Evaluation des Testverfahrens

Gütekriterien: a) Reliabilität b) Validität c) Objektivität

Reliabilität

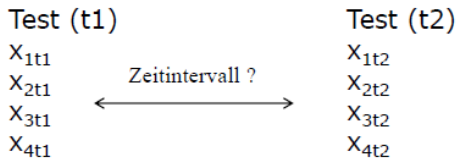
Methoden der Reliabilitätsbestimmung

1. Retestmethode
2. Paralleltestmethode
3. Konsistenzmethode

1. Retestmethode

1 Stichprobe	1 Test	2 Durchführungen

Berechnung: Korrelation der Testwerte aus Testdurchführung 1 mit den Testwerten aus Testdurchführung 2



Wahl der Länge des Intervalls:

- Beeinflusst die Höhe der Korrelation zwischen den Testwerten
- Kurzes Intervall: mögliche Überschätzung der Reliabilität durch z.B. Erinnerung, Lern-und Übungseffekte
- Großes Intervall: mögliche Unterschätzung der Reliabilität z.B. durch personenspezifische Veränderung durch neurologische Ausfälle, Medikamenteneinnahme usw.

Retestreliabilität

- Wird auch als Stabilität bezeichnet
- Stabilität des Merkmals über die Zeit

Länge des Intervalls

- Festlegung durch praktische Erwägungen (z.B. IQ zur Vorhersage von Berufserfolg -> mind. 2 Jahre Intervall)
- Annahme von Lern-und Übungseffekten (Intervall sollte nicht zu kurz sein)

Anwendung: Eher bei stabilen Merkmalen (Persönlichkeit, Intelligenz)

Metaanalyse (Charter, 2003)

Testkategorie	M	SD	25.-75. Perzentil	k
Klinische Verfahren, Erw. Persönlichkeit Berufseignung Intelligenz Pädagogischer Bereich	.74 .79 .76 .80 .79	.18 .10 .15 .14 .13	.64 - .86 .71 - .86 .68 - .85 .71 - .90 .72 - .87	36 64 46 36 35

2. Paralleltestmethode

1 Stichprobe	2 Tests	je 1 Durchführung

= 2 gleiche (sehr ähnliche) Tests werden korreliert. Vorteil: Zeitersparnis
Transfer-/Übungseffekte sind nicht ausgeschlossen

Berechnung:
Korrelation der Testwerte (z.B. Summenscore) aus Test A mit den Testwerten aus dem Paralleltest B.

Test A	Test B
x_{1t1}	x_{1t2}
x_{2t1}	x_{2t2}
x_{3t1}	x_{3t2}
x_{4t1}	x_{4t2}

Voraussetzung: Modell paralleler Messungen

- Paralleltest muss sich auf dieselben wahren Werte beziehen
- Messfehler (Messgenauigkeit) des Test A = Messfehler Test B
- Messfehler der Tests sind unkorreliert
- D.h. Korrelation der wahren Werte sollte 1 betragen
  ➔ Tests müssen äquivalent sein (Prüfung über Strukturgleichungsmodelle)

Anwendung: Leistungstest (Speed-und Power)

Störfaktoren:
- mangelhafte Parallelisierung
- Übungs-und Transfereffekte

Cross overdesign:
Gruppe 1: Test A –Test B
Gruppe 2: Test B –Test A

3. Konsistenzmethode

= schneller, bezieht sich nur auf den Inhalt des Tests

a) Testhalbierung

1 Stichprobe	1 Tests	1 Durchführungen

Berechnung:
Korrelation der Testwerte (z.B. Summenscore) der einen Testhälfte mit den Testwerten der anderen Testhälfte

Voraussetzung:
Modell paralleler Messungen (siehe Paralleltestethode)

Aufteilungstechniken:

- Odd-Even Methode (gerade/ungerade Itemnummer)
- Halbierung nach laufender Nummer
- Zufallsaufteilung
- Itemzwillinge (Schwierigkeit/Trennschärfe)
- Halbierung nach Testzeit (bei Speedtests)

Testhalbierung sollten möglichst identisch sein, wie bei Paralleltestung -> Schwierigkeit: Parallelität der Testhälften herstellen, z.B. bei heterogener Stichprobe -> Itemzwillinge

Problem: Korrelation der Testhälften unterschätzt die Reliabilität des Gesamttests (Länge und Reliabilität des Testverfahrens hängen zusammen, da mehrere Items Merkmale parallel testen

Korrektur:

Spearman-Brown-Formel

$$Rel_{korr} = \frac{m \cdot Rel}{1+((m-1)\cdot Rel)}$$

m = Faktor, um den sich die Itemzahl erhöht in diesem Fall m=2)

Testverdoppelung (bei parallelen Messungen) führt zu Verdoppelung der Messfehlervarianz und zu Vervierfachung der wahren Varianz.

b) Interne Konsistenz

1 Stichprobe	1 Tests	1 Durchführungen

Interkorrelationen zwischen Items -> Durchschnitt der Interkorrelationen -> Hochrechnung mit spearman brown ist in Cronbachs-alpha enthalten.

Berechnung:
Der Test wird im Prinzip in so viele Untertest zerlegt, wie er Items hat

Anwendung:

- Nur sinnvoll bei homogenen Tests
- wenn lediglich einmalige Erfassung sinnvoll (Befindlichkeitsfragebögen)

Cronbachs Alpha

= schätzt die durchschnittliche Korrelation zwischen allen Testitems, nach oben korrigiert um *m* durch die Spearman-Brown-Formel

$$\alpha = \frac{m \cdot \bar{r}}{1 + (m-1)\bar{r}}$$

m= Anzahl der Items (=parallele Messung)

r = durchschnittliche Interkorrelation zwischen den Items

$$\alpha = \frac{m}{m-1} \cdot \left(1 - \frac{\sum_{i=1}^{m} s_i^2}{s_t^2} \right)$$

$m =$ Anzahl der Items

$s_i^2 =$ Varinz des i-ten Items

$s_t^2 =$ Varianz des Tests t (Summenwert aller Items)

Zusammenfassung:

Retest und Parallel z.B. eher bei heterogenen Items,
Konsistenz eher bei homogenen Items da Schwankungen erfasst werden

	Re-test	**Parallel-test**	**Test-halbierung**	**Konsis-tenz**
Parallelform nötig	nein	ja	nein	nein
Zwei Testdurchführungen	ja	ja	nein	nein
Zwei Messzeitpunkte	ja	nein[a]	nein	nein
Überschätzung durch Erinnerung	ja	nein	nein	nein
Unterschätzung bei unsystematischen Merkmalsveränderungen	ja	nein[a]	nein	nein
Unterschätzung bei heterogenen Items	nein	nein[b]	ja[c]	ja

a - sofern Testformen direkt nacheinander vorgegeben werden
b - sofern Parallelität der Testformen sichergestellt ist
c - außer bei der Bildung tatsächlich paralleler Testhälften

Wie hoch sollte die Reliabilität sein?

So hoch wie möglich, doch tatsächlich hängt die Höhe von vielen Bedingungen ab, weshalb keine allgemeingültige Antwort gegeben werden kann.

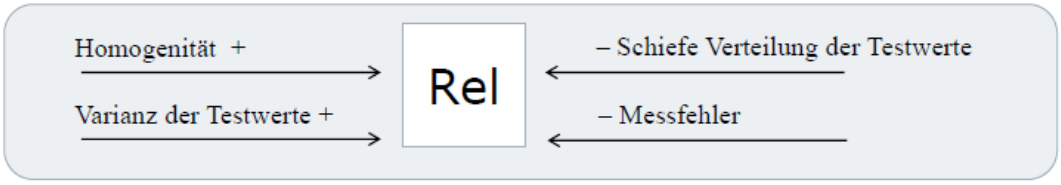
- Arte des Merkmals (Intelligenz versus Einstellung)
- Einsatzbereich: Individual-versus Kollektivdiagnostik

niedrig	<	.80
mittel		.80 - .90
hoch	>	.90

Möglichkeiten der Reliabilitätserhöhung

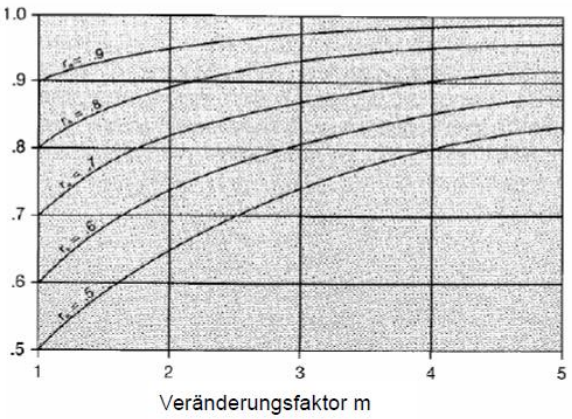
Einflussfaktoren:

1. Homogenität oder Heterogenität der Testitems (Trennschärfe)
2. Varianz und Verteilungsmerkmale der Testwerte
3. Verschiedene Arten von Messfehlern
   - vorübergehende (z.B. situationsbedingtes Antwortverhalten),
   - systematische (z.B. Übung, Antworttendenzen),
   - spezifische Messfehler (z.B. unterschiedliche Auffassung bezüglich gleicher Begriffe)



Reliabilitätserhöhung:

1. Messfehler reduzieren
   - klare Instruktionen
   - klare Formulierung der Items
   - standardisierte Testbedingungen
2. Homogenität erhöhen
   - Items mit geringer Trennschärfe entfernen (*alpha-Maximierung*)
     - Gefahr - facettenreiche Merkmale werden inhaltsarm
3. Testverlängerung
   - erhöht den Anteil wahrer Varianz



   - *Spearman-Brown-Formel:*

Vorhergesagte Reliabilität bei einer Verlängerung um den Faktor m

$$\mathrm{Rel}_{\text{Testverlängerung}} = \frac{m \cdot Rel}{1+((m-1)\cdot Rel)}$$

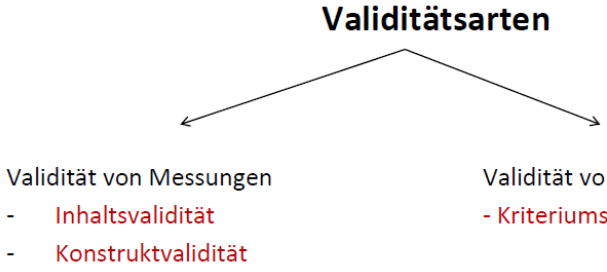
Validität

Definition: Validität

Was ist Validität? -> Wichtigstes Gütekriterium

- Unter Validität versteht man die Übereinstimmung der Testergebnisse mit dem, was der Test zu messen vorgibt.
- Maß der Genauigkeit, mit dem der Test dasjenige Merkmal misst, das er messen soll.
- **Beurteilung der Angemessenheit der Schlussfolgerungen** vom Testwert auf das Verhalten außerhalb der Testsituation oder auf die Ausprägung eines bestimmten Merkmals.

Arten von Validität



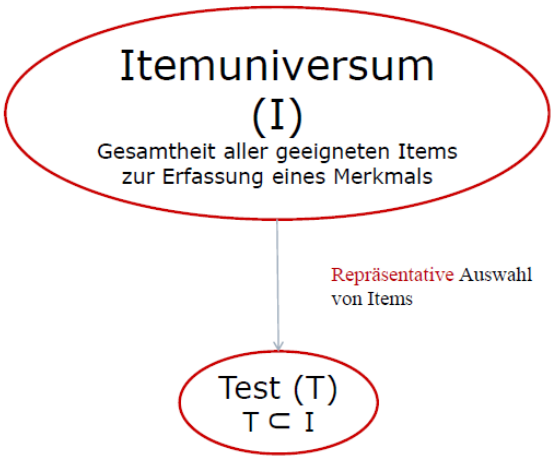
Inhalts- und Konstrukvalidität beziehen sich auf den Inhalt des Tests.

Kriteriumsvalidität bezieht sich auf die Entscheidung, die Konsequenzen für die getestete Person hat.

1. Inhaltsvalidität

- ist gegeben, wenn der Inhalt eines Tests (bzw. der Items) tatsächlich das interessierende Merkmal erfasst.
- Fragestellung: Wie repräsentativ ist der Inhalt eines Tests (sind die Items) für das zu messende Merkmal?
- Inhalt = Gesamtheit des Stimulusmaterials sowie aller Antwortalternativen
- Wird schon während der Konstruktion von Experten beurteilt

Idealfall: Der Test sollte das „Itemuniversum“ repräsentieren -> Repräsentative Untermenge



Beurteilung der Repäsentativität:

- Präzise Beschreibung des untersuchten **Inhaltsbereichs (content domain)**
- Bestimmung des Teils des Inhaltsbereiches, der durch jedes einzelne Item gemessen wird
- Vergleich der Struktur des Tests mit der des untersuchten Inhaltsbereichs
- Bestenfalls erfolgt eine systematisch Argumentation für den Schluss der Itembeantwortung auf das erfasste Konstrukt (und zwar bereits bei der Testkonstruktion)
  = Evidence-centered assessment design

Inhaltsbereich für Extraversion (NEO-PI-R)

(Gesamtheit der Verhaltensweisen zur Erfassung des Konstrukts)

Geselligkeit

Ich habe gerne viele Leute um mich herum.

Erlebnishunger

Ich bin gerne im Zentrum des Geschehens.

Aktivität

Ich bin ein sehr aktiver Mensch.

Frohsinn

Ich bin leicht zum Lachen zu bringen.

Herzlichkeit …

Durchsetzungsfähigkeit …

Argumentationskette:

1. Extravertierte Personen haben ein habituell niedriges kortikales Erregungsniveau.
2. Menschen empfinden ein mittleres Erregungsniveau als angenehm.
3. Es wird erwartet, dass extravertierte Personen stimulierende Situationen als angenehm erleben.
4. Das Item „Ich bin gerne im Zentrum des Geschehens." sollte daher von extravertierten Personen höher bewertet werden.

Fragen:

Ist der gesamte Inhaltsbereich durch Items erfasst?

Lassen sich alle Items dem Inhaltsbereich zuordnen (oder sind irrelevante Items vorhanden)?

Sind die Inhalte richtig gewichtet?

Lassen sich Evidenzen für den Schluss vom Item auf das Konstrukt finden?

Operational versus theoretisch definierte Merkmale

	Operational	**Theoretisch**
Merkmalsdefinition	Merkmal durch Testinhalt definiert	Merkmal im Rahmen einer Theorie definiert
	Merkmal direkt beobachtbar	Spezifikation eine latenten Variablen auf welche Unterschiede auf den manifesten Variablen rückgeführt werden
Anwendung	Leistungsbereich / Wissenstests	nicht beobachtbare Konstrukte
Item	Ist das Item Teil der interessierenden Gesamtheit möglicher Items?	Kann das interessierende theoretische Konstrukt Unterschiede in den beobachten Antworten erklären?

29

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Tests, die operational definierte Merkmale erfassen, sind dann inhaltsvalide, wenn die gewählten Aufgaben repräsentativ für das definierte Merkmal sind

z.B. Test zur Überprüfung des Lehrziels/Lernziels

- → Gewählte Aufgaben müssen eine Repräsentative Auswahl der Lehrinhalte darstellen (Sind alle relevanten Inhalte vorhanden?)
- → Aufgaben sollten repräsentativ für die jeweiligen Inhalte sein.
- → Aufgaben sollten in einem angemessenen Verhältnis zueinander stehen.
- → Aufgaben sollten keine irrelevanten Inhalte umfassen.

Die Feststellung der Inhaltsvalidität erfolgt in der Regel **nicht auf der Basis empirischer Untersuchungen.**

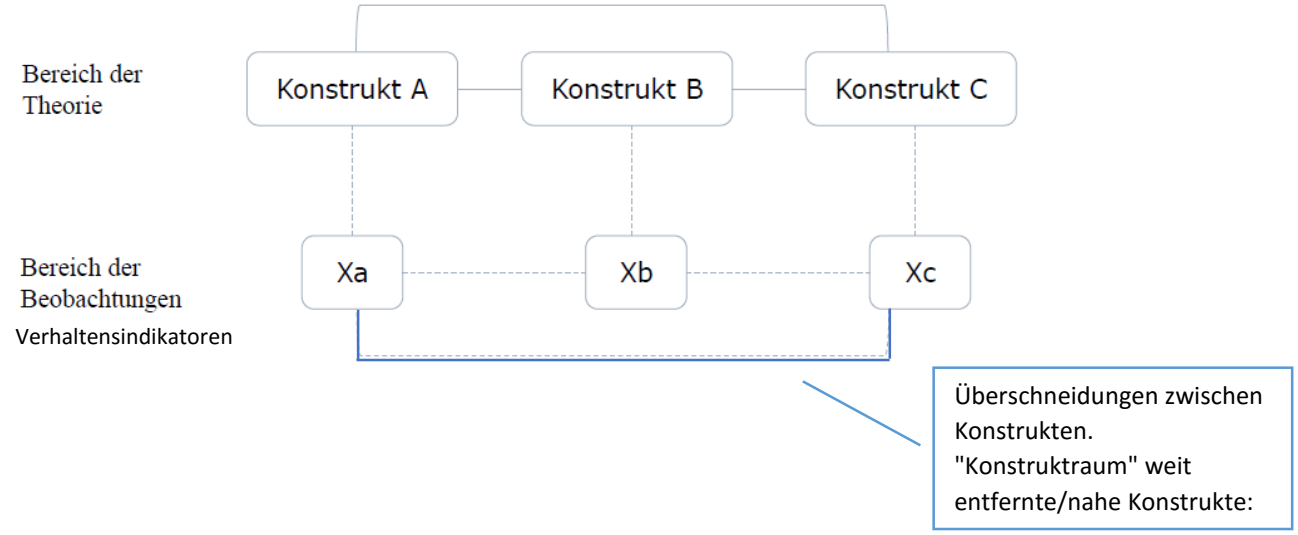
- Die Validität der Testinhalte ist durch theoretische Argumente sowie Expertenurteile gestützt und wird nicht empirisch quantifiziert.
- Inhaltsvalidität = theoretisch argumentativ (qualitatives Urteil auf Basis von Experten)

2. Konstruktvalidität

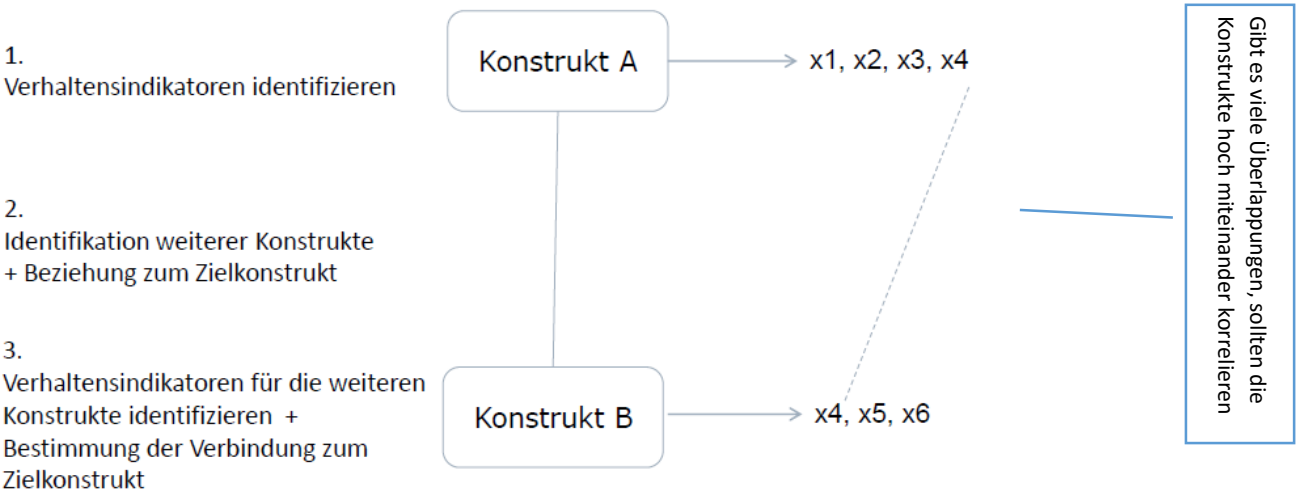
= ist ein empirischer Beleg (-> quantitative Bestimmung) dafür, dass der Test das **Konstrukt** misst, das er zu messen vorgibt – und nicht ein anderes.

Crohnbach und Meehl (1955) -> Die Idealvorstellung

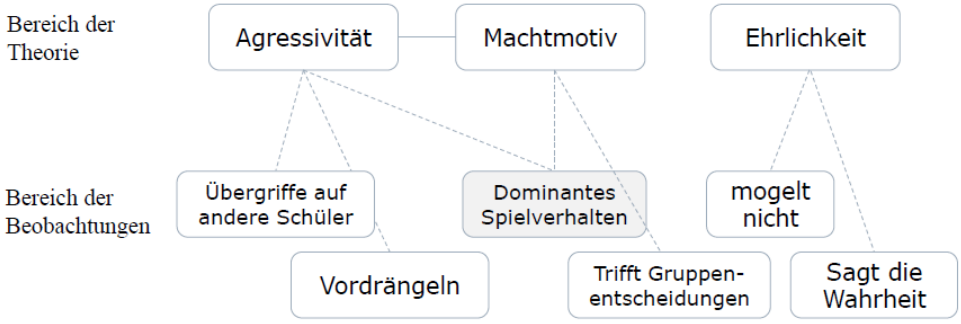
Das nomologische Netzwerk



Konstruktexplikation



Beispiel Agressivität bei Schülern



Verhaltensindikatoren	Erwartete Korrelation	Empirische Korrelation Test A	Empirische Korrelation Test B
Übergriffe auf andere Schüler	Stark positiv	.59	-.22
Vordrängeln	Stark positiv	.70	.14
Dominantes Spielverhalten	Stark positiv	.65	.02
Gruppenentscheidungen	Schwach positiv	.30	-.40
Mogelt nicht	Null	.09	.56
Sagt die Wahrheit	Null	-.04	.39
		Hohe Validität	Geringe Validität

Beispiel: Aggressivität: Die ersten drei Beobachtungsbereiche sollten stark positiv mit dem Testergebnis für Aggressivität korrelieren, da sie in der Theorie Indikatoren für Aggressivität sind. „Gruppenentscheidungen" indiziert schwächer über Konstrukt 2 -> „dom. Spielverhalten". Test B hat z.B. beim zweiten und dritten Indikator sehr schwache Korrelationen

Normalerweise wird nicht so ein Aufwand betrieben

Cronbach und Meehl (1955):
„There is no hope for developing in the short run the ‚nomological networks' we once envisioned"

➔ *Verstärkte Entwicklung formaler Theorien*
im Bereich der Psychologie ist für die Zukunft eher unwahrscheinlich.

Starker Ansatz:
Wenn ursprüngliches Ideal der Konstruktvalidität umgesetzt wird

Schwacher Ansatz:
Validierung ohne formale Theorie

Blinder Empirismus:
Weitgehend wahllose Korrelation – exploratives Erklären, wie die Ergebnisse zu interpretieren sind

Keine konkrete Einstufung, sondern Kontinuum zwischen „blinder Empirismus" und „starker Ansatz"

Empirische Bestimmung der Konstruktvalidität

Verschiedene Möglichkeiten, die sukzessiv durchgeführt werden können

- **Gruppenunterschiede** (z.B. Vergleich der Einstellung zur Kirche zwischen Kirchgängern und nicht Kirchgängern)
- **Korrelationen** (Tests zum gleichen Konstrukt sollten positive miteinander korrelieren)

  Konvergente Validität
  liegt vor, wenn hohe Korrelationen mit Test vorliegen, die dasselbe oder ähnliche Konstrukte erfassen

  Diskriminante Validität
  liegt vor, wenn geringer oder kein Zusammenhang zu Konstrukten vorliegt, zu denen theoretisch kein Zusammenhang angenommen wird.
- **Interne Struktur** (faktorielle Struktur – faktorielle Validität) -> *Faktorenanalyse*
- **Veränderungen** über die Zeit (hohe oder niedrige Stabilität lassen sich durch Retest-Reliabilität nachweisen)
- **Experimentelle Intervention** (z.B. Veränderung des Depressionswerts nach Therapie)
- Untersuchung des **Antwortprozesses** (Rechentest konfundiert mit Instruktionsverständnis)

Minderungskorrektur

(Implikation Reliabilität)

einfache Minderungskorrektur $corr\, r_{tc} = \frac{r_{tc}}{\sqrt{r_{cc}}}$

doppelte Minderungskorrektur $corr\, r_{tc} = \frac{r_{tc}}{\sqrt{r_{cc}} \cdot \sqrt{r_{tt}}}$

r_tc = Korrelation des Konstrukts mit anderem Test

r_cc = Reliabilität/Cronbachs Alpha des Konstrukts im Test A

r_tt = Reliabilität/Cronbachs Alpha des Konstrukts im Test B

Unterschied:
Die einfache Minderungskorrektur wird nur bei einem (eigenen) Test durchgeführt.
Die doppelte Minderungskorrektur wird bei zwei Tests (eigenem und anderem) angewandt.

Beispiel:

Extraversion(r_{tt})	BFI Extraversion (r_{cc})	Korrelation (r_{tc})
α = .803	α = .840	r = .713

Einfach Minderungskorrektur

$$corr\, r_{tc} = \frac{r_{tc}}{\sqrt{r_{cc}}} = \frac{.713}{\sqrt{.840}} = .78$$

Doppelte Minderungskorrektur

$$corr\, r_{tc} = \frac{r_{tc}}{\sqrt{r_{cc}} \cdot \sqrt{r_{tt}}} = \frac{.713}{\sqrt{.84}\ \sqrt{.803}} = .87$$

Validität ist eigentlich Korrelation zwischen zwei Testverfahren, die wiederrum durch Reliabilität beeinflusst wird.

Reliabilität ist "Deckel" für Validität.

Zusammenhang Rel. und Val ist U-förmig.

Durch Eingrenzung des Konstruktes wird ein Test z.B. nur Facette von breiterem Test und so sinkt die Korrelation wieder.

Einzige Möglichkeit der Validitätssteigerung über Rel. ist Testverlängerung

Minderungskorrektur:

- Information für die Frage nach perfekter Konstruktüberlappung
- zu berücksichtigen: Messfehler Streuung
- einfache Minderungskorrektur: Repräsentativster Wert, da Messfehler sowohl bei meinem Test als auch bei Vergleichstest vorhanden sind.
- Keine Anwendung bei Prädiktion, da Tuning der Validität gegen eigentlichen Sinn sprechen würde
- Normalerweise wird die originale Validitäten auch angegeben.

Multitrait Multimethod Ansatz (MTMM)

= systematische Variation von Traits und Methoden

Konvergente Validität
liegt vor, wenn Messungen eines Konstrukts, das mit verschiedenen Methoden erfasst wird, hoch miteinander korrelieren.

Diskriminante Validität
liegt vor, wenn Messungen verschiedener Konstrukte, die mit derselben Methode oder mit unterschiedlichen Methoden erfasst werden, nicht oder nur gering miteinander korrelieren.

Methodeneffekt
Wenn zwei unterschiedliche Konstrukte mit derselben Methode erfasst werden kann ein Teil der Korrelation möglicherweise auf die gemeinsame Methode zurückgeführt werden. MTMM erlaubt die Schätzung dieses des Methodenbias.

Korrelationsbasierte MTMM-Analyse

Trait		Methode A Selbsturteil		Methode B Freundeurteil		Methode C Elternurteil	
		1 Ver	2 Gew	1 Ver	2 Gew	1 Ver	2 Gew
Methode A Selbsturteil	1 Ver	(.90)					
	2 Gew	.47	(.89)				
Methode B Freundeurteil	1 Ver	**.20**	.08	(.93)			
	2 Gew	.01	.39	.40	(.89)		
Methode C Elternurteil	1 Ver	**.22**	.07	**.18**	.08	(.93)	
	2 Gew	.04	.35	-.05	.27	.48	(.92)

Bei unterschiedlichen Methoden sind Korrelationen oft niedrig. Schwierig Auf Validitätsbestimmung anzuwenden. Bei multimethod ist es immer schwierig hohe Korrelationen zu bekommen.

() = Montotrait-Monomethod (Reliabilität) Müsste eigentlich 1 sein, aber man trägt die Reliabilitäten ein.
X = Monotrait-Multimethod (konvergente Validität)
X = Heterotrait-Monomethod (diskriminante Validität)
X = Heterotrait-Heteromethod (diskriminante Validität)

Biesanz & West, 2004

- Nachweis der **konvergenten Validität**
  - Die Monotrait-Heteromethod-Koeffizienten sollten statistisch signifikant sein
  - Alternativ kann eine Mindestkorrelation festgelegt werden (unter Berücksichtigung der Reliabilitäten)
- Nachweis der **diskriminanten Validität**
  - Die Heterotrait-Heteromethod-Koeffizienten sind bei allen Vergleichen niedriger als die konvergenten Validitäten
  - Die Heterotrait-Monomethod-Koeffizienten sollten niedriger sein als die konvergenten Validitäten
  - Muster der Merkmalskorrel. sollte innerhalb sowie zwischen den Methoden konstant sein
- Nachweis des **Methodeneffekts**
  - Ein Methodeneffekt liegt vor, wenn Heterotrait-Monomethod-Koeffizienten signifikant höher sind als Heterotrait-Heteromethod-Koeffizienten

Kritik an der korrelationsbasierten MTMM-Analyse:
- Beurteilung der Validität lediglich auf Basis von Häufigkeitszählungen vieler Einzelvergleiche vorgenommen
- Keine exakten Entscheidungsregeln
- Reliabilität der Messung bleiben unberücksichtigt

➔ Prüfung des MTMM-Designs mittels **konfirmatorischer Faktorenanalyse**
- Trennung von Trait-, Methoden-und Messfehleranteil möglich
- Überprüfung der Eindimensionalität einzelner Traits

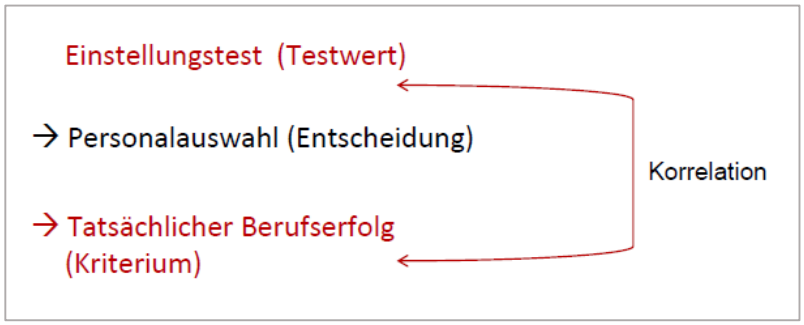
3. Kriteriumsvalidität

= Zusammenhang der Testleistung mit einem oder mehrerer Kriterien (z.B. Schulnote), mit denen der Test aufgrund seines Messanspruchs korrelieren sollte. Also muss Testleistung mit wirklicher Folge korreliert werden, um zu sehen, ob z.B. eine Prädiktion passt.
-> bedeutet, dass von einem Testergebnis auf ein für diagnostische Entscheidung praktisch relevantes Kriterium außerhalb der Testsituation geschlossen werden kann.

Kriterium = Maß an dem bestimmt werden kann, wie akkurat die aus den Testwerten abgeleiteten Entscheidungen sind.

Auswahl geeigneter Kriterien

- Kriterium muss für die zu treffende Entscheidung unmittelbar relevant sein (abgeleitet aus dem Anwendungszweck
- Kriterium sollte möglichst reliabel und valide operationalisierbar sein
- Beispiel: Studienauswahl
  - → Studienerfolg
    = wenn Studium nicht abgebrochen wir
    = wenn Studium in kurzer Zeit erfolgt
    = wenn Studium mit guten Noten abgeschlossen wird



Übereinstimmungsvalidität (concurrent validity) – *am besten/am meisten erwünscht*
Testwerte und externe Kriterien werden zeitlich parallel erfasst, z.B. nach der Entscheidung in einer intakten, vorselektierten Population

Vorhersagevalidität (predictive validity)
Testwerte werden vor der Entscheidung und das externe Kriterium zu einem späteren Zeitpunkt erhoben. *z.B. Aktuelle Mitarbeiter auch testen als Korrelationsgruppe -> Hinweiß auf Krit.-Val. aber nicht sehr gut, da kein Zeitintervall vorhanden ist.*

Retrospektive Validität - *problematisch*
Zusammenhänge mit zwischen Testwert und zeitlich vorher ermittelten Kriterium.

Inkrementelle Validität
Beitrag eines Tests zur Verbesserung der Vorhersage eines Krit. über einen anderen Test hinaus.

Probleme

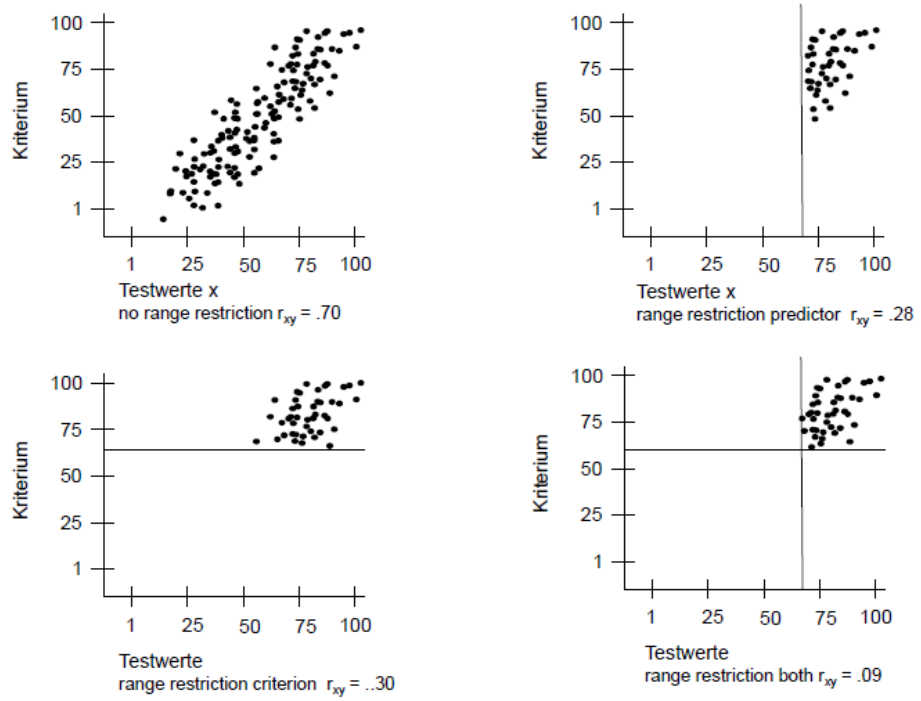
Probleme der Repräsentativität

- Wenn Tests zur Selektion verwendet werden (Personalauswahl, Zulassungstests, Einschulungstests, Entscheidung über Interventionen), dann bezieht sich die Korrelation zwischen Testwert und Kriterium bzw. die Validität nicht auf die gesamte Population auf welche der Test angewendet werden soll. -> *z.B. Abgelehnte können nicht mehr auf Vorhersage-Val. geprüft werden. Stichprobe wird homogener -> Varianzeinschränkung*

Statistisches Problem

- Selektion führt zur restriction of range (Spannweiteneinschränkung) der Testwerte
- Restriction of range führt zur Reduktion der Korrelation (Validitätsschätzung)
- Problem vor allem bei starker Selektion / geringen Auswahlquoten

Restriction of Range (statistisches Problem)



nach Murphy, K.R. & Davidshofer, C.O. (2005)

Formel zur Selektionskorrektur:

$$R_{tc} = \frac{\frac{r_{tc} \cdot S_x}{s_x}}{\sqrt{1 - r^2{}_{tc} + \frac{r^2{}_{tc} \cdot S^2{}_x}{s^2{}_x}}}$$

R_{tc} = korrigierter Validitätskoeffizient
r_{tc} = beobachteter Validitätskoeffizient
S_x = Messwertestreuung des Prädiktors ohne range restriction
s_x = beobachtete Messwertestreuung des Prädiktors

Streuung vor range restriction	Streuung nach range restriction	Geschätzte Korrelation in der Population (wenn beobachtete Korrelation r = .30)
.8	.7	.33
.8	.6	.38
.8	.5	.45
.8	.4	.53
.8	.3	.64
.8	.2	.78

Verschiedene range restrictions wurden angewandt.

Reliabilität

- Eine geringe Reliabilität des Prädiktors sowie des Kriteriums beeinträchtigt die Validitätsschätzung
- Einfache (unter Einbezug der Reliabilität des Kriteriums) oder doppelte Minderungskorrektur (unter Einbezug der Reliabilität des Kriteriums sowie des Prädiktors)

$$corr\, r_{tc} = \frac{r_{tc}}{\sqrt{r_{cc}}} \qquad corr\, r_{tc} = \frac{r_{tc}}{\sqrt{r_{cc}} \cdot \sqrt{r_{tt}}}$$

Vorsicht:
bei der Prädiktion ist der Messfehler ja schon integriert.

Interpretation des Validitätskoeffizienten (für Kriteriumsvalidität)

- Koeffizient kann zwischen 0 und 1 variieren
- Validitätskoeffizienten fallen häufig klein aus (ein Koeffizient von .5 ist bereits ein gutes Ergebnis) Validitätskoeffizient bestimmt das Ausmaß in dem der Test die Qualität von Entscheidungen verbessern kann.
- Ziel = so viele korrekte Entscheidungen wie möglich treffen
- Um die genaue Entscheidungsgüte eines Tests zu bestimmen muss außerdem die Basisrate (baserate) und die Selektionsrate (selection ratio) berücksichtigt werden

Zusammenhang zwischen Intelligenzleistung und Validitätskriterien

Herkunft der Studien	Validitätskriterien			
	Berufserfolg	Ausbildungs-erfolg	Bildungsniveau	Schulerfolg
International Europa Deutschland	(.51) .27 (.53) .33 (.62)	(.56) .29 (.53) .37 (.59)	.46 (.56)	.69

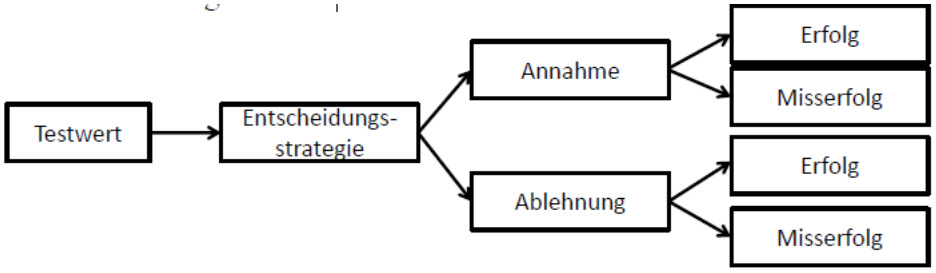
Korrelationen in Klammern sind korrigiert für Varianzeinschränkungen (range restriction) und Reliabilität von Prädiktor und Kriterium

Basisrate beeinflusst das Ergebnis der Entscheidung

- Basisrate = Häufigkeit potentielle Erfolgreicher in der Population
- Je höher die Basisrate maximiert die Anzahl angenommener erfolgreicher Bewerber, führt jedoch auch dazu, dass viele potentielle erfolgreiche Bewerber abgelehnt werden.

Selektionsrate (selection ratio) Anzahl der Stellen im Verhältnis zur Anzahl der Bewerber

- Selektionsrate = je geringer die Selektionsrate, desto größer ist die Wahrscheinlichkeit erfolgreiche Bewerber anzunehmen.



- Bestimmung der Bedeutung der Validität unter Berücksichtigung der Basisrate und der Selektionsrate

Table 9-4 TAYLOR–RUSSELL TABLE SHOWING THE EXPECTED PROPORTION OF SUCCESSES WITH A BASE RATE OF .50

	Selection ratio									
Validity	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90
.00	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50
.10	.54	.54	.53	.52	.52	.51	.51	.51	.51	.50
.20	.57	.54	.61	.59	.58	.56	.55	.53	.53	.52
.30	.74	.71	.67	.64	.62	.60	.58	.56	.54	.52
.40	.82	.78	.73	.69	.66	.63	.61	.58	.56	.53
.50	.88	.84	.78	.74	.70	.67	.63	.60	.57	.54
.60	.94	.90	.84	.79	.75	.70	.66	.62	.59	.45
.70	.98	.95	.90	.85	.80	.75	.70	.65	.60	.55
.80	1.00	.99	.95	.90	.85	.80	.73	.67	.61	.55
.90	1.00	1.00	.99	.97	.92	.86	.78	.70	.62	.56
1.00	1.00	1.00	1.00	1.00	1.00	1.00	.83	.71	.63	.56

Die Bedeutung eines Test und seiner Validität für die Güte von Entscheidungen steigt bei abnehmender Basisrate und Selektionsrate.

Auswahl und Beurteilung der Validität

Wahl einer geeigneten Validierungsmethode

1. Schwerpunkt: Repräsentationsschluss
   - Verallgemeinerung der Testwerte auf große Menge möglicher Aufgaben („Itemuniversum“)
   - Test zur Erfassung von Fähigkeiten, Lernzielen usw.
   - Inhaltsvalidität
2. Schwerpunkt: theoriebasierte Testwertinterpretation
   - Psychologische Forschung (theoretische Konstrukte)
   - Konstruktvalidität
3. Schwerpunkt: diagnostische Entscheidung
   - Entscheidung
   - Kriteriumsvalidität

*Welche Faktoren beeinflussen die Validität? -> **Einflussgrößen***

- Methodenfaktoren
  - Verwendung unterschiedlicher Erhebungsmethoden senkt die Korrelation (z.B. standardisierter Fragebogen vs. Verhaltensbeobachtung)
- Gemeinsame Methodenvarianz
- Kriteriumskontamination und –defizit
- Mangelnde Symmetrie zwischen Prädiktor und Kriterium
- Streuungsrestriktion
- Mangelnde Reliabilität Kriterium/Prädiktor

Einflussfaktoren

Merkmale des Tests
- Reliabilität des Tests

Merkmale des Kriteriums
- Reliabilität des Kriteriums
- Validität des Kriteriums

Gemeinsame Merkmale von Test und Kriterium
- Gemeinsame Methodenvarianz
- Konfundierung mit dem gleichen Merkmal

Merkmale der untersuchten Personen
- Stichprobenumfang
- Merkmale der Stichprobe

Wie hoch sind Validitätskoeffizienten?

Beurteilung durch Vergleichswerte

Persönlich-keitsmerkmal	**Validitätskriterium**					
	Verhalten im Alltag	Fremd-beurteilung	Schulerfolg	Ausbildungs-erfolg	Studien-erfolg	Berufs-erfolg
Neurotizismus	.53	(.51)	(.20)	.05 (.09)	(.01)	.06 (.13)
Extraversion	.42	(.62)	(.18)	.13 (.28)	(-.01)	.06 (.15)
Verträglichkeit	.55	(.46)	(.30)	.07 (.14)	(.06)	.06 (.13)
Gewissenhaftigkeit	.48	(.56)	(.28)	.13 (.27)	(.23)	.12 (.27)
Offenheit	.56	(.59)	(.24)	.14 (.33)	(.07)	.03 (.07)

() = minderungskorrigiert

Objektivität

= Ausmaß, in dem die Ergebnisse eines Tests unabhängig von der Person des Untersuchungsleiters sind.

- Wird gesichert durch Standardisierung der einzelnen Phasen des diagnostischen Prozesses.

Arten von Objektivität:

- Durchführungsobjektivität
- Auswertungsobjektivität
- Interpretationsobjektivität

Durchführungsobjektivität

- Sicherung durch maximale Standardisierung der Testsituation
- Beispiel:

Instruktion für den Anwender: Aus dem HAWIE (Wechsler, 1964):
„Bei der Durchführung des Tests muß der VL unbedingt die Anweisung befolgen. Sie müssen wörtlich auswendig gelernt werden. Der VL soll die VP während des Tests nicht in ein Gespräch verwickeln; erlaubt sind nur notwendige Ermunterungen der VP. Die Anweisung dürfen so oft wie erforderlich wiederholt, jedoch nicht erklärt werden."

Auswertungsobjektivität

- Gegeben bei eindeutiger Quantifizierung des Verhaltens
- Geringer bei freiem als bei gebundenem Antwortformat
- Quantitative Bestimmung der Auswertungsobjektivität:
  Vergleich der Ergebnisse von mindestens zwei Auswertern
- Bestimmung der Auswertungsobjektivität mittels Intraklassenkorrelation (Maß für Beurteilerübereinstimmung)

Aus dem „Lern- und Gedächtnistest (LGT 3)"(Bäumler, 1974)
In der 3. Aufgabe werden dem Probanden 20 Gegenstände gezeigt (bildlich), z.B. Kleiderbügel, Hammer, Roller.
Zur Auswertung werden folgende Hilfen angeboten:

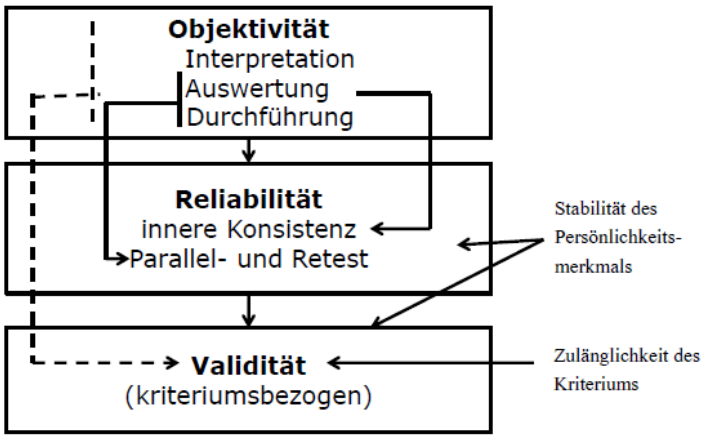
Richtig	***Noch gültig***	***Ungültig***
Kleiderbügel	*Aufhänger, Haken*	*Bogen, Bumerang*
Hammer	*Schlegel*	*Werkzeug*
Roller	*Zweirad*	*Fahrrad, Rad*

Interpretationsobjektivität

- Gegeben wenn aus gleichen Scores von verschiedenen Auswertern die gleichen Schlüsse gezogen werden

Zusammenfassung: Gütekriterien

Wechselbeziehung zwischen den Gütekriterien Objektivität, Reliabilität und Validität (nach Lienert)



Objektivität als Basis für Reliabilität, insbesondere die Auswertungs- und Durchführungsobjektivität.
Sind diese nicht gegebene, wirkt sich das auf innere Konsistenzen und Parallel- und Retest-Reliabilität aus.

Reliabilität als „Deckel" für Validität.

weitere Gütekriterien:
z.B: Kulturfairness, Ökonomie, Anwendbarkeit, Handhabbarkeit

Beurteilung der Höhe von Testkennwerten (nach Weise, 1975)

Kennwert		Kürzel	Niedrig	Mittel	Hoch
Schwierigkeit		P*)1	<.20	.20 - .80	> .80
Trennschärfe	(korrigiert)	r_{itc}	< .30	.30 - .50	> .50
Objektivität	(Auswerter)	r_k	< .60	.70 - .90	> .90
Reliabilität		r_{tt}	< .80	.80 - .90	> .90
Validität *)1	(unkorrigiert)	r_{tc}	< .40	.40 - .60	> .60
Eichstichprobe		N	< 150	150 - 300	> 300

*)1 Validitätskoeffizient besagt wenig über die Bedeutung des Tests, wenn man nur den absoluten Betrag bewertet. Beachtet werden muss auch der Beitrag, den ein Test zur Lösung einer gegebenen Fragestellung leisten kann (z.B. Basis-, Selektionsrate und Validität)
*)2 p = relativer Anteil von Probanden, die ein Item „richtig" beantworten.

Beispiel für Richtlinien zur Bewertung der Reliabilität sowie des Umfangs der Normstichprobe (Evers, 2001)

	Reliabilität[1]			Umfang Normstichproben		
	Niveau 1	Niveau 2	Niveau3	Niveau 1	Niveau2	Niveau3
unzureichend	< .80	< .70	< .60	< 300	< 200	< 100
ausreichend	.80 - .90	.70 - .80	.60 - .70	300 – 400	200 – 300	100 – 200
gut	> .90	> .80	> .70	> 400	> 300	> 200

Anmerkung: Niveaus (1) Tests für wichtige Entscheidungen auf der individuellen Ebene (z.B. Personalauswahlentscheidungen), (2) Tests für weniger bedeutsame Entscheidungen auf individueller Ebene (z.B. Fortschrittskontrolle), (3) Tests für Untersuchung auf Gruppenniveau.
[1] Für Paralleltest-Reliabilität, interne Konsistenz, Test-Retest-Reliabilität und Interrater-Reliabilität