Institut für Medizinische Biometrie und Informatik,
Universitätsklinikum Heidelberg

Heidelberg, 24.11.2005

Evaluation von Fragebogen in der Medizin

- Theoretische Grundlagen und methodische Umsetzung

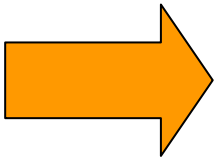
Dipl.-Psych. Wilmar Igl

Institut für Psychotherapie und Medizinische Psychologie,
Arbeitsbereich Rehabilitationswissenschaften
Universität Würzburg

I. Einleitung

Der Fragebogen als medizinisches Messinstrument

- Evidence-based Medicine erfordert Beurteilung der Wirksamkeit einer Behandlung
- mögliche Maße zur Beurteilung der Wirksamkeit
  - biometrische Maße (z.B. Blutdruck, Gelenkbeweglichkeit)
  - psychometrische Maße (z.B. Depressivität, subjektive Gesundheit)
- Bedeutung psychometrischer Maße insbesondere bei Patienten mit chronischen Krankheiten (z.B. Herzinsuffizienz, Diabetes)
- wesentlicher Einfluss der Güte der Messinstrumente auf Messung des Therapieerfolgs (z.B. Aussagekraft der Studie) und Studienplanung (z.B. Fallzahlplanung)



Evaluation von Fragebogen notwendig!

Überblick

I. Einleitung
II. Grundlagen der Konstruktion und Evaluation von Fragebogen
    1. Was ist ein Fragebogen?
    2. Konstruktion von Fragebogen
    3. Klassische Testtheorie (KTT)
III. Spezielle Methoden zur Bestimmung von Gütekriterien
    1. Objektivität
    2. Reliabilität
    3. Validität
    4. Änderungssensitivität
IV. Beispiel für eine Studie zur Fragebogenevaluation
V. Zusammenfassung

II. Grundlagen der Konstruktion und Evaluation von Fragebogen

1. Was ist ein Fragebogen?

Begriffsbestimmung „Fragebogen“

- Ein Fragebogen ist eine Sammlung von Fragen für eine systematische Befragung einer Stichprobe.

- Beispiele für Konstrukte:
  - Lebensereignisse (z.B. Heirat, Kinder, Beruf)
  - Verhaltensweisen (z.B. Fernsehgewohnheiten, Hobbies)
  - aktueller Zustand (z.B. Stimmung, subjektive Gesundheit)
  - u.v.m

- (standardisierte) Fragebogen werden vorwiegend zur Untersuchung von aggregierten Werten (z.B. Mittelwerte von Gruppen) in der Forschung angewendet!

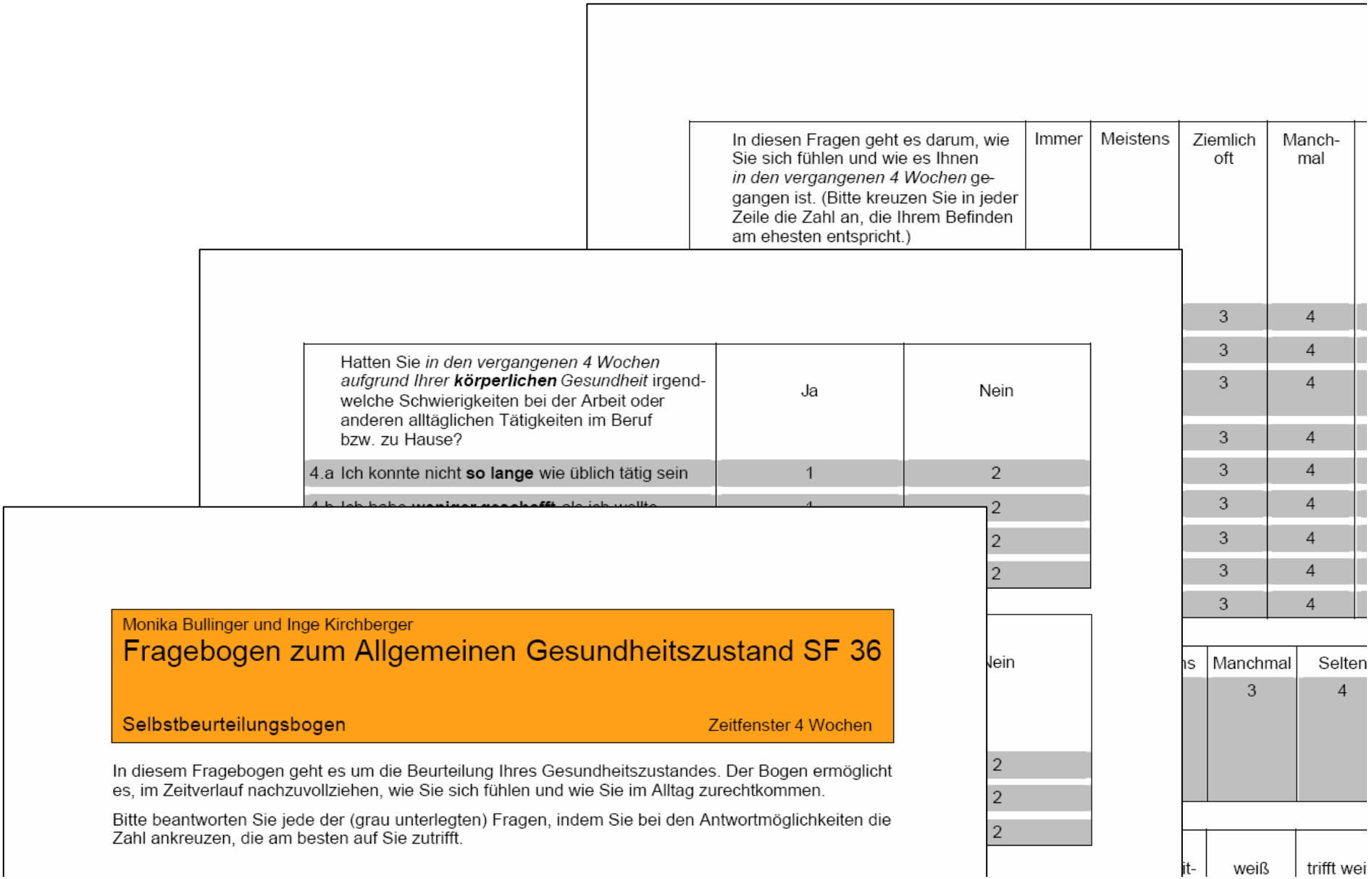
Begriffsbestimmung „Test“

- Ein Test macht quantitative Aussagen über den relativen Grad der individuellen Merkmalsausprägung.

- Wichtige Konstrukte:
  - Persönlichkeit, z.B. Extraversion, emotionale Stabilität, ...
  - Leistung, z.B. Intelligenz, berufliche Eignung,...

- Tests werden unter Verwendung von Normwerten häufig zur Einzelfalldiagnostik im klinischen Bereich eingesetzt!

SF-36 (Bullinger & Kirchberg, 1998)



In diesen Fragen geht es darum, wie Sie sich fühlen und wie es Ihnen *in den vergangenen 4 Wochen* gegangen ist. (Bitte kreuzen Sie in jeder Zeile die Zahl an, die Ihrem Befinden am ehesten entspricht.)	Immer	Meistens	Ziemlich oft	Manch-mal
			3	4
			3	4
			3	4
			3	4
			3	4
			3	4
			3	4
			3	4

Hatten Sie *in den vergangenen 4 Wochen aufgrund Ihrer* ***körperlichen*** *Gesundheit* irgendwelche Schwierigkeiten bei der Arbeit oder anderen alltäglichen Tätigkeiten im Beruf bzw. zu Hause?	Ja	Nein
4.a Ich konnte nicht **so lange** wie üblich tätig sein	1	2
	1	2
		2
		2

	Manchmal	Selten
	3	4

Monika Bullinger und Inge Kirchberger

Fragebogen zum Allgemeinen Gesundheitszustand SF 36

Selbstbeurteilungsbogen — Zeitfenster 4 Wochen

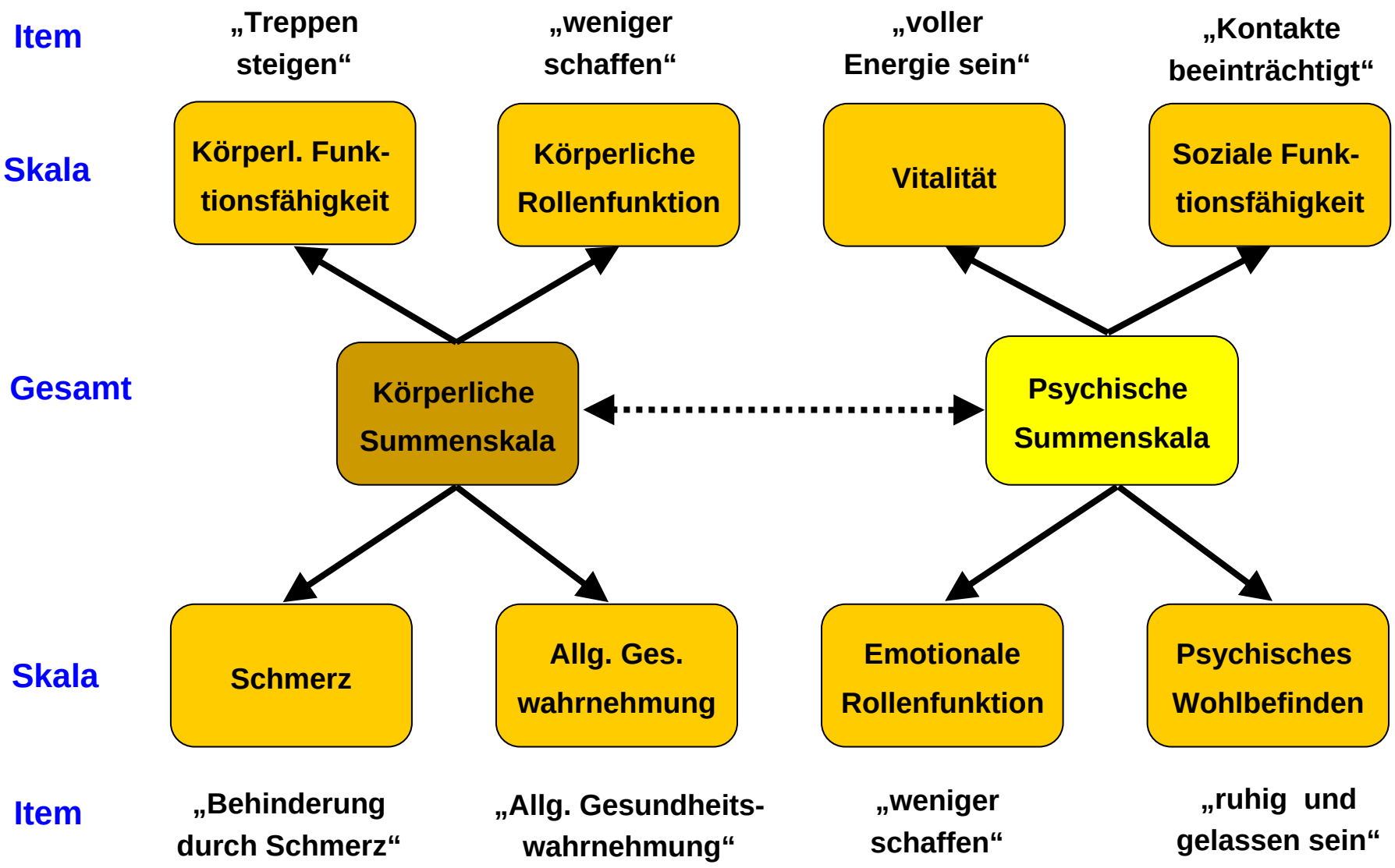
In diesem Fragebogen geht es um die Beurteilung Ihres Gesundheitszustandes. Der Bogen ermöglicht es, im Zeitverlauf nachzuvollziehen, wie Sie sich fühlen und wie Sie im Alltag zurechtkommen.

Bitte beantworten Sie jede der (grau unterlegten) Fragen, indem Sie bei den Antwortmöglichkeiten die Zahl ankreuzen, die am besten auf Sie zutrifft.

SF-36: Steckbrief

- Konstrukt: subjektive Gesundheit
- Beurteiler: Selbst-, Fremdbeurteilung, Interview
- Beurteilungszeitraum: 1 Woche bzw. 4 Wochen
- Itemzahl: 36 Fragen
- Antwortformat: mehrstufig (2 bis 6 Stufen)
- Aggregation: 2 Summenskalen, 8 Skalen

SF-36: Struktur



2. Konstruktion von Fragebogen

Vorbereitung

- Suche nach vorhandenen Messinstrumenten
  (evtl. auch in anderer Sprache)

- möglichst genaue Definition des interessierenden Konstrukts

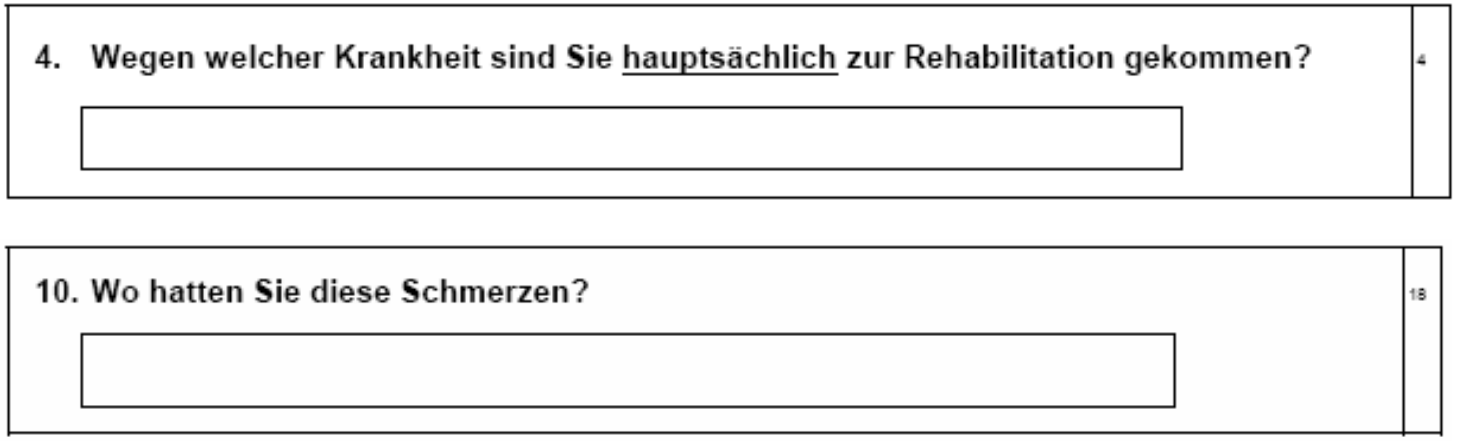
- Festlegung des Verhaltensbereichs

- Bestimmung der Zielpopulation

- Generierung von Items (z.B. über Literaturrecherche von Tests)

Itemformulierung

- Items mit offener Beantwortung

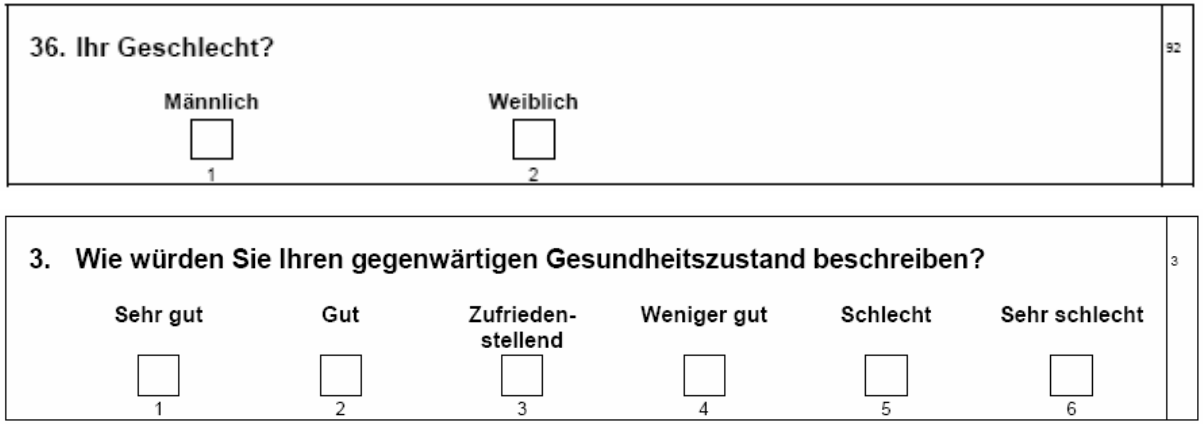


Beispiele aus IRES-3 (Bührlen et al. 2005)

- Vorteile:
  - keine Beschränkung der Antwort
  - wichtige Informationen für Fragebogenkonstruktion
- Nachteil:
  - keine quantitative Aussage möglich

Itemformulierung

- Items mit Antwortvorgaben



36. Ihr Geschlecht?

92

Männlich ☐ 1 Weiblich ☐ 2

3. Wie würden Sie Ihren gegenwärtigen Gesundheitszustand beschreiben?

3

Sehr gut ☐ 1 Gut ☐ 2 Zufrieden-stellend ☐ 3 Weniger gut ☐ 4 Schlecht ☐ 5 Sehr schlecht ☐ 6

Beispiele aus IRES-3 (Bührlen et al. 2005)

- Vorteile: objektiv, ökonomisch, (quantitativ auswertbar)
- Nachteile:
  - Antwortgaben entsprechen ggf. nicht Ansicht des Probanden
  - Boden-, Deckeneffekte

Rating-Skalen

	Numerische Skala				
Dimension	**1**	**2**	**3**	**4**	**5**
Zustimmung	völlig falsch	ziemlich falsch	unentschieden	ziemlich	völlig richtig
	stimmt gar nicht	stimmt wenig	stimmt teils-teils	stimmt ziemich	stimmt völlig
	trifft eindeutig nicht zu	trifft nicht zu	trifft weder zu noch nicht zu	trifft zu	trifft eindeutig zu
Häufigkeit	nie	selten	gelegentlich	oft	immer
Intensität	gar nicht	kaum	mittelmäßig	ziemlich	außer-ordentlich
Wahrschein-lichkeit	keinesfalls	wahrschein-lich nicht	vielleicht	ziemlich wahr-scheinlich	ganz sicher
	Verbaler Anker				

Anzahl der Skalenstufen

- ungerade: neutrale Mittelkategorie
- gerade: Zwang zur (tendenziellen) Entscheidung
- Anzahl der Skalenstufen: Hinweise auf die geringe Bedeutung für Reliabilität und Validität
- „Prominenzstruktur des Dezimalsystems“ bei zu feiner Differenzierung (100 Stufen ⇨ 5er- oder 10er-Kategorien)
- Empfehlung: 5- bis 6-stufige Lösung meist gut geeignet

Exkurs: Messtheoretische Probleme

- Skalenniveau:
  - Puristen: keine Intervallskala, sondern höchstens Ordinalskala
  - Pragmatiker: Intervallskala, da keine großen Abweichungen

- Angemessenheit eines statistischen Verfahrens: messtheoretisches Interpretationsproblem vs. mathematisch-statistische Voraussetzungen

- Robustheit parametrischer statistischer Tests bei nicht exakt intervallskalierten Daten

- Kriterium: inhaltlich sinnvolle Ergebnisse, die sich in der Praxis bewähren

Itemanalyse

- zentrales Instrument der Testkonstruktion und Testbewertung an Eichstichprobe (vgl. Zielstichprobe)

- statistische Methoden:
  - Rohwerteverteilung (Mittelwert, Streuung, Normalverteilung)
  - Itemschwierigkeiten
  - Trennschärfe
  - Homogenität (vgl. Interne Konsistenz)
  - Dimensionalitätsprüfung (vgl. Faktorenanalyse)

Itemschwierigkeit(sindex)

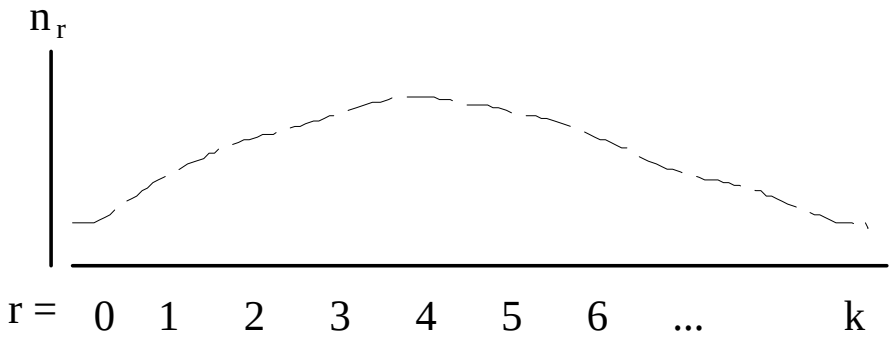
$$p_i = \frac{\sum_{m=1}^{n} x_{im}}{k_i \cdot n}$$

- („psychometrische“) Schwierigkeit = Anteil der Personen, welche die Aufgabe richtig „lösen“
- („psychologische“) Schwierigkeit = Anteil der Personen, welche die Aufgabe nicht (!) richtig „lösen“

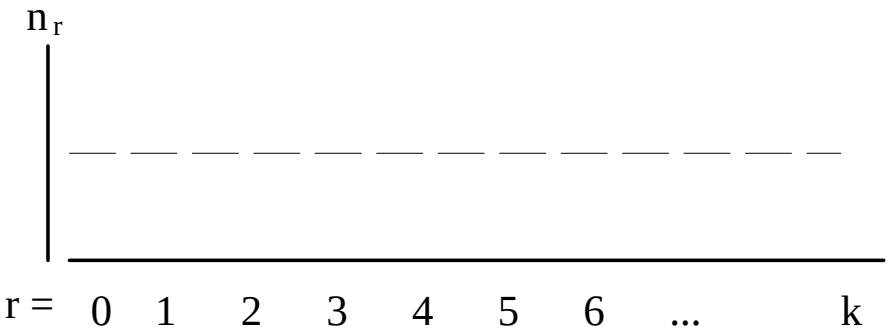
- Ziel:
  - hohe Streuung von Itemschwierigkeiten
  - mittlere Schwierigkeit (p=[0,20; 0,80]), da extrem schwierige Items keine Personenunterschiede anzeigen

Verteilung von Itemwerten

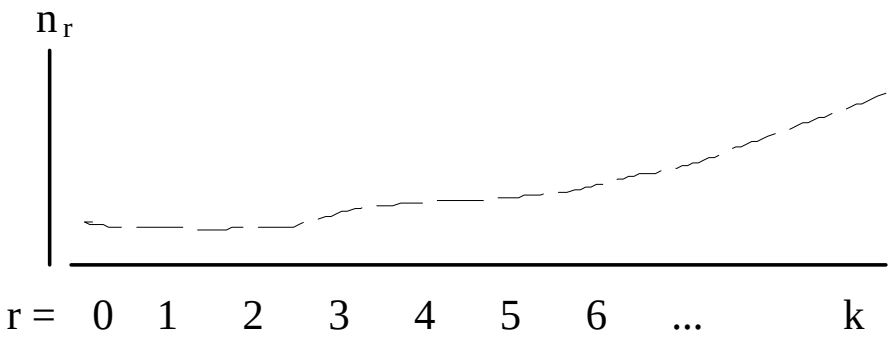
Eingipflige (unimodale) Verteilung



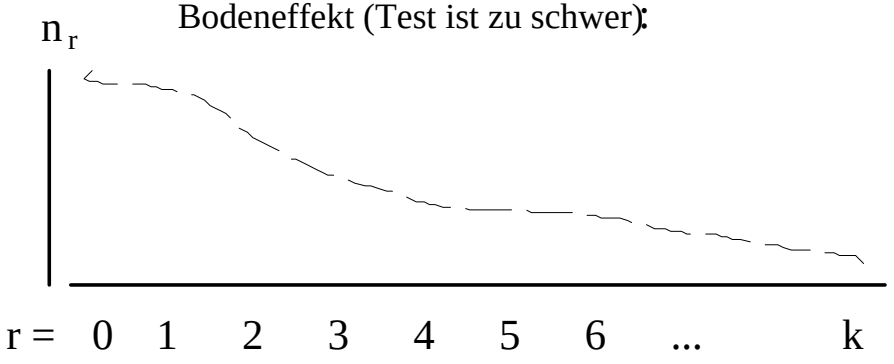
Gleichverteilung:



Deckeneffekt (Test ist zu leicht):



Bodeneffekt (Test ist zu schwer):



Quelle: Rost 2004, S. 92

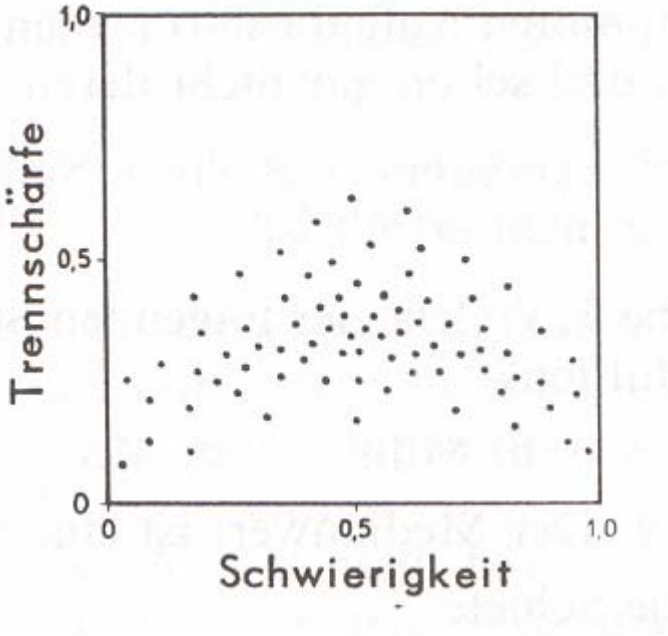
Trennschärfe(koeffizient)

$$r_{i,t\text{-}i} = \frac{\mathrm{cov}(i,t\text{-}i)}{s_i \cdot s_{t\text{-}i}}$$

- Korrelation eines Items mit dem Gesamttest
- part-whole korrigierter Trennschärfekoeffizient
  ohne Berücksichtigung des betrachteten Items im Gesamtwert
- Wertebereich: $r_{i,t\text{-}i}=[-1;+1]$
- Bewertungsheuristik:
  - „hoch“: $r_{i,t\text{-}i} > 0.50$
  - „mittel“: $r_{i,t\text{-}i}=[0.30; 0.50]$
  - „niedrig“: $r_{i,t\text{-}i}=[0.00; 0.30]$
- Beispiel: Depressivitätstest
  Frage 1:„Ich denke oft daran, mir das Leben zu nehmen.“ (Ja vs. Nein)
  Frage 2:„Ich bin mit mir oft unzufrieden.“ (Ja vs. Nein)

Schwierigkeit und Trennschärfe

Je extremer die Schwierigkeit, desto geringer die Trennschärfe!



Quelle: Lienert & Raatz 1994, S. 31

Homogenität(sindex)

$$\bar{r}_{\text{homogen, Item}} = \frac{\sum_{j}^{k-1} r(i,j)}{k-1}$$

- itemspezifische Homogenität: mittlere Korrelation eines Items mit allen anderen Items
- skalenspezifische Homogenität: mittlere Korrelation aller Itempaare einer Skala (~ Cronbach‘s Alpha)
- Bewertungsheuristik: „akzeptabel“: $r_{homogen}$= [0,20; 0,40]

Beispiel	Item 1	Item 2	Item 3	Item 4	Gesamt
Item 1	1.00				
Item 2	0,05	1,00			
Item 3	0,17	0,42	1,00		
Item 4	0,12	0,37	0,54	1,00	
Homogenitäten	0,11	0,28	0,38	0,34	0,27

Quelle: Bortz 1995, S. 201

Dimensionalität

- Bestimmung der Zahl der erfassten Konstrukte mittels der Faktorenanalyse

- Zusammenfassung von mehreren ähnlichen (korrelierenden) Items zu Skalenwerten
  - homogene Faktorenladungen
    ⇨ ungewichteter, additiver Gesamtwert
  - ungleiche Faktorenladungen
    ⇨ gewichteter, additiver Gesamtwert

3. Klassische Testtheorie

Testtheorie allgemein

- Eine Testtheorie versucht Bedingungen zu formulieren, welche erfüllt sein müssen, um von der Beantwortung einzelner Fragen auf die Ausprägung übergeordnete Konstrukte zu schließen.

Beispiel:

- Konstrukt: „Wie hoch ist die subjektive Gesundheit von orthopädischen Rehabilitanden bei Beginn einer Rehabilitationsbehandlung?“

- Frage: „Wie stark sind Sie durch Ihren derzeitigen Gesundheitszustand bei anstrengenden Tätigkeiten, z.B. schnellem Laufen, eingeschränkt?“

Klassische Testtheorie

- naturwissenschaftliches Messmodell als Grundlage
  Beispiele: Thermometer, Strommessgerät

- deterministischer Ansatz (vs. probabilistischer Ansatz):
  Das Testergebnis entspricht (abgesehen vom Messfehler) der latenten Merkmalsausprägung.

- Klassische Testtheorie ist sowohl auf „Tests“ als auch „Fragebogen“ anwendbar.

„Axiome“ der Klassischen Testtheorie

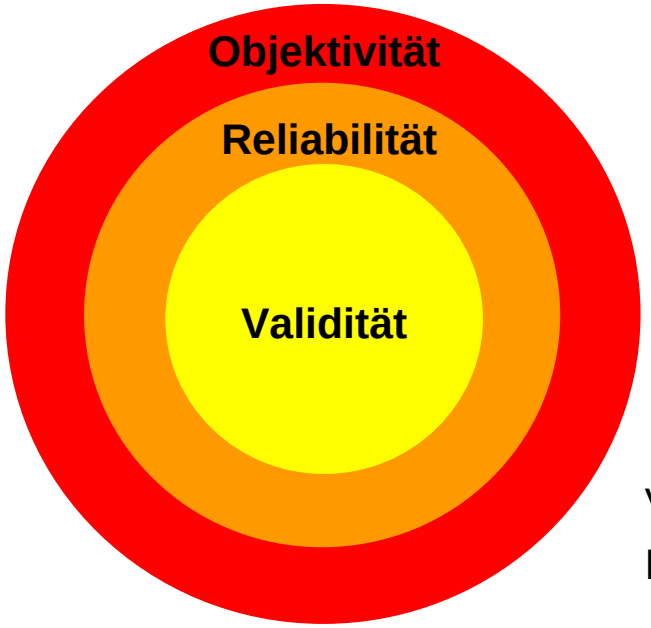
1. $X_1 = T_1 + E_1$
   Das Testergebnis setzt sich additiv aus dem wahren Wert („True Score“: T) und dem Messfehler („Error Score“: E) zusammen.
2. $\mu (E_1) = 0$ bzw. $\mu (X_1) = T_1$
   Bei Messwiederholung heben sich die Fehler gegenseitig auf.
3. $\rho (T_1, E_1) = 0$
   Die Größe des Messfehlers ist unabhängig vom Ausprägungsgrad des untersuchten Merkmals.
4. $\rho (T_2, E_1) = 0$
   Die Größe des Messfehlers ist unabhängig vom Ausprägungsgrad eines anderen Merkmals.
5. $\rho (E_{11}, E_{12}) = 0$
   Messfehler von verschiedenen Testanwendungen sind voneinander unabhängig.

Unabhängigkeit der Messfehler

Hauptgütekriterien (I)

- Objektivität: Grad, indem die Ergebnisse unabhängig vom Untersucher sind
- Reliabilität: Grad der Genauigkeit, mit dem ein Merkmal gemessen wird, unabhängig davon, ob er dieses Merkmal messen soll
- Validität: Grad der Genauigkeit, mit dem das Merkmal gemessen wird, das gemessen werden soll
- Beispiel: Messung von „Intelligenz“ mit Fieberthermometer

„Hierarchie“ der Gütekriterien



Venn-Diagramm

Hauptgütekriterien (II)

- Definition der Güte eines Tests über Varianzanteile

- Objektivität:
- Reliabilität:

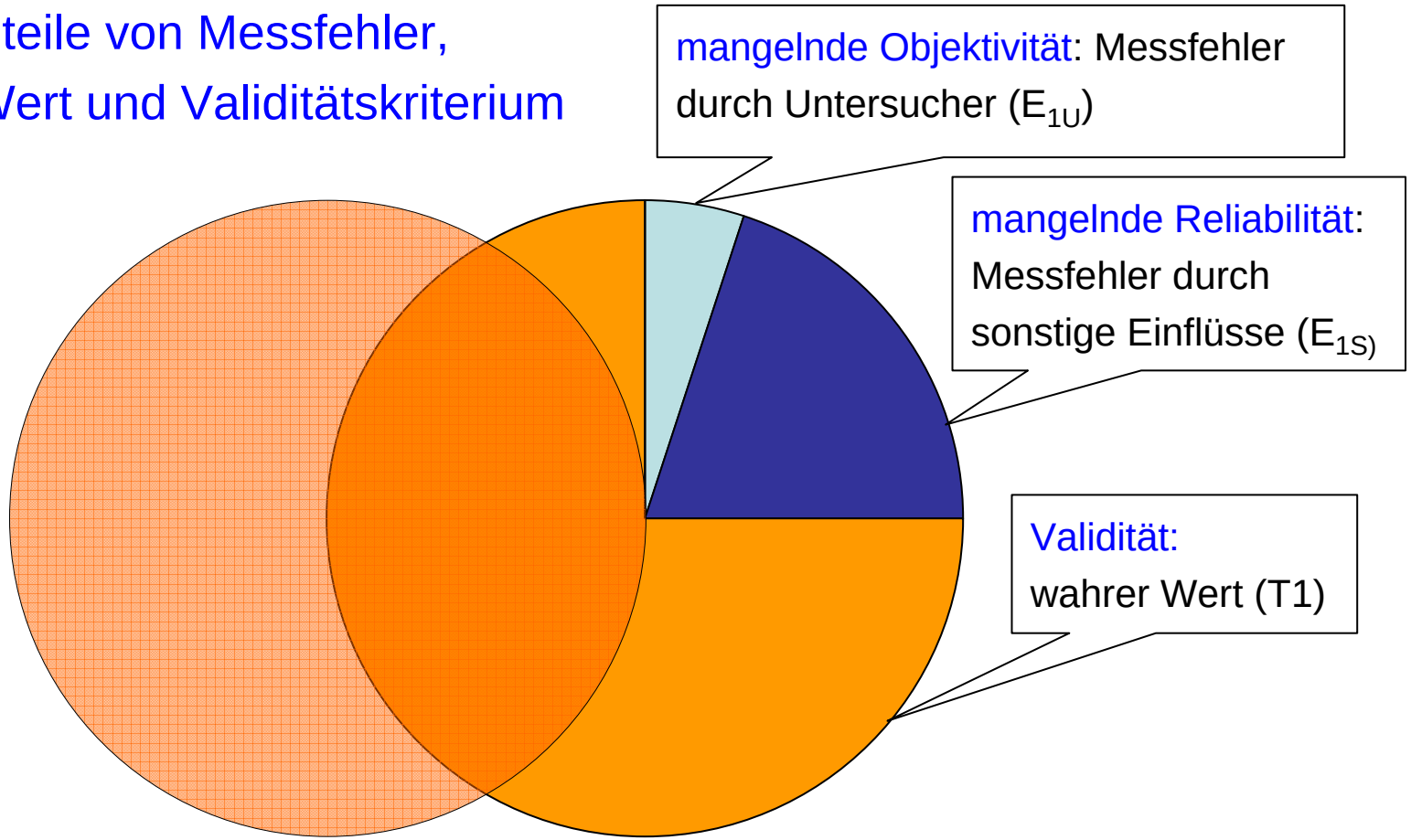
$$\mathrm{Rel}(X_1) = r_{tt}(X_1) = \frac{\sigma^2(T_1)}{\sigma^2(X_1)} = \frac{\sigma^2(T_1)}{\sigma^2(T_1) + \sigma^2(E_1)}$$

- Validität:

$$\mathrm{Val}(X_1) = r_{tc}^2(X_1) = \frac{\sigma(X_1, C)}{\sigma(X_1) \cdot \sigma(C)}$$

Hauptgütekriterien (III)

Varianzanteile von Messfehler,
wahrem Wert und Validitätskriterium



latentes Merkmal
z.B. „subjektive Lebensqualität“

Testergebnis
z.B. SF-36

Nebengütekriterien

- Normierung: Einordnung von Testergebnissen in ein Bezugssystem von Werten einer Vergleichspopulation
- Vergleichbarkeit: Vorliegen von anderen Messinstrumenten des gleichen Konstrukts
- Ökonomie: Aufwand an Zeit, Kosten, Material etc.
- Nützlichkeit: Relevanz des Konstrukts für die Praxis, Fehlen alternativer, besserer Messinstrumente

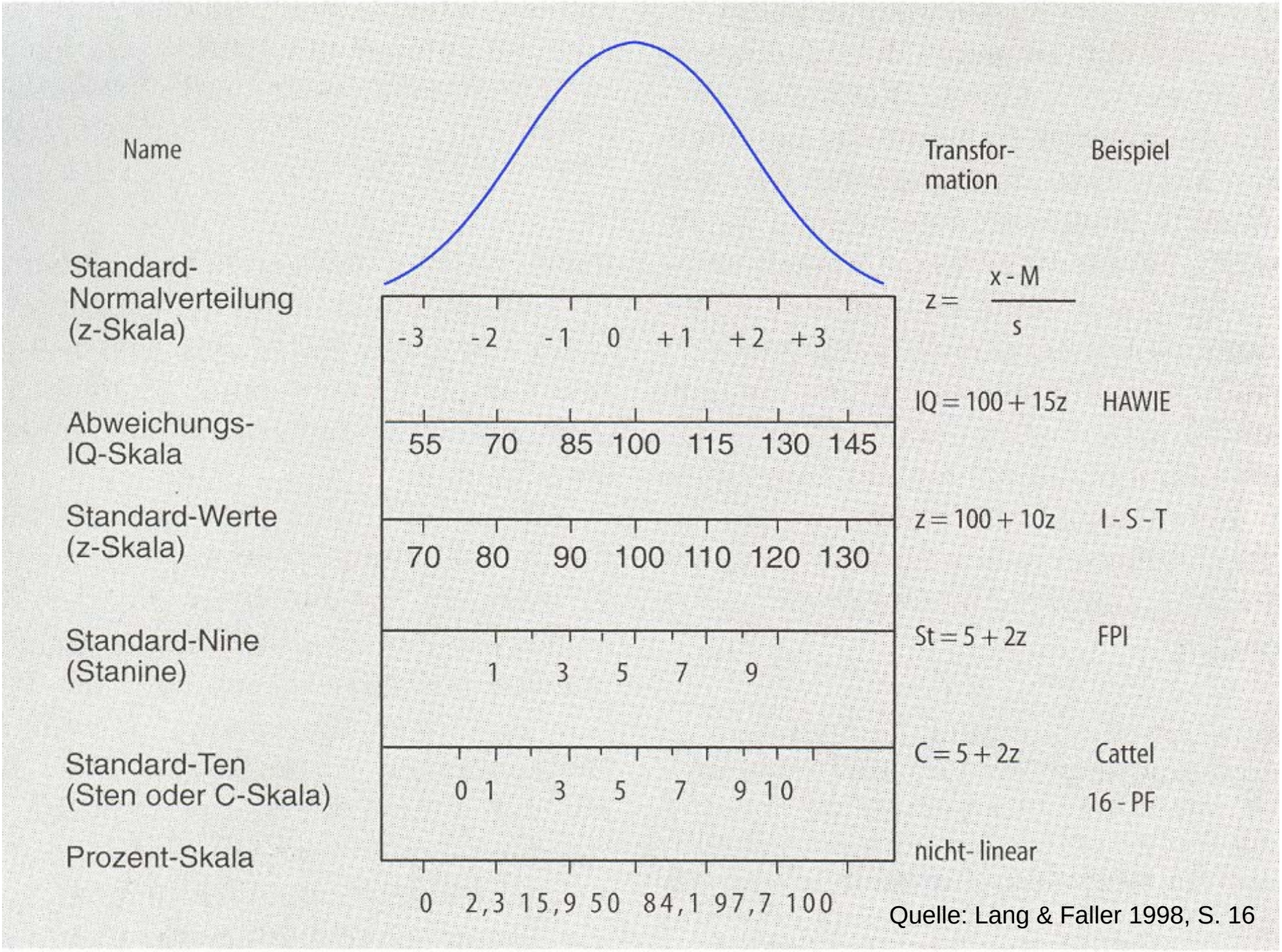
Normierung

- Vorgehen:
  - Erhebung von Testwerten an einer Vergleichsstichprobe (z. B. bevölkerungsrepräsentative Eichstichprobe)
  - Skalentransformation der Rohwerte
    - Z-Werte: $\mu = 100$, $\sigma = 10$
    - IQ-Skala: $\mu = 100$, $\sigma = 15$
    - T-Werte: $\mu = 50$, $\sigma = 10$
    - Stanine-Skala: $\mu = 5$, $\sigma = 2$
    - Prozentränge [0;100%]

allgemeine Formel:

$$X_T = \frac{\sigma}{s} \cdot X + \left(\mu - \frac{\sigma}{s} \overline{X} \right)$$

- Skalentransformation ermöglicht den Vergleich eines Individuums mit einer Vergleichsstichprobe unabhängig vom Schwierigkeitsgrad und der Aufgabenzahl eines Tests.



Quelle: Lang & Faller 1998, S. 16

III. Spezielle Methoden zur Bestimmung von Gütekriterien (Fragebogenevaluation)

1. Objektivität

Objektivität

- Kennwert: Korrelation, Intra-Klassen-Korrelation (ICC), Cohen‘s Kappa
- Bewertungsheuristik: „hoch“ ⇨ $r_{Obj} \sim 1$
- Rechenbeispiel:

	Test (Untersucher A)	Test (Untersucher B)
Proband 1	10	20
Proband 2	20	10
Proband 3	30	30
Proband 4	40	40
Proband 5	50	50

Inter-Rater-Reliabilität: $r_{Obj} = 0.90$

Objektivität

- Durchführungsobjektivität: Unabhängigkeit vom Untersucher
  Umsetzung: Standardisierung der Durchführung, Manualisierung

- Auswertungsobjektivität: Unabhängigkeit vom Auswerter
  Umsetzung: Leitfaden für Auswertung, Auswertungsschablonen/ -programm

- Interpretationsobjektivität: Unabhängigkeit vom Interpreter
  Umsetzung: genaue Beschreibung der Skalen, Normierung durch repräsentative Stichproben (z.B. Alter, Geschlecht, Bildung)

2. Reliabilität

Reliabilität

$$\text{Rel} = r_{tt} = \frac{\sigma^2(\text{T})}{\sigma^2(\text{X})} = \frac{\sigma^2(\text{T})}{\sigma^2(\text{T}) + \sigma^2(\text{E})}$$

- Synonyme: Zuverlässigkeit (*reliability*), Präzision (*precision*)
- Bedeutung: Anteil der Varianz der wahren Merkmalsausprägung an der gesamten Varianz der Testergebnisse
- Prinzip der Reliabilitätsberechnung: Korrelation zweier Messwerte erfasst deren gemeinsame „systematische Varianz“
- Bewertungsheuristik:
  - „hoch“: Rel = [0,90; 1,00]
  - „mittel“: Rel = [0,80; 0,90]
  - „niedrig“: Rel = [0,70; 0,80]
  - „unzureichend“: Rel < 0,70

Reliabilitätsarten

- Test-Retest-Reliabilität:
  Übereinstimmung von wiederholten Messungen
  (nur bei zeitlich stabilen Merkmalen!)

- Paralleltest-Reliabilität:
  Übereinstimmung von Messungen mit „parallelen“ Tests

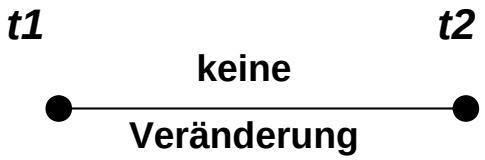
- Testhalbierungs-Reliabilität:
  Übereinstimmung der Testergebnisse von zwei zufällig gewählten Testhälften

- Interne Konsistenz (meist Cronbachs Alpha):
  Übereinstimmung von Fragen einer (homogenen) Skala

Test-Retest-Reliabilität

$$Rel_{Retest} = r_{t_1,t_2} = \frac{\sigma(t_1, t_2)}{\sigma(t_1) \cdot \sigma(t_2)}$$

- Studiendesign: zweimalige Vorgabe des Tests an einer Stichprobe in angemessenem zeitlichen Abstand



- Voraussetzung:
  - zeitliche Stabilität des Merkmals
  - Vernachlässigung von Erinnerungseffekten

- Nachteile:
  - Überschätzung der Reliabilität z.B. durch Erinnerung
  - Unterschätzung der Reliabilität bei zeitabhängigen Merkmalen; Konfundierung von wahren Veränderungen und mangelnder Reliabilität
  - zeitlicher und untersuchungstechnischer Aufwand (Drop-Out!)

Paralleltest-Reliabilität

$$\mathrm{Rel}_{Par} = r_{t_A,t_B} = \frac{\sigma(t_A,t_B)}{\sigma(t_A)\cdot\sigma(t_B)}$$

- Studiendesign: Vorgabe von zwei parallelen Testversionen in kurzer Folge in einer Sitzung an einer Stichprobe
- Vorteile:
  - Verringerung von Unterschleif durch parallele Versionen (A, B)
  - Gütekriterien gelten für beide Tests
- Nachteil: Aufwand zur Entwicklung von zwei parallelen Test-Versionen
- Konstruktion von parallelen Tests:
  - Itempool: Sammlung von Items
  - Itemanalyse: Ermittlung von Schwierigkeit und Trennschärfe an einer Eichstichprobe S1
  - Item-Zwillinge: gleiche Schwierigkeit, gleiche Trennschärfe
  - Paralleltests: zufällige Zuordnung der „Item-Zwillinge“
  - Evaluierung der Reliabilität an neuer (!) Stichprobe S2

Split-Half-Reliabilität

$$\text{Rel}_{\text{Split-Half}} = r_{t_{1/2},t_{2/2}} = \frac{\sigma(t_{1/2},t_{2/2})}{\sigma(t_{1/2}) \cdot \sigma(t_{2/2})}$$

- Studiendesign: einmalige Vorgabe des Tests an einer Stichprobe
- Durchführung: Einteilung der Testitems zu zwei Testhälften nach bestimmten Prinzip (Zufall, Reihenfolge, Odd-Even,...)
- Sonderform der Paralleltest-Reliabilität
- Vorteil: geringer untersuchungstechnischer Aufwand
- Nachteil: Unterschätzung der Reliabilität durch reduzierte Testlänge (Korrektur durch Spearman-Brown-Prophecy-Formel)

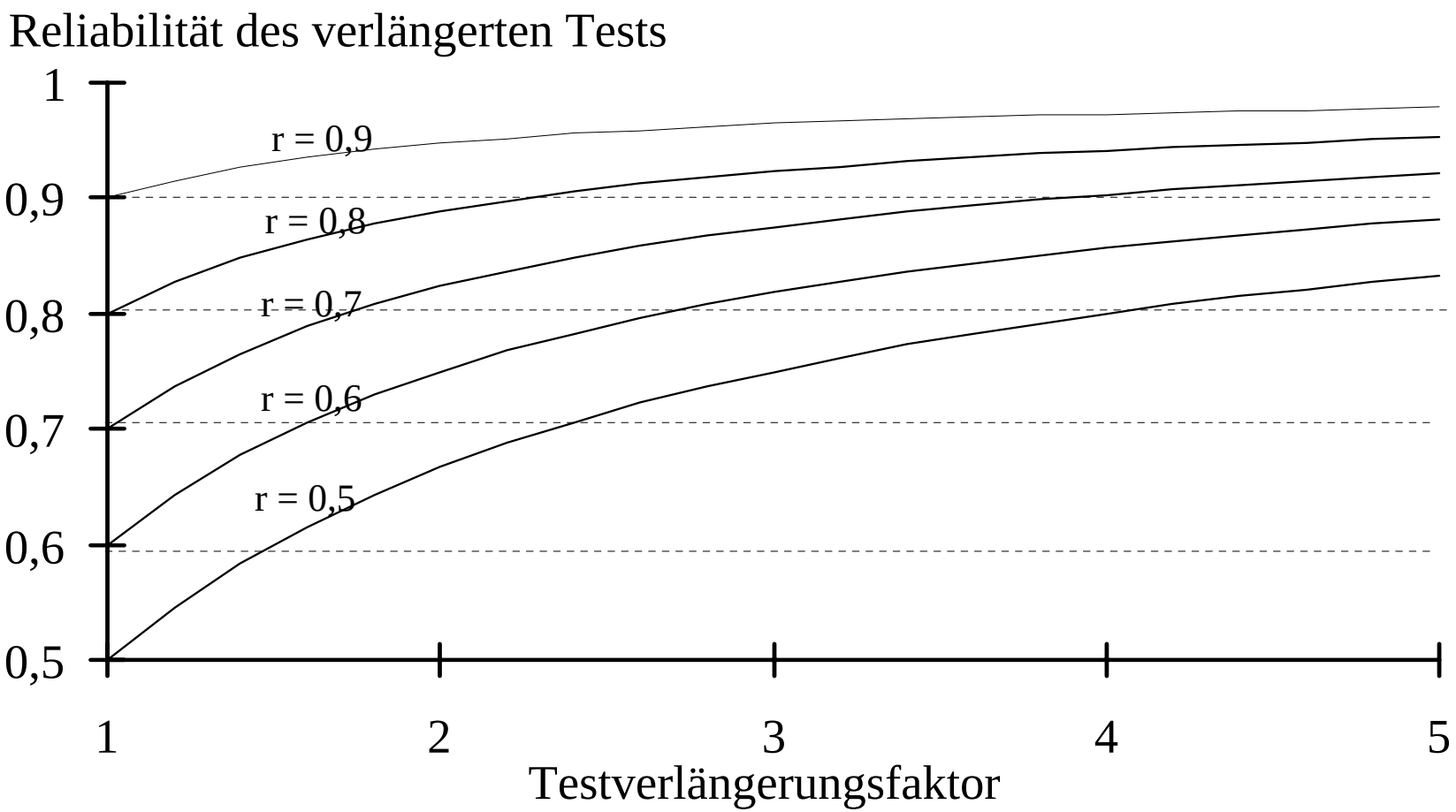
Exkurs: Testlänge und Gütekriterien (I)

- Reliabilität ~ Testlänge (Itemzahl)
- Testhalbierung führt zur Unterschätzung der Reliabilität
- Korrektur durch Spearman-Brown-Prophecy-Formel:

$$Rel_{korr} = \frac{2 \cdot Rel_{Split-Half}}{1 + Rel_{Split-Half}}$$

- Erhöhung der Testlänge kann in Sonderfällen zur Verringerung der Gütekriterien führen (vgl. Yousfi 2005a)

Exkurs: Testlänge und Gütekriterien (II)



Rost 2004, S. 383

Interne Konsistenz

$$\alpha = \frac{p}{p-1}\left(1 - \frac{\sum_{i=1}^{p} s^2_{\text{Item}}}{s^2_{\text{Testwert}}}\right)$$

- Studiendesign: einmalige Vorgabe an einer Stichprobe
- Prinzip: Unterteilung in soviele „Sub-Tests“ wie Items, Erweiterung der Testhalbierungsreliabilität
- Kennwerte:
  - Formel von Kuder und Richardson (1939)
  - Alpha-Koeffizient von Cronbach (1951)
- Cronbachs Alpha
  - Bedeutung: mittlere Testhalbierungsreliabilität (Korrelation) über alle möglichen Testhälften (eigtl. Homogenitäts-Index)
  - Anwendbarkeit: dichotome und polytome Items
- Vorteil: geringer Aufwand, stabilere Schätzung als Testhalbierung
- Nachteil: Unterschätzung der Reliabilität bei heterogenen bzw. mehrdimensionalen Tests

3. Validität

Validität

- wichtigstes Testgütekriterium
- Synonyme: Gültigkeit (*validity*), Unverzerrtheit (*unbiasedness*), Richtigkeit (*accuracy*)
- Bedeutung: Anteil der gemeinsamen Varianz des Testergebnisses mit Zielkonstrukt
- Kennwert: Korrelationskoeffizient, η^2 (eta-quadrat), Kontingenzkoeffizient C
- Bewertungsheuristik:
  - „hoch“: $r_{tc} > 0{,}60$
  - „mittel“: $r_{tc} = [0{,}40; 0{,}60]$
  - „niedrig“: $r_{tc} < 0{,}40$
  - „bedeutsam“: $r_{tc} >> 0$ (H_0: $r_{tc} = 0$, $p < 0{,}05$)

Inhaltsvalidität

- Synonyme: Augenscheinvalidität (face validity), Logische Validität

- Inhalt eines Tests erfasst die wichtigsten Aspekte des Konstrukts

- Beispiele:
  - einfache motorische, sensorische Tests (z.B. Tastaturschreiben)
  - Erfassung der Rechenfähigkeit mit Grundrechenarten
  - Erfassung von körperlichem Wohlbefinden mit der Frage „Haben Sie Schmerzen?“

Kriteriumsvalidität

$$\text{Val} = r_{tc} = \frac{\sigma(T,C)}{\sigma_T \cdot \sigma_C}$$

- Übereinstimmung mit einem beobachtbaren Indikator des Zielkonstrukts
- Beispiel: Berufseignungstest und Jahreseinkommen in fünf Jahren als Indikator für beruflichen Erfolg

- Kennwert: Korrelation zwischen Testergebnis und Kriterium

- Probleme:
  - häufig kein adäquates Außenkriterium vorhanden, z.B. Messung von Religiosität über die Häufigkeit des Kirchgangs
  - Schwierigkeit der Interpretation der Validität bei Operationalisierung über unreliables, unvalides Außenkriterium

Arten der Kriteriumsvalidität

- konkurrente Validität: gleichzeitige Messung von Testergebnis und Kriterium

- prognostische Validität: Messung des Testergebnisses vor dem Kriterium
  (Prädiktor soll Kriterium vorhersagen)
  Beispiel: Psychotizismus-Fragebogen ⇨psychiatrische Diagnose

- prädiktive Validität: Messung des Testergebnisses vor dem Kriterium
  (Prädiktor repräsentiert Teil des Kriteriums)
  Beispiel: Schuleingangstest ⇨ Abiturnote

- diskriminative Validität: Definition des Kriteriums über die Zugehörigkeit zu einer bestimmten Gruppe (Kranke vs. Gesunde)

Konstruktvalidität

- Prüfung einer Struktur von Hypothesen, welche aus dem Zielkonstrukt abgeleitet werden können

- Beispiel: „subjektive Gesundheit“
  Zusammenhang von Skalen zur körperlichen Gesundheit unterschiedlicher Fragebogen sollten untereinander höher korrelieren als mit Skalen zur psychischen Gesundheit.

- elaborierte Prüfung der Konstruktvalidität mit Multitrait-Multimethod-Methode (konvergente und diskriminante Validität)

- Nachteil:
  - gut gesicherte Instrumente erforderlich
  - Hypothesen müssen bestätigt werden
  - Konfundierung der Gültigkeit der Hypothesen und der Validität des Kriteriums

Weitere Validitätsarten

- differentielle Validität: Gültigkeit des Testergebnisses in unterschiedlichen Gruppen (auch: vgl. diskriminative Validität)
- faktorielle Validität: Zusammenhang (Faktorladung) von Item und latentem Faktor (Faktorenanalyse)
- externe Validität: Verwendung eines validen Außenkriteriums
- interne Validität: Gültigkeit des Testmodells für Daten (Item-, Personenparameter)
- Validitätsbegriffe in experimentellen Untersuchungen!
  - externe Validität: Verallgemeinerbarkeit von Aussagen einer Untersuchung auf eine bestimmte Population von Individuen
  - interne Validität: Kausalbeziehung zwischen der Variation in den abhängigen Variablen (Ergebnis) und der Variation der unabhängigen Variablen (Behandlung) in einem Experiment

Multitrait-Multimethod-Ansatz (Campell & Fiske, 79)

- Multitrait: Verwendung mehrerer Konstrukte
- Multimethod: Verwendung mehrerer Erhebungsmethoden
- Vorgehen: systematische, regelgeleitete Prüfung der wechselseitigen Beziehungen zwischen Konstrukt und Methode
- konvergente Validität: Grad der Übereinstimmung mehrerer Methoden zur Erfassung des selben Konstrukts
  Beispiel:
  Ratingsskala zur Gesundheit ~ mehrdimensionaler Fragebogen
- diskriminante Validität: Grad der Differenzierung zwischen Zielkonstrukt und anderen Konstrukten
  Beispiel: Fragebogen zur Depressivität, zugleich mit Fragebogen zur Ängstlichkeit und Einsamkeit
- Ergebnisse: MTMM-Matrix der Zusammenhangsmaßen.

Datensatz „Subjektive Gesundheit“ (IRES-2)

	Patient			Arzt		
Patient	Somat. Status	Psychosoz. Status	Funktion. Status	Somat. Status	Psychosoz. Status	Funktion. Status
1	4	7	6	6	7	4
2	5	5	3	3	6	4
...	...	...	...	...	...	...
n	6	7	4	4	7	6

Monotrait-Heteromethod-Block

		Patient			Arzt		
		SOMA	PSYCH	FUNK	SOMA	PSYCH	FUNK
Patient	SOMA						
	PSYCH						
	FUNK						
Arzt	SOMA	0,63					
	PSYCH		0,83				
	FUNK			0,58			

Heterotrait-Monomethod-Block

		Patient			Arzt		
		SOMA	PSYCH	FUNK	SOMA	PSYCH	FUNK
Patient	SOMA						
	PSYCH	0,44					
	FUNK	0,55	0,52				
Arzt	SOMA						
	PSYCH				0,41		
	FUNK				0,64	0,51	

Heterotrait-Heteromethod-Block

		Patient			Arzt		
		SOMA	PSYCH	FUNK	SOMA	PSYCH	FUNK
Patient	SOMA						
	PSYCH						
	FUNK						
Arzt	SOMA		0,19	0,42			
	PSYCH	0,14		0,37			
	FUNK	0,29	0,29				

Vollständige MTMM-Matrix

		Patient			Arzt		
		SOMA	PSYCH	FUNK	SOMA	PSYCH	FUNK
Patient	SOMA	1					
	PSYCH	0,44	1				
	FUNK	0,55	0,52	1			
Arzt	SOMA	0,63	0,19	0,42	1		
	PSYCH	0,14	0,83	0,37	0,41	1	
	FUNK	0,29	0,29	0,58	0,64	0,51	1

Kriterien

Konvergente Validität:

Kriterium 1: Korrelationen des **Monotrait-Heteromethod-Blocks** (1=rot) > 0

Diskriminante Validität:

Kriterium 2: Korrelationen des **Heterotrait-Monomethod-Blocks** (2=grün) <
Korrelationen des **Monotrait-Heteromethod-Blocks** (1=rot)

Kriterium 3: Korrelationen des **Heterotrait-Heteromethod-Blocks** (3=orange) <
Korrelationen des **Heterotrait-Monomethod-Blocks** (2=grün)

Konstruktvalidität:

Kriterium 4: Rangreihe der Trait-Interkorrelationen ist identisch in allen Teilmatrizen

Probleme von Validierungsstudien

- Nachweis der Validität ist schwierig aufgrund von methodischen und theoretischen Einschränkungen (Latenz der Konstrukte!)

- MTMM-Analysen sind aufwendig und werden selten durchgeführt

- reduzierte MTMM-Analyse: statt verschiedener Methoden werden nur verschiedene Indikatoren für das selbe Konstrukt erhoben

- Validität kann nicht endgültig geklärt werden.
  Beispiel: Patient und Arzt setzen somatisches Befinden mit Schmerzfreiheit gleich.

- pragmatisches Kriterium der Validität:
  bessere Vorhersagen bei Anwendung des Tests

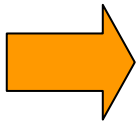
4. Änderungssensitivität

Hintergrund

- Anwendungsziele von klinischen Messinstrumenten (Kirshner & Guyatt, 1985):
  - Diskrimination: Messung stabiler Konstrukte z.B. Intelligenz
  - Prädiktion: Vorhersage von Werten z.B. Mortalität
  - Evaluation: Beurteilung von Veränderungen z.B. Reha-Erfolg

- Messinstrumente zur Diskrimination oder Prädiktion: Klassische Gütekriterien (Objektivität, Reliabilität, Validität) sind ausreichend

- Messinstrumente zur Evaluation: Mängel von klassischen Gütekriterien bzw. Kennwerten (z.B. Test-Retest-Reliabilität)



Kennwerte der Änderungsensitivität notwendig!

Änderungssensitivität

- Definition „Änderungssensitivität“ (engl. *sensitivity to change*):
  „Änderungssensitivität bezeichnet die Eigenschaft eines Messinstruments, ‚wahre‘ Veränderungen eines (psychologischen) Konstrukts abzubilden.“
  (vgl. auch Responsivität, longitudinale Validität)

Beispiel:

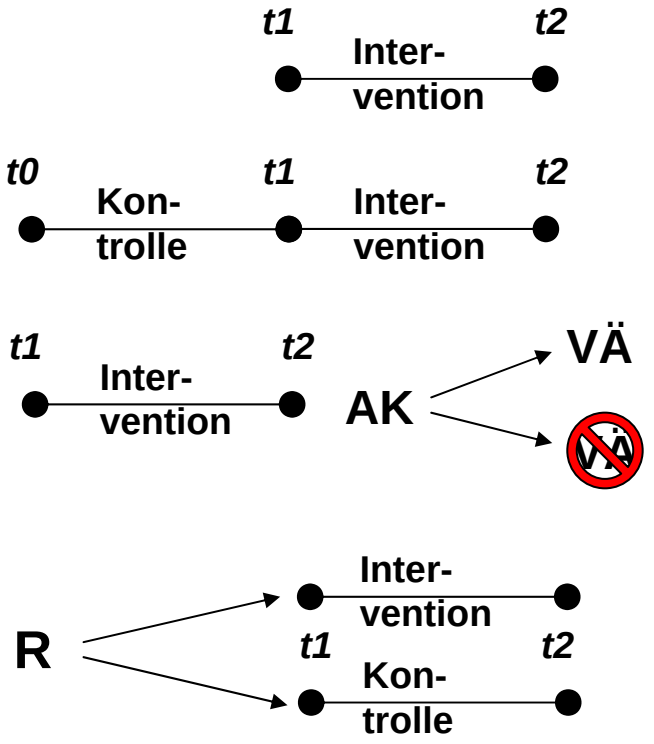
Frage: „Haben Sie jemals versucht, sich das Leben zu nehmen?“

Antwort: ☐ Ja ☐ Nein

- Frage ist geeignet, Gruppen von psychisch unterschiedlich belasteten Personen zu identifizieren (Diskrimination)
- Frage ist ungeeignet, Veränderungen aufgrund einer Intervention zu messen (Evaluation)

Studiendesigns

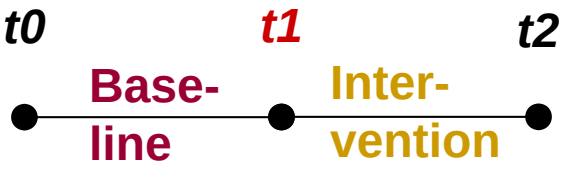
- Designs zur Ermittlung der Änderungssensitivität:
  - Ein-Gruppen-Designs
    - Einfaches Prä-Post-Design
    - Prä-Post-Design mit Baseline
    - Einfaches Prä-Post-Design mit Außenkriterium
  - Zwei-Gruppen-Design (mit Experimental- und Kontrollgruppe)



Kennwerte

- t-Tests und Varianzanalysen (t-Werte, F-Werte)
- Korrelationen mit veränderlichen Außenkriterien
- Anteil von Patienten mit reliabler Veränderung
- Effektstärken
  - nicht abhängig von Stichprobengröße
  - häufig eingesetzt
  - Interpretation ist methodenabhängig
- Norman-Koeffizienten
  - analog zu Intraklassen-Koeffizienten
  - Berücksichtigung von Varianz- und Kovarianzanteilen
  - bessere Interpretierbarkeit (Bereich von 0 bis 1)
  - selten benutzt

Effektstärken



$$SES = \frac{M(t2) - M(t1)}{SD[x(t1)]}$$

- SD der Messwerte x(t1) der Ausgangslage
- hohe Werte bei kleiner Ausgangsvariabilität

$$SRM = \frac{M(t2) - M(t1)}{SD[x(t2) - x(t1)]}$$

- SD der Differenz x(t2) - x (t1) der Interventionsphase
- hohe Werte bei kleiner Variabilität der Differenzen x(t2) - x (t1)

$$GRI = \frac{M(t2) - M(t1)}{SD[x(t1) - x(t0)]}$$

- SD der Differenz x(t1) - x(t0) der Baselinephase
- hohe Werte bei kleiner Variabilität der Differenzen x(t1) - x (t0)
- Berücksichtigung von Veränderung und Stabilität!

IV. Beispiel für Validierungsstudie

Quelle**:** Wollmerstedt N et al. (2005). Reliabilitäts-, Validitäts- und Änderungs-sensitivitätsprüfung des Funktionsfragebogen Bewegungsapparat (SMFA-D) in der stationären Rehabilitation von Patienten mit konservativ behandelter Rheumatoider Arthritis. Aktuelle Rheumatologie, 30 (4), 215-222

SMFA-D

- Short Musculoskeletal Function Assessment Questionnaire – deutsche Version
- Konstrukt: Funktionszustand des Bewegungsapparats (untere und obere Extremitäten)
- 46 Items
- 2 Skalen
  - Funktionsindex:
    tägliche Aktivitäten (10 Items), emotionaler Zustand (7 Items), Mobilität (9 Items), Arm-Hand-Funktion (8 Items)
  - Beeinträchtigungsindex:
    12 Items zu Arbeit, Familie, Hobbies, Freizeit, Ruhe, Schlaf
- Wertebereich der Skalen: 0 bis 100

Stichprobenbeschreibung

- Stichprobengröße: N=56
- Diagnose: Rheum. Arth. (kons.)
- Geschlecht: 33 w, 23 m
- Alter: m=49 Jahre
- Sprachkenntnisse: deutsch
- soziodemographische Variablen
- Einschluss-/Ausschlusskriterien

Tab. 1 Soziodemografische Standarddaten gemäß den Empfehlungen des Robert Koch-Instituts [1]

		Anzahl	*Prozent*
Familienstand	ledig:	5	9%
	verheiratet bzw. Partner:	45	80,0%
	geschieden oder getrennt:	5	9%
	verwitwet:	1	2%
Haushalt	1-Personen-Haushalt:	6	11%
	mit Partner oder mehr:	50	89%
Schulabschluss	Hauptschulabschluss:	44	79%
	Realschulabschluss:	6	11%
	Fachhochschule:	3	6%
	anderer Abschluss:	3	6%
Erwerbstätigkeit	berufstätig:	36	64%
	Hausfrau/-mann:	7	13%
	arbeitslos:	6	11%
	anderes:	3	5%
	Berufsunfähigkeitsrente:	3	5%
	keine Angaben:	1	2%
Berufsausbildung	Lehre:	34	61%
	Fachschule:	2	4%
	Fachhochschule:	2	4%
	anderes:	6	11%
	keine Berufsausbildung:	12	21%
Berufsgruppe	Arbeiter:	43	77%
	Angestellte:	7	13%
	Beamte:	1	2%
	selbstständig:	2	4%
	sonstiges:	2	4%
	keine Angaben:	1	2%

Studiendesign und Messinstrumente

- Studiendesign:
  - t0: 1 Woche vor Reha-Beginn
  - t1: Reha-Beginn
  - t2: Reha-Ende
  - t3: Katamnese (3 Monate nach Reha-Beginn)

- Behandlung: stationäre orthopädische/rheumatologische Rehabilitation

- Messinstrumente:
  - SMFA-D
  - SF-36 (Short-Form Health Survey 36)
  - HAQ (Health Assessment Questionnaire)
  - FFbH-P (Funktionsfragebogen Hannover Polyarthritis)
  - Patientenratings
  - Arztratings

Ergebnisse

- Rohwertverteilung

- Reliabilität
  - interne Konsistenz: Cronbachs Alpha
  - Retest-Reliabilität: Intra-Klassen-Korrelationskoeffizient (ICC)

- Validität:
  - Konstruktvalidität: Korrelationen (SMFA-D, FFbH-P)
  - Kriteriumsvalidität: Korrelationen (SMFA-D, Arzt-/Patientenurteil)

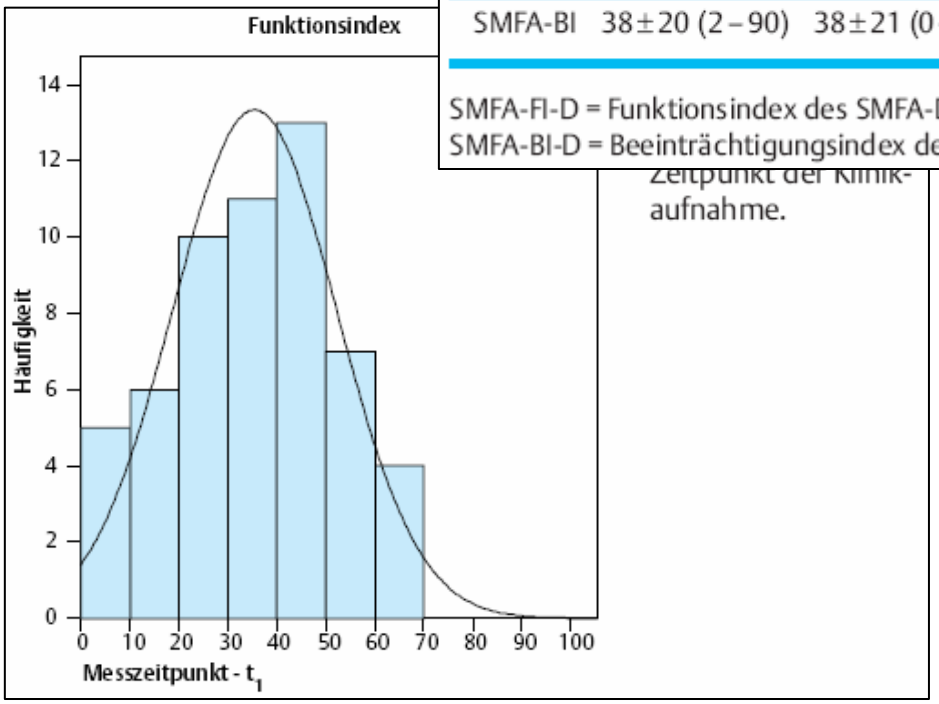
- Änderungssensitivität: Effektstärken

Rohwertverteilung

Tab. **2** Verteilungskennwerte bei konservativ behandelten Rheumapatienten in der stationären Rehabilitation zu drei Messzeitpunkten

	7–10 Tage vor Aufnahme *n = 29* *MW ± SD (Bereich)*	*Aufnahme* *n = 56* *MW ± SD (Bereich)*	*Entlassung* *n = 54* *MW ± SD (Bereich)*	*3 Monate* *n = 43* *MW ± SD (Bereich)*
SMFA-FI	36±18 (4–67)	35±17 (4–70)	31±16 (1–73)	40±19 (0–78)
SMFA-BI	38±20 (2–90)	38±21 (0–		

SMFA-FI-D = Funktionsindex des SMFA-D
SMFA-BI-D = Beeinträchtigungsindex des
Zeitpunkt der Klinikaufnahme.



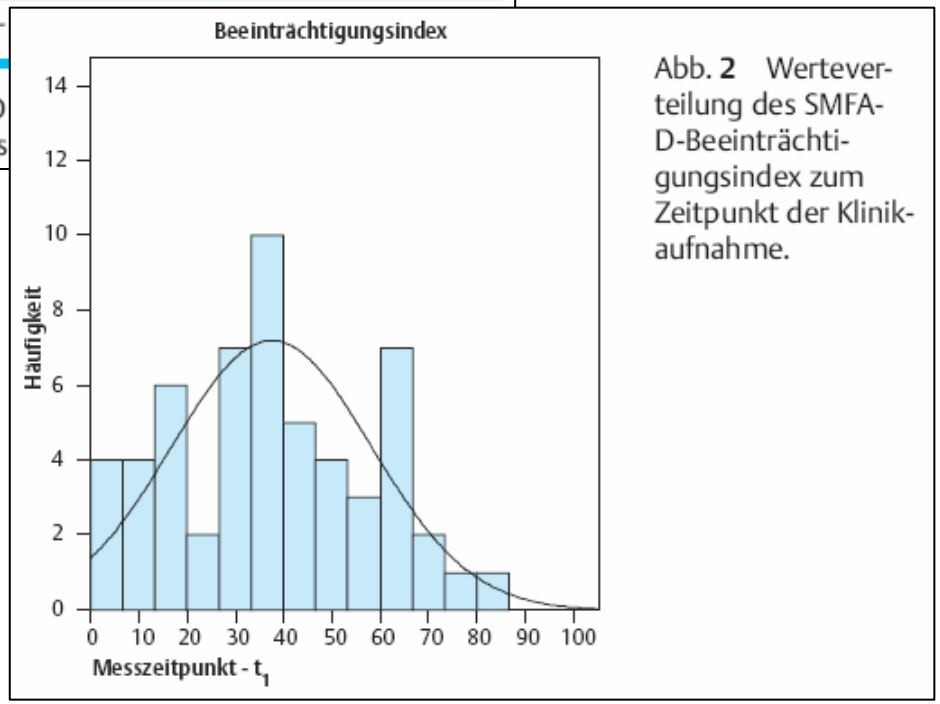


Abb. **2** Wertevertteilung des SMFA-D-Beeinträchtigungsindex zum Zeitpunkt der Klinikaufnahme.

Reliabilität

- Interne Konsistenz: Cronbachs Alpha = [0,93;0,98]

Tab. 3 Interne Konsistenz, Cronbachs α zu vier Messzeitpunkten

	7–10 Tage prästationär	*Klinikauf-nahme*	*Klinikent-lassung*	*3 Monate nach Klinikaufnahme*
SMFA-FI	0,97	0,97	0,97	0,98
SMFA-BI	0,93	0,94	0,94	0,95

SMFA-FI = Funktionsindex des SMFA-D
SMFA-BI = Beeinträchtigungsindex des SMFA-D

- Retest-Reliabilität:
  - ICC (FI) = 0,93
  - ICC (BI) = 0,87

Konstrukt- und Kriteriumsvalidität

Tab. **4** Zusammenhänge zwischen den Skalen des SMFA-D und HAQ, FFbH-P, SF-36, Gehgeschwindigkeit, Schmerzbeurteilung sowie Funktionsbeurteilung

		Aufnahme		*Entlassung*		*3-Monats-Befragung*	
	Skala	*SMFA-FI*	*SMFA-BI*	*SMFA-FI*	*SMFA-BI*	*SMFA-FI*	*SMFA-BI*
Konstruktvalidität	HAQ	0,84[1]	0,75[1]	0,82[1]	0,64[1]	0,85[1]	0,73[1]
	FFBH-P	−0,86[1]	−0,80[1]	−0,84[1]	−0,72[1]	−0,75[1]	−0,69[1]
	SF-36 – körperliche Funktionsfähigkeit	−0,73[1]	−0,70[1]	−0,84[1]	−0,69[1]	−0,83[1]	−0,76[1]
	SF-36 – körperliche Rollenfunktion	−0,59[1, 2]	−0,64[1, 2]	−0,68[1, 2]	−0,67[1, 2]	−0,67[1, 2]	−0,74[1, 2]
	SF-36 – körperliche Schmerzen	−0,74[1]	−0,74[1]	−0,80[1]	−0,77[1]	−0,74[1, 2]	−0,76[1, 2]
	SF-36 – allg. Gesundheitswahrnehmung	−0,64[1]	−0,64[1]	−0,51[1]	−0,49[1]	−0,62[1]	−0,68[1]
	SF-36– Vitalität	−0,62[1]	−0,68[1]	−0,66[1]	−0,65[1]	−0,62[1]	−0,67[1]
	SF-36 – soziale Funktionsfähigkeit	−0,68[1]	−0,73[1]	−0,62[1, 2]	−0,64[1, 2]	−0,48[1]	−0,67[1]
	SF-36 – emotionale Rollenfunktion	−0,40[1, 2]	−0,54[1, 2]	−0,60[1, 2]	−0,67[1, 2]	−0,38[1, 3]	−0,57[1, 2]
	SF-36 – psychisches Wohlbefinden	−0,54[1]	−0,62[1]	−0,49[1]	−0,59[1]	−0,35[3]	−0,49[1]
	SF-36 – körperliche Summenskala	−0,74[1]	−0,68[1]	−0,80[1]	−0,68[1]	−0,81[1]	−0,71[1]
	SF-36 – psychische Summenskala	−0,45[1]	−0,56[1]	−0,49[1]	−0,61[1]	−0,27 n. s.	−0,49[1]
Kriteriumsvalidität[1]	Schmerzbeurteilung – Patient	0,62[1]	0,67[1]	0,75[1]	0,57[1]	0,67[1]	0,64[1]
	Funktionsbeurteilung – Patient	0,62[1]	0,71[1]	0,71[1]	0,59[1]	0,70[1]	0,67[1]
	allg. Gesundheitsbeurteilung – Patient	0,43[1]	0,53[1]	0,60[1]	0,50[1]	0,60[1]	0,62[1]
	Symptomatikbeurteilung – Arzt	0,56[1]	0,45[1]	0,62[1]	0,44[1]		
	Funktionsbeurteilung – Arzt	0,50[1]	0,35[3]	0,52[1]	0,41[1]		

[1] $p < 0,01$
[2] Spearman-Rho;
SMFA-FI = Funktionsindex des SMFA-D

Diskriminative Validität

- Unterscheidung zwischen konservativ und operativ behandelten Patienten mit rheumatoider Arthritis

Tab. **5** Known-groups validity – Unterscheidung von konservativ und operativ versorgten Patienten mit Rheumatoider Arthritis anhand des Funktions- und Beeinträchtigungsindex

	Behandlungsgruppe	*n*	*MW*	*SD*	*p (2-seitig)*
t_1 – SMFA – Funktionsscore	konservativ	56	35	17	< 0,001
	operativ	55	48	18	
t_1 – SMFA – Beeinträchtigungsscore	konservativ	56	38	21	0,008
	operativ	55	47	18	

Änderungssensitivität

- Effektstärken:
  - SES: Standardized Effect Size
  - SRM: Standardized Response Mean

- Veränderungszeiträume:
  - Reha-Beginn bis Reha-Ende
  - Reha-Beginn bis Katamnese

Tab. 6 Änderungssensitivität zur Entlassung aus der Klinik und drei Monate nach Klinikaufnahme

	Effektstärken Aufnahme vs. Entlassung		*Effektstärken Aufnahme vs. nach 3 Monaten*	
	SES[1]	*SRM*[2]	*SES*[1]	*SRM*[2]
SMFA-FI	0,25	0,53	−0,32	−0,58
SMFA-BI	0,27	0,51	−0,14	−0,22
HAQ	0,19	0,28	−0,07	−0,08
Funktionsfragebogen Hannover	0,12	0,18	−0,35	−0,35
SF-36 – körperliche Funktionsfähigkeit	0,22	0,27	−0,31	−0,31
SF-36 – körperliche Rollenfunktion	0,34	0,41	0,00	−0,01
SF-36 – körperliche Schmerzen	0,50	0,79	0,17	0,19
SF-36 – allg. Gesundheitswahrnehmung	0,13	0,15	−0,17	−0,18
SF-36 – Vitalität	0,50	0,72	0,03	0,04
SF-36 – soziale Funktionsfähigkeit	0,32	0,37	0,03	0,04
SF-36 – emotionale Rollenfunktion	0,09	0,09	−0,03	−0,03
SF-36 – psychisches Wohlbefinden	0,47	0,69	−0,06	−0,06
SF-36 – körperliche Summenskala	0,29	0,43	−0,14	−0,16
SF-36 – psychische Summenskala	0,30	0,39	0,08	0,10

[1] Standardisierung der Mittelwertsdifferenz an der Standardabweichung der Prä-Werte (standardized effect size [SES])
[2] Standardisierung der Mittelwertdifferenz an der Standardabweichung der Prä-Post-Differenzen (standardized response mean [SRM])
SMFA-FI = Funktionsindex des SMFA-D
SMFA-BI = Beeinträchtigungsindex des SMFA-D

IV. Zusammenfassung

Zusammenfassung

- Fragebogen als vielseitiges Messinstrument auch in der Medizin
- Beurteilung der Güte von Fragebogen auf der Basis der Klassischen Testtheorie (Objektivität, Reliabilität, Validität, Änderungssensitivität) üblich
- Vielzahl von Methoden vorhanden
  - Objektivität (Durchführung, Auswertung, Interpretation)
  - Reliabilität (Retest, Paralleltest, Interne Konsistenz)
  - Validität (Inhalt, Kriterium, Konstrukt, Änderungssensitivität)
- Auswahl der Methode abhängig von Fragestellung (Stichprobe, Studiendesign, Kennwert,...)
- Messinstrument muss für andere Patientengruppen jeweils neu evaluiert werden!

Literatur

deutsch:

- Bortz, J. (2002). Forschungsmethoden und Evaluation (3. Auflage). Berlin: Springer
- Köbberling, J., Richter, K., Trampisch, H. J. & Windeler, J. (1991). Methodologie der medizinischen Diagnostik. Springer, Heidelberg.
- Lienert, G. & Raatz, U. (1994). Testaufbau und Testanalyse. Weinheim: PVU.
- Rost, J. (2004). Lehrbuch Testtheorie – Testkonstruktion. Bern: Huber.
- Steyer, R. & Eid, M. (2001). Messen und Testen (2. Auflage). Berlin: Springer.
- Yousfi, S. (2005a). Mythen und Paradoxien der klassischen Testtheorie (I) - Testlänge und Gütekriterien. Diagnostica, 51 (1), 1-11.
- Yousfi, S. (2005b). Mythen und Paradoxien der klassischen Testtheorie (II) - Trennschärfe und Gütekriterien. Diagnostica, 51 (2), 55-66.

Literatur

englisch:

- Aiken, L. R. (2000). Psychological Testing and Assessment (10th ed.). Boston: Allyn and Bacon
- Anastasi, A. & Urbina, S. (1997). Psychological Testing (7th ed.). Upper Saddle River, NJ: Prentice Hall.
- Kraemer, H. C. (1992). Evaluating medical tests. Newbury Park: Sage.
- Lord, F. M. & Novick, M. R. (1976). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Links

- Testzentrale des Hogrefe-Verlags (http://www.testzentrale.de/)

- Zentrum für Psychologische Information und Dokumentation - ZPID (www.zpid.de):
  - Datenbank Psyndex Tests (kostenpflichtig)
  - Testarchiv (frei erhältliche Tests als pdf-Dateien)
  - Testbibliotheken
  - Testanbieter

- Datenbank des IQPR (http://www.assessment-info.de/)

- Beispielartikel zur Fragebogenvalidierung des Funktionsfragebogen Bewegungsapparat SMFA-D (http://www.smfa-d.de/)

Vielen Dank

für Ihre Aufmerksamkeit!

Kontakt: wilmar.igl@mail.uni-wuerzburg.de