

# Grundlagen der Testtheorie

## WS 2020/21

1. Grundlagen & Gütekriterien

02.11.2020

Prof. Dr. Eunike Wetzel

# Lernziele

- Theoretischen und methodischen Grundlagen der Erfassung psychologischer Merkmale verstehen
- Psychologische Tests entwickeln und evaluieren können

# Semesterplan

Sitzung	Termin	Thema
1	02.11.	Grundlagen & Gütekriterien
2	09.11.	Schritte der Testkonstruktion: Übersicht Konstruktdefinition & Itemgenerierung
3	16.11.	Erstellung eines Testentwurfs
4	23.11.	Klassische Testtheorie
5	30.11.	Item Response Theorie
6	07.12.	Exploratorische Faktorenanalyse 1
7	14.12.	Exploratorische Faktorenanalyse 2

# Semesterplan

Sitzung	Termin	Thema
8	04.01.	Itemanalyse
9	11.01.	Itemselektion & Testrevision
10	18.01.	Objektivität
11	25.01.	Reliabilität
12	01.02.	Validität
13	08.02.	Normierung Standards für psychologisches Testen

# Literatur

- Moosbrugger, H. & Kelava, A. (2012). *Testtheorie und Fragebogenkonstruktion* (2. Aufl.). Heidelberg: Springer.
- Weitere Kapitel aus anderen Lehrbüchern
- Zu jeder Sitzung wird die prüfungsrelevante Literatur angegeben

# Übung zur VL

- Gehalten von Maria Jalynskij und Rebekka Kupffer
- Vertiefung der Inhalte der VL durch Durchführung von Übungen

# Grundlagen & Gütekriterien

# Grundlagen & Gütekriterien

- Grundlagen
  - Item und Konstrukt
  - Definition „psychologischer Test“
  - Was und wofür ist Testtheorie?
  - Ziele und Anwendungsbereiche psychologischer Tests
  - Arten psychologischer Tests
- Gütekriterien



# Item und Konstrukt

- **Item:** Aufgabe/Frage in einem Test
  - Die Antworten auf Items sind beobachtbar (manifest)

Ich mag es, im Voraus zu planen.

Starke Ablehnung	Ablehnung	Zustimmung	Starke Zustimmung
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

$$68 \quad \underline{\quad ? \quad} \quad 32 \quad \underline{\quad ? \quad} \quad 4 \quad = \quad 25$$

- **Konstrukt:** Eigenschaft, Fähigkeit, Merkmal
  - Konstrukte sind latente Variablen, da sie nicht direkt beobachtet werden können
  - Bsp.: Extraversion, Intelligenz

# Definition psychologischer Test

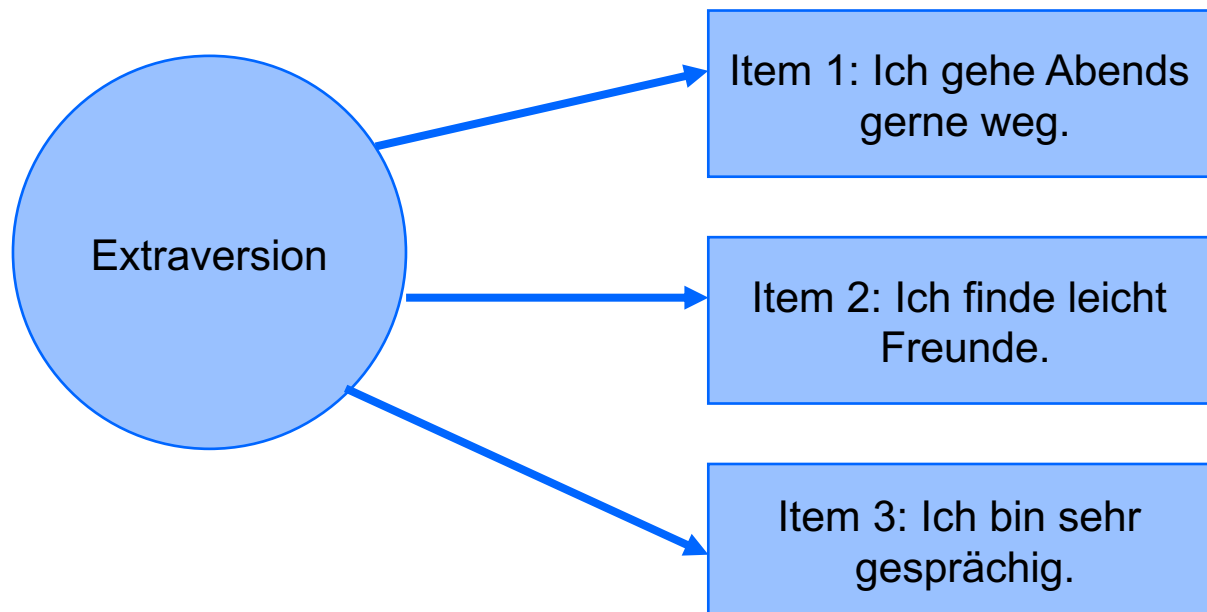
- Psychologische Tests erfassen Fähigkeiten, Eigenschaften, Fertigkeiten und Zustände einer Person
- Moosbrugger & Kelava (2012, S. 2):  
„Ein Test ist ein wissenschaftliches Routineverfahren zur Erfassung eines oder mehrerer empirisch abgrenzbarer psychologischer Merkmale mit dem Ziel einer möglichst genauen quantitativen Aussage über den Grad der individuellen Merkmalsausprägung.“

# „Ein Test ist ein...

- „wissenschaftliches“: Test muss Qualitätsstandards genügen
- „Routineverfahren“: standardisiertes Vorgehen
- „zur Erfassung eines oder mehrerer empirisch abgrenzbarer psychologischer Merkmale“: Erfassung klar definierter Konstrukte
- „mit dem Ziel einer möglichst genauen quantitativen Aussage über den Grad der individuellen Merkmalsausprägung“: Ausprägung des Merkmals (quantitativ) oder Vorhandensein/Art des Merkmals (qualitativ)

# Messung von Konstrukten

- Konstrukte äußern sich in beobachtbaren Verhalten
- In psychologischen Tests werden Items als beobachtbare Indikatoren des Konstrukts verwendet

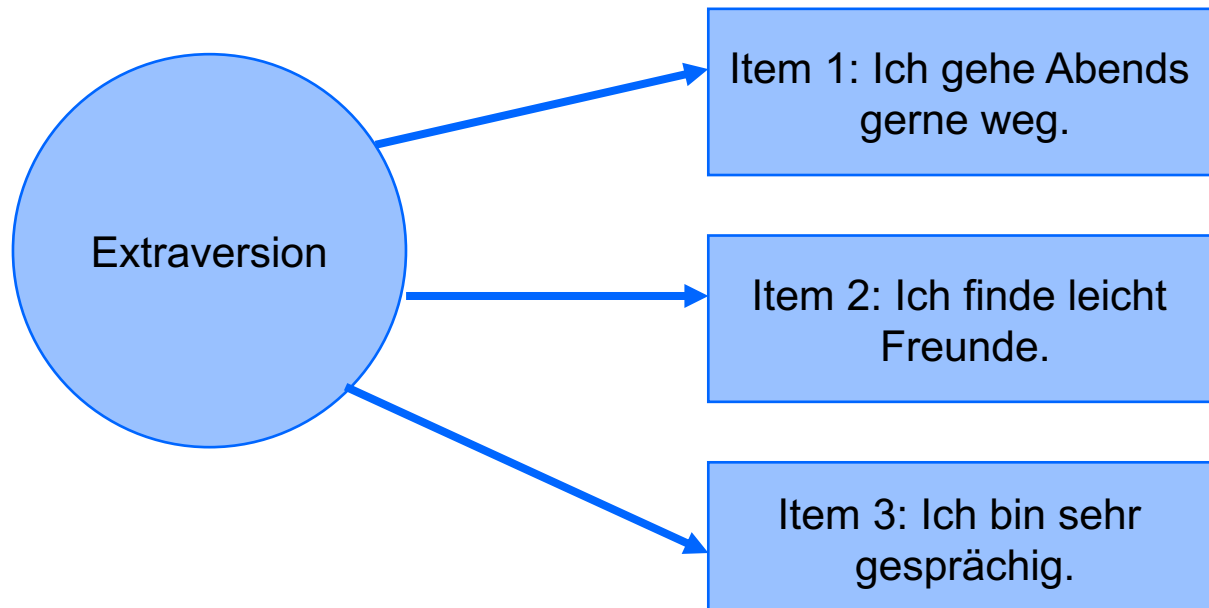


# Was ist Testtheorie?

- Testtheorie beschäftigt sich mit dem Zusammenhang zwischen dem Antwortverhalten im Test und dem zu erfassenden Konstrukt
- Theoretischer Hintergrund zur Konstruktion und Interpretation von Testverfahren

# Warum Testtheorie?

- Entspricht das Antwortverhalten direkt dem interessierenden Merkmal, benötigt man keine Testtheorie
- Bsp.: Treiben Sie regelmäßig Sport?
- Bei psychologischen Konstrukten ist eine Testtheorie vonnöten, da von dem Antwortverhalten im Test auf das latente Konstrukt geschlossen wird



# Ziele psychologischer Tests

- **Querschnittlich:**

- Position des Individuums innerhalb einer Gruppe feststellen
- Unterschiede in der Merkmalsausprägung zwischen Individuen/Gruppen erfassen
- Klassifikation: Feststellung des Vorhandenseins/Nicht-Vorhandenseins eines Merkmals oder einer über einem Kriterium liegenden Merkmalsausprägung
- Profil: Erfassung individueller Merkmalskombinationen (z.B. Persönlichkeitsprofil)

- **Längsschnittlich:**

- Erfassung von Merkmalsveränderungen über die Zeit (Verlauf)

# Anwendungsbereiche

Psychologische Tests werden eingesetzt

- Zur Diagnose psychischer Störungen
- In der Beratung (z.B. Erziehungsberatung)
- In der staatlichen Verwaltung (z.B. Berufsberatung, Verkehrseignung)
- Im forensischen Bereich
- In Unternehmen (z.B. Personalauswahl und -entwicklung)
- Im Pädagogischen Bereich (z.B. Schulreife, Intelligenzdiagnostik)
- In der Forschung
- ...



# Arten psychologischer Tests

- **Leistungstests**
  - Es gibt richtige und falsche Antworten/Lösungen
  - Ziel: maximal mögliches Verhalten erfassen
- **Psychometrische Persönlichkeitstests**
  - Es gibt keine richtigen und falschen Antworten
  - Ziel: typisches Verhalten erfassen
- **Persönlichkeitsentfaltungs-Verfahren/Projektive Verfahren**
  - Verfahren, keine Tests i.S. der eben vorgestellten Definition!
  - Sollen Projektionen hervorrufen, die dann Rückschlüsse über Einstellungen, Motive, Wünsche etc. erlauben

# Arten psychologischer Tests


- Leistungstests
  - Intelligenztests
  - Entwicklungstests
  - Eignungstests
  - Schultests
- Psychometrische Persönlichkeitstests
  - Persönlichkeitsstrukturtests
  - Einstellungstests
  - Interessentests
  - Klinische Tests
- Persönlichkeitsentfaltungs-Verfahren/Projektive Verfahren
  - Formdeutungsverfahren
  - Verbal-thematische Verfahren
  - Zeichnerische und Gestaltungsverfahren

# Beispiel Leistungstests

Konzentrationstest **d2** (Brickenkamp, 2002)

Aufgabe: jedes d mit 2 Strichen durchstreichen

„Arbeiten Sie so schnell wie möglich – aber natürlich auch ohne Fehler!“



1	d d p d d d p p d p d d d d d p d p d d d p p d d d d d d p d p d d p p d d d d p p d p d d p
2	p d p p d d d p d p d d d p d d p d p d p d p d d d d p d p d p d p d d d d p d p d d
3	d d d d p p d p d p p p d d p d p d d p d p d d p d p p d d d d p d d p d p d d d d d p d
4	d d p d d d p p d p d d d d d p d p d d d p p d d d d d d p d p d d p p d d d d p p d p d d p

GZ	F <sub>1</sub>	F <sub>2</sub>	KL

# Beispiel Intelligenztests

Intelligenz-Struktur-Test (I-S-T) 2000 R (Amthauer, Brocke, Liepmann & Beauducel, 2001)

*Tab. 2.1 Intelligenzmaße des I-S-T 2000 R*

<b>Grundmodul</b>	
<b>(1) verbale Intelligenz</b>	(V; sprachgebundene Intelligenzfähigkeiten)
<b>(2) numerische Intelligenz</b>	(N; zahlengebundene Intelligenzfähigkeiten)
<b>(3) figurale Intelligenz</b>	(F; figural-räumliche Intelligenzfähigkeiten)
<b>(4) Merkfähigkeit</b>	(M)
<b>(5) schlußfolgerndes Denken</b>	(SD; schlußfolgerndes Denken mit Wissensanteilen)

- Erfasst mit dem Grundmodul die fluide Intelligenz und mit dem Erweiterungsmodul die kristallisierte Intelligenz (Wissen)

# I-S-T 2000 R: Beispielaufgaben

- Untertest „Analogien“

Gramm : Gewicht = Stunde : ?

a) Minute   b) Pause   c) Uhr   d) Tage   e) Zeit

- Untertest „Zahlenreihen“

33   30   15   45   42   21   63   ?

- Untertest „Matrizen“



?



a



b



c



d



e

# Arten psychologischer Tests

- Leistungstests
  - Intelligenztests
  - Entwicklungstests
  - Eignungstests
  - Schultests
- Psychometrische Persönlichkeitstests
  - Persönlichkeitsstrukturtests
  - Einstellungstests
  - Interessentests
  - Klinische Tests
- Persönlichkeitsentfaltungs-Verfahren/Projektive Verfahren
  - Formdeuteverfahren
  - Verbal-thematische Verfahren
  - Zeichnerische und Gestaltungsverfahren

# Beispiel Persönlichkeitsstrukturtests



- Revised NEO Personality Inventory (Costa & McCrae, 1992)
- Deutsche Version von Ostendorf und Angleitner (2004)
- Erfasst die Big Five auf der übergeordneten Domänebene: Neurotizismus, Extraversion, Offenheit für Erfahrungen, Verträglichkeit und Gewissenhaftigkeit

# NEO-PI-R

- Jede Big Five Domain ist unterteilt in 6 Facetten
- Facetten von Extraversion:
  - Herzlichkeit („Ich bin als herzliche und freundliche Person bekannt.“)
  - Geselligkeit („Ich habe gerne viele Leute um mich herum.“)
  - Durchsetzungsfähigkeit („Es fällt mir schwer, eine führende Rolle zu übernehmen.“)
  - Aktivität („Ich bin ein sehr aktiver Mensch.“)
  - Erlebnishunger („Manchmal habe ich etwas nur wegen des Nervenkitzels getan.“)
  - Frohsinn („Ich halte mich nicht für besonders fröhlich.“)
- Jede Facette wird mit 8 Items erfasst, d.h. pro Domain gibt es 48 und insgesamt 240 Items
- 5-stufige Ratingskala: *starke Ablehnung* – *Ablehnung* – *Neutral* – *Zustimmung* – *starke Zustimmung*



# Beispiel Klinische Tests

## Allgemeine Depressionsskala (ADS; Hautzinger & Bailer, 1993)

Bitte kreuzen Sie bei den folgenden Aussagen die Antwort an, die Ihrem Befinden während der letzten Woche am besten entspricht/entsprochen hat.

Antworten:    0    selten oder überhaupt nicht    (weniger als 1 Tag)  
                  1    manchmal    (1 bis 2 Tage lang)  
                  2    öfters    (3 bis 4 Tage lang)  
                  3    meistens, die ganze Zeit    (5 bis 7 Tage lang)

Während der letzten Woche ...	selten 0	manchmal 1	öfters 2	meistens 3
1. haben mich Dinge beunruhigt, die mir sonst nichts ausmachen	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2. konnte ich meine trübsinnige Laune nicht loswerden, obwohl mich meine Freunde/Familie versuchten, aufzumuntern	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3. hatte ich Mühe, mich zu konzentrieren	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

# Arten psychologischer Tests

- Leistungstests
  - Intelligenztests
  - Entwicklungstests
  - Eignungstests
  - Schultests
- Psychometrische Persönlichkeitstests
  - Persönlichkeitsstrukturtests
  - Einstellungstests
  - Interessentests
  - Klinische Tests
- Persönlichkeitsentfaltungs-Verfahren/Projektive Verfahren
  - Formdeutungsverfahren
  - Verbal-thematische Verfahren
  - Zeichnerische und Gestaltungsverfahren

# Beispiel Formdeuteverfahren

Rorschach Test (Rorschach, 1992)

„Was könnte dies sein?“



- Keine Standardinstruktion
- Antworten werden protokolliert
- Auswertung nach formalen (z.B. Anzahl Antworten, Reaktionszeiten) und inhaltlichen (z.B. Deutung als menschliche Figuren, Tiere oder Pflanzen) Aspekten

# Beispiel zeichnerische und Gestaltungsverfahren

Scenotest (von Staabs, 1995; 2004)



- Zur Erfassung unbewusster Probleme bei Kindern und Jugendlichen
- Darstellung von Szenen, in denen Alltags- und Beziehungserleben, Ängste, Wünsche etc. erkennbar werden können

# Qualitätsanforderungen an Tests

- Hauptgütekriterien
  - Objektivität
  - Reliabilität
  - Validität
  - Skalierung
- Nebengütekriterien
  - Normierung
  - Ökonomie
  - Nützlichkeit
  - Zumutbarkeit
  - Unverfälschbarkeit
  - Fairness

# Objektivität

Ein Test ist dann objektiv, wenn die Durchführung und Auswertung des Tests sowie die Interpretation des Testergebnisses unabhängig von dem/der Testleiter\*in ist.

## 1. Durchführungsobjektivität

- Standardisierung der Durchführungsbedingungen (z.B. Instruktion, Zeitbegrenzung)

## 2. Auswertungsobjektivität

- Auswertung mit Schablonen, Computern
- Bei Ratingskalen, Multiple Choice i.d.R. leicht zu erreichen
- Bei offenen Antwortformaten Festlegung exakter Auswertungsregeln
- Lässt sich angeben als Grad der Übereinstimmung zwischen Testauswerter\*innen

# Objektivität

## 3. Interpretationsobjektivität

- Liegt vor, wenn verschiedene Testauswerter\*innen bei identischen Testergebnissen dieselben Schlussfolgerungen ziehen
- Normtabellen zur Einordnung der Testperson im Vergleich zu relevanten Bezugsgruppen

# Reliabilität

- Die Reliabilität gibt den Grad der Messgenauigkeit eines Messwerts an.
- Das Ausmaß der Reliabilität wird über Reliabilitätskoeffizienten angegeben, die einen Wertebereich von 0 bis 1 haben
- Verfahren zur Bestimmung der Reliabilität
  1. Retest-Reliabilität
    - Test wird zu zwei verschiedenen Zeitpunkten durchgeführt
    - Korrelation der beiden Testergebnisse
    - Probleme: Korrelation kann in Abhängigkeit vom Zeitintervall variieren (Übungs- und Erinnerungseffekte, Merkmalsveränderungen)



# Reliabilität

## 2. Paralleltest-Reliabilität

- Korrelation der Testwerte aus „parallelen“ Testformen, die dasselbe Konstrukt erfassen und aus inhaltlich möglichst ähnlichen Items bestehen
- Probleme: schwierig, parallele Testformen herzustellen; Verzerrungseffekte durch Wiederholungsmessung

## 3. Testhalbierungs-Reliabilität

- Test wird in möglichst gleiche Hälften geteilt
- Korrelation der Testhälften
- Korrekturfaktor für Testlänge → aufgewertete Testhalbierungs-Reliabilität

## 4. Interne Konsistenz

- Jedes Item wird als eigenständiger Testteil angesehen
- Korrelation der Testteile (Items) unter Berücksichtigung der Testlänge

# Validität

Die Validität gibt an, ob der Test auch wirklich das misst, was er zu messen beansprucht.

## 1. Inhaltsvalidität

- Ausmaß, in dem ein Test oder ein Testitem das zu messende Merkmal repräsentativ erfasst
- Repräsentationsschluss: Testitems als repräsentative Stichprobe aus dem „Universum“ von Items, die das interessierende Merkmal abbilden
- Bestimmung aufgrund logischer und fachlicher Überlegungen (z.B. Expertenbefragung, Big Five: psycholexikalischer Ansatz)
- Eng verbunden mit der Augenscheinvalidität eines Tests

# Validität

## 2. Kriteriumsvalidität

- Zusammenhang des Testwerts mit Kriterien
- Vom Verhalten innerhalb der Testsituation (Testwert) wird auf Verhalten außerhalb der Testsituation (Kriterium) geschlossen
- Arten:
  1. Vorhersagevalidität
    - Korrelation mit zeitlich später erhobenem Kriterium
    - Z.B. Leistung im Intelligenztest zu Beginn des Studiums wird mit der Bachelor-Abschlussnote korreliert
  2. Übereinstimmungsvalidität
    - Korrelation mit zeitgleich erhobenem Kriterium
    - Z.B. Leistung im Konzentrationstest wird mit Erfolg in der Führerscheinprüfung korreliert

# Validität

## 3. Retrospektive Validität

- Korrelation mit zeitlich vorher erhobenem Kriterium
- Z.B. Leistung im Intelligenztest während des Studiums wird mit Abiturnote korreliert

## 4. Inkrementelle Validität

- Beitrag eines Tests zur Verbesserung der Vorhersage eines Kriteriums
- Z.B. Gewissenhaftigkeit zur Vorhersage des Berufserfolgs inkrementell zur Leistung im Intelligenztest

# Validität

## 3. Konstruktvalidität

- Liegt vor, wenn die Schlussfolgerungen, die aufgrund des Testwerts über das zugrundeliegende Konstrukt gemacht werden, wissenschaftlich fundiert sind
- Arten:
  1. Konvergente Validität
    - Korrelation mit Tests, die das gleiche oder ein ähnliches Konstrukt erfassen → Erwartung hoher Zusammenhänge
  2. Diskriminante Validität
    - Korrelation mit Tests, die ein anderes Konstrukt erfassen → Erwartung niedriger Zusammenhänge
  3. Faktorielle Validität
    - Prüfung der Struktur mit Verfahren der Faktorenanalyse (exploratorisch und konfirmatorisch) und Item Response Theorie

# Skalierung

- Das Gütekriterium der Skalierung ist erfüllt, wenn die laut Verrechnungsregel resultierenden Testwerte die empirischen Merkmalsrelationen adäquat abbilden
- D.h. Personen mit einer höheren Ausprägung auf dem Konstrukt müssen höhere Testwerte erhalten als Personen mit einer niedrigeren Ausprägung auf dem Konstrukt
- Abhängig vom Skalenniveau des Tests

# Qualitätsanforderungen an Tests

- Hauptgütekriterien
  - Objektivität
  - Reliabilität
  - Validität
  - Skalierung
- Nebengütekriterien
  - Normierung
  - Ökonomie
  - Nützlichkeit
  - Zumutbarkeit
  - Unverfälschbarkeit
  - Fairness

# Nebengütekriterien

- Normierung
  - Einordnung des individuellen Testwerts einer Person in eine Referenzgruppe
  - Referenzgruppe sollte Testperson hinsichtlich relevanter Merkmale (z.B. Alter, Geschlecht, Schulbildung) ähneln
  - Normstichprobe sollte möglichst groß und repräsentativ sein



# NEO-PI-R Normtabellen und Profilbogen<sup>5</sup>

Vorder- und Rückseiten der 10 Profilbogen:

NEO-PI-R	Beschriftung der Kopfzeile des Profils	N
<b>Form S</b>		
1.	a) Gesamtstichprobe b) Bevölkerungsrepräsentative Gesamtstichprobe <sup>a</sup>	11 724 871
2.	a) Männer / 16–20 Jahre b) Frauen / 16–20 Jahre	480 1 686
3.	a) Männer / 21–24 Jahre b) Frauen / 21–24 Jahre	1 358 1 925
4.	a) Männer / 25–29 Jahre b) Frauen / 25–29 Jahre	943 1 189
5.	a) Männer / 30–49 Jahre b) Frauen / 30–49 Jahre	1 035 1 992
6.	a) Männer / $\geq 50$ Jahre b) Frauen / $\geq 50$ Jahre	403 713
7.	a) Bevölkerungsrepräsentative Stichprobe <sup>a</sup> /Männer b) Bevölkerungsrepräsentative Stichprobe <sup>a</sup> /Frauen	423 448

Abb. aus NEO-PI-R Manual (Ostendorf & Angleitner, 2004)

# Nebengütekriterien

- Normierung
  - Einordnung des individuellen Testwerts einer Person in eine Referenzgruppe
  - Referenzgruppe sollte Testperson hinsichtlich relevanter Merkmale (z.B. Alter, Geschlecht, Schulbildung) ähneln
  - Normstichprobe sollte möglichst groß und repräsentativ sein
  - DIN 33430: Normen sollten alle 8 Jahre überprüft und ggf. eine Neunormierung vorgenommen werden
  - Nicht erforderlich bei kriteriumsorientiertem Testen (z.B. PISA)

# Nebengütekriterien

- Ökonomie
  - Liegt vor, wenn die Kosten (Zeit, Geld, ...) gemessen am diagnostischen Erkenntnisgewinn relativ gering sind
  - Hohe Ökonomie darf nicht zulasten der anderen Kriterien gehen (v.a. Validität)
- Nützlichkeit
  - Ein Test ist nützlich, wenn er ein Merkmal erfasst oder vorhersagt, das praktische Relevanz besitzt und es nicht bereits einen Test für das Merkmal gibt, der die übrigen Gütekriterien genauso gut erfüllt

# Nebengütekriterien

- Zumutbarkeit
  - Ein Test ist zumutbar, wenn er die Testperson in zeitlicher, psychischer sowie körperlicher Hinsicht nicht über Gebühr belastet

# Nebengütekriterien

- Unverfälschbarkeit
  - Ein Test ist unverfälschbar, wenn die Testperson durch gezieltes Testverhalten die konkreten Ausprägungen ihrer Testwerte nicht steuern bzw. verzerren kann
  - Verfälschung (z.B. im Sinne der sozialen Erwünschtheit) v.a. möglich bei Tests, bei denen leicht zu erkennen ist, was gemessen wird
  - Bsp. „Ich komme immer pünktlich in die Vorlesung.“

# Nebengütekriterien

- Fairness
  - Ein Test ist fair, wenn die resultierenden Testwerte zu keiner systematischen Benachteiligung bestimmter Personen aufgrund ihrer Zugehörigkeit zu ethnischen, soziokulturellen oder geschlechtsspezifischen Gruppen führen
  - Z.B. computergestützte Diagnostik bei jüngeren vs. älteren Personen

# Interpretation Gütekriterien

- Gütekriterien sind keine Eigenschaft des Tests
  - Z.B. nicht der Test an sich ist valide/nicht valide, sondern die Schlussfolgerungen, die aus den Testergebnissen gezogen werden
- Gütekriterien liegen nicht für einen Test vor, sondern genau genommen für die Testkennwerte in bestimmten Stichproben
  - Z.B. „Die Reliabilitäten nach Cronbachs alpha für Summenscores auf den Big Five waren .93 für Neurotizismus, .89 für Extraversion,...“

# Literatur zu dieser Sitzung

- Moosbrugger & Kelava (2012): Kapitel 1 und 2