechung weiter.

Konsistenzanalyse

Allgemeines Vorgehen

Der Test wird bei einer Stichprobe von Probanden einmal durchgeführt. Der Test wird dann nicht nur in zwei vergleichbare Hälften, sondern in drei, vier oder in ebenso viele Teile untergliedert, wie Aufgaben vorhanden sind.

Hinweis:

Der Cronbach-Alpha-Koeffizient wird sehr häufig zur Bestimmung der Konsistenz herangezogen.

Besonderheit

Die Konsistenzanalyse ist eine Verallgemeinerung der Testhalbierung, ist dieser aber in vielerlei Hinsicht überlegen. Sie kann im Unterschied zur Halbierungsmethode bereits auf den Daten der Aufgabenanalyse aufbauen und bedarf deshalb nicht unbedingt einer speziellen Darbietung.

Nur wenn folgende Bedingungen eingehalten werden, liefert die Konsistenzanalyse eindeutige Ergebnisse:

- a.** Die Testaufgaben müssen homogen sein. Sind sie mehr oder weniger heterogen, so resultiert ein entsprechend niedriger Konsistenzkoeffizient. In einem solchen Fall kann eine Testwiederholung eine verlässlichere Reliabilitätsschätzung ermöglichen.
- b.** Der Test muss ein Niveautest sein. Die Schnelligkeitskomponente verursacht eine Scheinerhöhung des Konsistenzkoeffizienten.

Störquellen

Durch die unterschiedlichen Verfahren zur Bestimmung der Reliabilität ergeben sich die aufgeführten Störquellen, die die Fehlervarianz vergrößern und damit die Reliabilität verringern.

- Störquelle 1: Die Ungenauigkeit des Tests als Messinstrument (se^2 consist)
- Störquelle 2: Die Veränderlichkeit der Bedingungen der Testdurchführung (se^2 cond)
- Störquelle 3: Beim Einsatz eines Paralleltests ist davon auszugehen, dass die beiden Tests niemals vollständig äquivalent sein werden. Dadurch ergibt sich ein weiterer Fehleranteil (se^2 äqui).
- Störquelle 4: Bei Testwiederholungen wirken sich Wiederholungseinflüsse (se^2 rep) und Erinnerungseinflüsse (se^2 mem) aus.
- Störquelle 5: Die Merkmalsfluktuation (se^2 fluk) sogenannter aktueller Persönlichkeitsmerkmale (Emotionen, Dispositionen, Bedürfnisse), während habituelle Persönlichkeitsmerkmale eine relativ höhere Konstanz aufweisen.

Tabellarischer Überblick

Störquellen bei den verschiedenen Reliabilitätsschätzungen

	Testhalbierung und Konsistenzanalyse	Testwiederholung (sofort)	Paralleltest (sofort)	Testwiederholung (später)	Paralleltest (später)
se^2 (consist)	X	X	X	X	X
se^2 (äqui)			X		X
se^2 (cond)		X	X	X	X
se^2 (rep)		X	X	X	X
se^2 (mem)		X		X	
se^2 (fluk)				X	X

Fazit

Der Konsistenzkoeffizient gibt eine optimistischere Auskunft über die Reliabilität eines Tests als Retest- oder Paralleltestkoeffizienten.

Der Konsistenzkoeffizient sollte dann gewählt werden, wenn das untersuchte Merkmal offensichtlich inkonstant ist (aktuelles Persönlichkeitsmerkmal).

Das Paralleltestverfahren liefert im allgemeinen niedrigere Reliabilitätskoeffizienten.

Beim Retestverfahren sollte zur Testwiederholung ein Zeitpunkt gewählt werden, zu dem die Erinnerungseinflüsse weitestgehend unwirksam geworden sind, andererseits die Merkmalskonstanz noch als gegeben angenommen werden kann.

Grundsätzlich darf man nicht den einen Koeffizienten für besser und den anderen für schlechter halten, sondern muss jeden Koeffizienten im Hinblick auf das untersuchte Persönlichkeitsmerkmal und das zur Ermittlung benutzte experimentelle Verfahren beurteilen.

Bedeutung der Koeffizienten

Klicken Sie bitte auf die nebenstehenden Stichwörter, um weitere Informationen zu erhalten.

Konsistenz- und Halbierungskoeffizienten

Konsistenz-Koeffizienten und Halbierungskoeffizienten lassen situative Einflüsse und die Inkonstanz des Merkmals unberücksichtigt. Sie messen also im wesentlichen das, was ein Reliabilitätskoeffizient gemäss seiner Bestimmung messen soll, nämlich die Qualität des Tests selber. Die Retest- und Paralleltestmethode wird gegenüber diesem mehr theoretischen Interesse mehr den praktischen Bedürfnissen der Diagnostik gerecht.

Paralleltest-Koeffizient

Ein hoher Paralleltest-Koeffizient mit langem Zeitintervall deutet auf eine hohe Merkmalskonstanz hin.

Bei einem kurzen Zeitintervall deutet die Höhe des Koeffizienten auf die Bedingungskonstanz hin.

Bei gleichem Zeitabstand müsste der Paralleltest-Koeffizient noch etwas niedriger sein als der Retest-Koeffizient, weil er noch Fehlervarianz aufgrund mangelnder Parallelität enthalten könnte.

Retest-Koeffizient

Ein hoher Retest-Koeffizient mit langem Zeitintervall deutet auf eine hohe Merkmalskonstanz hin.

Bei einem kurzen Zeitintervall deutet die Höhe des Koeffizienten auf die Bedingungskonstanz hin.

Erbringt die kurzfristige Testwiederholung einen hohen, die langfristige aber einen niedrigen Reliabilitätskoeffizienten, so besteht entweder eine geringe Merkmalskonstanz oder aber der Test unterliegt einer sogenannten Funktionsfluktuation. Dieser Begriff meint, dass der Test nach einer gewissen Zeit etwas anderes prüft als vorher, oder bei unterschiedlichen Bedingungen jeweils einen anderen Aspekt des zu untersuchenden Merkmals misst.

Zuverlässigkeitsforderungen

An die Zuverlässigkeit eines Tests sind folgende Forderungen zu stellen. In der Forschung sind bereits Tests mit einem Zuverlässigkeitskoeffizienten um .50 verwendbar. In der Individualdiagnostik sind dagegen Tests mit Zuverlässigkeitskoeffizienten um .70 nur bedingt brauchbar und solche ab .80 als ausreichend, ab .90 als gut zu bezeichnen. Für die instrumentelle Zuverlässigkeit wird man in der Regel solche um .90, für die Stabilität (Retest-, Paralleltest) bereits solche um .80 als gut bezeichnen.

Reliabilitätsverbesserung

Um die Reliabilität eines Tests zu verbessern, bieten sich die beiden nebenstehenden Möglichkeiten an. Klicken Sie für weitere Informationen auf einen der Begriffe.

Testverlängerung

Sind die neu hinzunehmenden Aufgaben von der gleichen Art wie im Test (über eine nochmalige Aufgabenanalyse festzustellen), so kann man die Spearman-Brown-Beziehung (prophecy-formula) anwenden, um vorauszusagen, um wie viel die Reliabilität voraussichtlich größer werden wird, wenn man eine bestimmte Anzahl hinzunimmt (gilt auch für Testverkürzung).

Hinweis:

Ein bestehender Test kann natürlich nur in gewissen Grenzen verlängert werden, da er ökonomisch bleiben muss.

Ersatz schwieriger Aufgaben

Sind die im Test enthaltenen Aufgaben sehr unterschiedlich in der Schwierigkeit (d. h. eine große Streuung der Itemschwierigkeitsverteilung), dann sollte man einen Teil der Aufgaben mit extremer Schwierigkeit durch solche von mittlerer Schwierigkeit ersetzen. Dieses Vorgehen ist darin begründet, dass mittelschwere Items in der Regel reliablere Tests ergeben als solche mit extremer Schwierigkeit.



3. Bestimmung der Validität

Definition allgemein

Die Validität ist der Grad der Genauigkeit, mit der der Test das misst, was er messen soll. Anders ausgedrückt soll der Test das Merkmal, das er erfassen soll, in seinem ganzen Bedeutungsumfang oder in einem repräsentativen Umfang abbilden.

Definition konkret

Da es beim Testen niemals um das Testverhalten selbst geht, muss man das Testverhalten stets auf irgendwas "übertragen", auf ein Konstrukt (Begriff) oder auf ein Kriterium, das außerhalb des Testverhaltens liegt. Der Begriff der Validität bezieht sich auf die Frage, ob der Sprung vom Testverhalten zum Konstrukt oder Kriterium gerechtfertigt ist. Die Zusammenstellung von Beweismitteln für diese Rechtfertigung ist der Validierungsprozess. Der Grad, in dem diese Rechtfertigung nachzuweisen ist, wird durch die Validität angegeben.

Zeitpunkt

Meist wird die Validitätskontrolle parallel mit der Reliabilitätskontrolle durchgeführt. Allerdings kann diese auch noch im Anschluss an die Reliabilitätskontrolle erfolgen. Bei Eignungstests sollte aber schon in der Phase der Aufgabenanalyse der Test auf Validität geprüft werden.

Stichprobe

Ist eine kriterienbezogene Validierung geplant, dann sollte die Stichprobe repräsentativ für die zu testende Population sein. Die Größe kann zwischen 30 und einigen 100 Probanden variieren (abhängig von den eigenen Präzisionsansprüchen).

Formen der Validität

Eine Untergliederung der Validität ergibt sich aus einer Differenzierung der Rolle des Kriteriums bei den verschiedenen Testanwendungen. Es kommt darauf an, ob das Kriterium selbst das Ziel ist (kriterienbezogene Validität), oder ob es nur die Operationalisierung eines hypothetischen Begriffs ist (Konstruktvalidität).

Hinweis:

Für weitere Informationen, klicken Sie auf einen der Begriffe.

Kriterienbezogene Validität

Innere Validität

Erklärung

Der zu validierende Test wird mit anderen Tests korreliert, die für dasselbe Persönlichkeitsmerkmal als valide anerkannt sind. Dies setzt allerdings voraus, dass diese Tests einen möglichst hohen Validitätskoeffizienten gegenüber einem Außenkriterium aufweisen.

Beispiel

Ein neu entwickelter Gedächtnistest wird an einem bereits bewährten Test validiert. Eine repräsentative Stichprobe bearbeitet sowohl den neuen als auch den bewährten Test. Eine Korrelation aus Test- und Kriteriumswert ergibt den Validitätskoeffizienten.

Äußere Validität

Erklärung

Hierbei wird der Testpunktwert mit einem äußeren Kriterium (objektive Kriteriumsleistungen oder subjektive Schätzurteile) korreliert.

Beispiel

Ein äußeres Kriterium könnte in der Beurteilung durch eine Gruppe von Sachverständigen bestehen - etwa die Beurteilung des Schulerfolges durch Lehrer. Die Beziehungen zwischen Testbewertungen und einem Außenkriterium ließen sich besonders anschaulich durch bivariate Häufigkeitstabellen darstellen.

Tabelle: Bivariate Häufigkeitstabelle für die Variablen Gesamt-IQ (HAWIK) und Lehrerbeurteilungen

IQ	Lehrerurteil					N
	1	2	3	4	5	
140-149	3	1				4
130-139	12	14				26
120-129	12	85	2			99
110-119	2	167	84			253
110-109		102	293	7		402
90-99		16	290	77		383
80-89			55	155	7	217
70-79			3	62	19	84
60-69				7	17	24
50-59				2	4	6
40-49					2	2

Die Tabelle zeigt, dass kein Schüler mit einer Lehrerbeurteilung von "sehr gut" einen IQ < 110 erreichte. Insgesamt erzielen die von den Lehrern gut beurteilten Schüler hohe und die als schlecht beurteilten Schüler niedrige Intelligenzwerte. Der aus der o. g. Tabelle errechnete Validitätskoeffizient beträgt $r = 0,83$.

Übereinstimmungsvalidität

Erklärung

Bei der Übereinstimmungsvalidität werden Testwerte und Kriteriumswerte etwa gleichzeitig erfasst. Bei dieser Validierungsmethode werden die erhobenen Testresultate einer bestimmten Stichprobe mit außerhalb des Tests liegenden Vergleichsdaten (Außenkriterium) hinsichtlich des gleichen psychischen Merkmals verglichen. Im günstigsten Falle müssten dann beide Messwertreihen in der Weise übereinstimmen, dass die Individuen, die im Test den höchsten /niedrigsten Wert erhalten haben, auch in bezug auf das Außenkriterium die höchsten/niedrigsten Messwerte erhalten haben.

Beispiel

In der klinischen Praxis wird ein Test zur Anpassungsstörung eingesetzt. Gleichzeitig werden als Kriterienwerte Schätzurteile von Psychiatern und Lehrern, die die Probanden kennen, erfasst. Die anschließende Korrelationsberechnung zwischen Test- und Kriteriumswert ergibt den Validitätskoeffizienten.

Vorhersagevalidität

Erklärung

Bei der Bestimmung der Vorhersagevalidität wird erst hinterher überprüft, ob die aufgrund der Testresultate gehegten Erwartungen tatsächlich eingetreten sind.

Beispiel

Es wird ein Hochschuleingangstest für Psychologen dadurch validiert, dass am Ende des Vordiploms festgestellt wird, ob diejenigen Studenten, die im Test am besten abgeschnitten haben, auch das Vordiplom mit dem größten Erfolg durchlaufen haben.

Allgemeiner Hinweis

An dieser Stelle ergibt sich die Frage, warum man nicht nur das Kriterium erhebt und sich den Einsatz eines Testverfahrens erspart. Es sind vor allem ökonomische Gründe die für Testverfahren sprechen, denn die Erfassung der Kriterien ist mühsamer, aufwendiger, teurer, schwieriger oder dauert länger als die Erfassung von Testwerten (Prädiktoren). Außerdem sind bestimmte Kriterien nicht unbedingt direkt vorhanden und erst in weiter Zukunft verfügbar.

Konstruktbezogene Validität

Kriteriums- vs. Konstruktvalidität

Bei der Konstruktvalidierung steht nicht der weitgehend doch oberflächliche und praxisorientierte Vergleich zwischen Testergebnissen und Kriteriumswerten im Vordergrund. Vielmehr geht es um die umfassende inhaltliche psychologische Klärung dessen, was der Test zu messen vorgibt: also seinen psychologischen Gültigkeitsbereich. Konstruktvalidität ist daher in wesentlich höherem Maße als kriterienbezogene Validität in der persönlichkeitspsychologischen Grundlagenforschung verankert.

Nomologisches Netzwerk

Es wird nach und nach ein Netz von Prädiktoren und ihren empirischen Beziehungen geknüpft, das dann über den tatsächlichen Bedeutungsumfang des jeweiligen Konstrukts Aufschluss gibt. Konstruktvalidierung ist somit eine schrittweise Annäherung zwischen dem Konstrukt und dem Testverfahren. Je dichter das Netzwerk geknüpft ist, desto größer ist die Eindeutigkeit, dass ein Test tatsächlich den Merkmalsbereich erfasst, den er erfassen soll.

Statistische Möglichkeiten

- ❖ Korrelation des Tests mit mehreren Außenkriterien
- ❖ Korrelationen des Tests mit Tests ähnlichen Validitätsanspruches
- ❖ Korrelationen mit Tests, die andere Persönlichkeitsmerkmale erfassen
- ❖ Faktorenanalyse des zu validierenden Tests gemeinsam mit Außenkriterien, validitätsverwandten und validitätsdivergenten Tests
- ❖ Analyse interindividueller Unterschiede in den Testresultaten
- ❖ Analyse intraindividuelle Veränderungen bei wiederholter Durchführung mit und ohne systematische Variation der Durchführungsbedingungen

Inhaltsvalidität

Es gibt eine Reihe von Tests, deren Aufgabeninhalt keinen Zweifel darüber lässt, was durch den Test erfasst wird. Zum Beispiel erübrigt es sich bei einem Test zur Erfassung der Additionsfähigkeit, die Validität zu überprüfen, wenn seine Aufgaben ausschließlich das Zusammenzählen von Zahlen abverlangen. In diesem Fall stimmt also der Inhalt der Testaufgabe mit dem Testzweck selber überein.

Ein Koeffizient für die Inhaltsvalidität lässt sich nicht berechnen. Zur groben Orientierung kann bei solchen Tests deshalb ersatzweise der Reliabilitätskoeffizient herangezogen werden.

Wahl des Validitätskriteriums

Die Wahl eines für das Merkmal repräsentativen Validitätskriteriums gehört zu den schwierigsten Aufgaben der Testentwicklung. Es können drei Gruppen von Kriterien unterschieden werden. Klicken Sie dazu auf einen der Begriffe, um weitere Informationen zu erhalten.

Produktkriterien

Das sind beispielsweise Stückfertigung, Gehalt, verkaufte Einheiten pro Zeiteinheit, etc..

Aktionskriterien

Hierunter fällt etwa die Schnelligkeit (Zeit), mit der eine bestimmte Tätigkeit ausgeführt wird.

Subjektive Kriterien

Hierunter fallen alle Leistungen, die nicht objektiv zu bewerten sind. Zur Bewertung eignet sich vor allem die Rating-Skala. Am besten haben sich nach allgemeiner Erfahrung 5-, 7- und 9-stufige Skalen bewährt.

Problem: Durch die unterschiedliche Bewertungstendenz der einzelnen Beurteiler wird der Validitätskoeffizient vermindert. Es empfiehlt sich einen gewissen Ausgleich zu schaffen, in dem jedes Schätzurteil als Differenz von seinem zugehörigen Mittelwert eingestuft wird. Sollten allerdings die Schätzwerte auch eine unterschiedliche Streuung aufweisen (ein Beurteiler bewertet oft gut bzw. schlecht, ein anderer Beurteiler hält sich mehr an den Durchschnittswerten), so sollten die standardisierten Schätzwerte (z-Wert) als Kriteriumswerte herangezogen werden (diese Vorgehensweise ist jedoch nur gültig bei der Annahme einer intervallskalierten Beurteilung).

Individuelles Schätzverfahren

Jeder Proband wird von einem Beurteiler anhand einer festgelegten Skala eingestuft.

Vorteil: Schnell, ökonomisch.

Nachteil: Aufgrund großer Zufallsfehler der Schätzwerte wird der Validitätskoeffizient entsprechend niedrig ausfallen.

Kollektives Schätzverfahren

Jeder Proband wird von mehreren Beurteilern (2 - 5) gemeinsam (Konsens-Urteil) auf einer festgelegten Skala eingestuft.

Vorteil: Höherer Validitätskoeffizient ist zu erwarten (geringere Zufallsfehler).

Nachteil: Größerer Aufwand; Problemfall Gruppeneffekt durch Konsensbildung.

Mittelungs-Schätzverfahren

Jeder Proband wird von mehreren Beurteilern getrennt - keiner weiß etwas über das Urteil des anderen - auf einer festgelegten Skala eingestuft. Anschließend werden die verschiedenen Schätzwerte über den einzelnen Probanden gemittelt.

Vorteil: Unabhängige Beurteilung; höherer Validitätskoeffizient ist zu erwarten

Nachteil: Größerer Aufwand

Rangordnungsverfahren

Bei kleineren Stichproben werden die Probanden von mehreren Beurteilern unabhängig voneinander in eine Rangordnung gebracht. Anschließend werden die verschiedenen Rangplätze gemittelt und in T-Werte transformiert (nur dann lassen sich Maßkorrelationen berechnen).

Nachteil: Kaum durchzuführen bei größeren Stichproben.

Paarvergleichsverfahren

Bei größeren Stichproben werden alle möglichen Paarkombinationen der Probanden von mehreren Beurteilern unabhängig voneinander verglichen. In den einzelnen Paarvergleichen wird nur beurteilt, ob ein Proband bzw. seine Leistung besser oder schlechter war als die des anderen. Hieraus lassen sich Prozenträge ermitteln, die anschließend in z-Werte transformiert werden (nur dann lassen sich Maßkorrelationen berechnen).

Vorteil: Nicht so hohe Anforderung an die Beurteiler.

Statistische Methoden

Wenn ein Validitätskriterium klar definiert werden kann, ist die rechnerische Bestimmung eines Validitätskoeffizienten keine Schwierigkeit. Es ist lediglich zu unterscheiden, auf welchem Skalenniveau Testergebnis und Kriteriumswert variieren.

Hinweis:

Klicken Sie auf einen der Koeffizientennamen, um die Berechnung in SPSS zu verfolgen.

Reliabilität des Kriteriums

Der Validitätskoeffizient eines Tests hängt neben einigen speziellen Bedingungen vor allem von der Zuverlässigkeit des Tests und der Zuverlässigkeit des Kriteriums ab. Während die Zuverlässigkeit des Tests in der Regel recht gut abgeschätzt werden kann, kann das Kriterium leider nur in den seltensten Fällen auf seine Zuverlässigkeit geprüft werden. Letzteres bedeutet eine systematische Beeinträchtigung des Gültigkeitsmaßes. Der Validitätskoeffizient wird aufgrund der mangelnden Zuverlässigkeit des Merkmalskriteriums systematisch erniedrigt. Abhilfe schafft hier die Berechnung des sogenannten minderungskorrigierten Validitätskoeffizienten.

Hinweis:

Für weitere Informationen, klicken Sie bitte auf den Begriff "Minderungskorrigierte Validität".

Minderungskorrigierte Validität lässt sich berechnen, wenn die Reliabilität eines Validitätskriteriums empirisch ermittelt werden kann. Der minderungskorrigierte Validitätskoeffizient gibt Auskunft darüber, welche Validität der Test haben würde, wenn das Kriterium absolut reliabel wäre.

Zentrale Aspekte der Testkonstruktion

Berechnung

Eigentlich sollte eine zweimalige Gewinnung desselben Kriteriums an derselben Stichprobe vorgenommen werden. Aber diese durch die Wiederholung der Messung doch sehr aufwendige Verfahrensweise kann dadurch abgekürzt werden, indem die innere Konsistenz der subjektiven Schätzwerte berechnet wird.

Einschränkung

Der Einsatz der Minderungskorrektur ist nur dann gerechtfertigt, wenn man annimmt, dass das gemessene Persönlichkeitsmerkmal größere Konstanz aufweist als das stattdessen erhobene Validitätskriterium (Bsp. Intelligenztest und Lehrerurteil).

Forderung

Ein minderungskorrigierter Validitätskoeffizient ist stets im Testmanual anzugeben.

Die partielle Inkompatibilität

Die Reliabilität eines Tests ist eher durch homogene Aufgaben, die empirische Validität eher durch heterogene Aufgaben gewährleistet (Dilemma der partiellen Inkompatibilität). Fazit: Indem man das eine anstrebt, gefährdet man das andere. Dies wird insbesondere bei der Verteilung der Aufgabenschwierigkeit deutlich. Bei mittlerer Aufgabenschwierigkeit ist die Reliabilitätserwartung am größten. Für die Validität wirken sich aber gerade unterschiedliche Aufgabenschwierigkeiten am günstigsten aus. Für dieses Dilemma bieten sich die folgenden Lösungen an.

Lösungen

1. Variation der Trennschärfekoeffizienten
Die besten Chancen für optimale Reliabilitäts- als auch Validitätswerte hat man, wenn die Trennschärfekoeffizienten etwa von 0,3 bis 0,8 variieren.
2. Testbatterie
Das Dilemma der Inkompatibilität wird am besten gelöst, wenn man sich für eine Testbatterie entscheidet. Hier wahren die Teiltests die Reliabilität, die Batterie der Teiltests sichert die Validität.

Anforderung an die Validität

Für den Validitätskoeffizienten lassen sich keine starren Normen einführen wie beispielsweise für den Reliabilitätskoeffizienten. Statistisch müsste man Koeffizienten von $\geq .70$ verlangen; in der Praxis ist man allerdings schon mit Koeffizienten mit $.60$ sehr zufrieden. Ganz wichtig für die Validität eines Tests ist sein Verwendungszweck. So sollte etwa bei der Bestimmung eines Persönlichkeitsmerkmals oder Eignung eines Probanden der Test einen Validitätskoeffizienten $> .70$ aufweisen. Bei einer nichtindividuellen Auslese oder aber bei Gruppenvergleichen sind die Anforderungen an den Validitätskoeffizienten nicht so hoch.

