



Grundlagen der Testtheorie

BSc Psychologie WiSe 2025/26

Prof. Dr. Dirk Ostwald

(1) Psychologische Tests

Definition

Testarten

Antwortformate

Testgütekriterien

Testtheorien

Selbstkontrollfragen

Definition

Testarten

Antwortformate

Testgütekriterien

Testtheorien

Selbstkontrollfragen

Test (engl.) *Probe, Prüfung, Kontrolle*

Bedeutungen

- Ein Verfahren zur Untersuchung eines Persönlichkeitsmerkmals
- Der Vorgang der Durchführung einer Untersuchung
- Die Gesamtheit der zur Durchführung einer Untersuchung nötigen Materialien
- Ein statistisches Prüfverfahren (Hypothesentests)
- Kurze, außerplanmäßige “Zettelarbeiten” im Schulunterricht
- u.v.a.m.

Lienert and Raatz (1998)

Test (engl.) *Probe, Prüfung, Kontrolle*

test = 1. (*logic*) any criterion or procedure used to determine the truth or falsity of a hypothesis, either by empirical evidence or by reasoning; 2. (*psychol.*) a routine examination administered to individuals belonging to the same group, in order to determine the relative position of a given individual in the group with respect to one or more mental traits, motor abilities, etc. or in order to compare one group with another in these characteristics. [In practice a series or battery of psychological tests (2) is generally used, and the scores are arranged or plotted on a definite scale; cf. **mental test, scale** (1).] **tester**ⁿ (contr. rarely w. **tested**, the individual tested). — Fr. (1) *épreuve*, (2) *test*; Ger. (1) *Prüfung*, (2) *Test, Probe*.

Warren (1934)

Definition

Definitionsversuche

Definition eines **Tests** nach Moosbrugger and Kelava (2012)

“Ein *Test* ist ein wissenschaftliches Routineverfahren zu Erfassung eines oder mehrerer empirisch abgrenzbarer psychologischer Merkmale mit dem Ziel einer möglichst genauen quantitativen Aussage über den Grad der individuellen Merkmalsausprägung”

Definition eines **psychometrischen Tests** nach Bühner (2010)

“Ein *psychometrischer Test* ist ein wissenschaftliches Routineverfahren zur Untersuchung eines oder mehrerer empirisch abgrenzbarer Persönlichkeitsmerkmale (vgl. Lienert and Raatz (1998), S.1). Das Ziel eines psychometrischen Tests besteht darin, die absolute oder relative Ausprägung einer Eigenschaft, einer Fähigkeit oder eines Zustands bei einer oder mehreren Personen zu messen oder eine qualitative Aussage zu treffen, welcher Personenklasse Personen zugeordnet werden können (vgl. Rost (2004)). Psychometrische Tests sind nach der Klassischen oder probabilistischen Testtheorie entwickelt, sind theoretisch fundiert und genügen genau definierten Gütekriterien (Haupt- und Nebengütekriterien).”

Definition eines **Psychologischen Tests** nach Krauth (1995)

“Ein *psychologischer Test* besteht aus eine Menge von Reizen mit den zugehörigen zugelassenen Reaktionen, d.h. aus einer Menge von manifesten Variablen, und einer Vorschrift (Skala), die den Reaktionsmustern der manifesten Variablen Ausprägungen einer oder mehrerer latenter Variablen zuordnet. Anders ausgedrückt ist ein Test ein Messinstrument zur Messung von nicht direkt beobachtbaren (latenten) Variablen, deren Existenz man bei Personen (gelegentlich auch bei Tieren) postuliert.”

Definition

Definition eines **Psychologischen Tests** nach Lienert (1968) kommentiert von Rost (2004)

“Ein Test ist ein wissenschaftliches Routineverfahren zur Untersuchung eines oder mehrerer empirisch abgrenzbarer Persönlichkeitsmerkmale mit dem Ziel einer möglichst quantitativen Aussage über den relativen Grad der individuellen Merkmalsausprägung.”

- “Wissenschaftlich”: Existenz einer Theorie darüber unter welchen Bedingungen aus den Testergebnissen welche Aussagen über die Testpersonen abgeleitet werden können (eben eine Testtheorie)
- “Routineverfahren”: Durchführung und Auswertung sind an einer Normstichprobe erprobt und dokumentiert
- “Persönlichkeitsmerkmale”: Erfassung eines relativ stabilen und konsistenten Merkmals der Testperson, das für das im Test gezeigte Verhalten verantwortlich ist
- “Quantitativ”: Relativer Grad der Merkmalsausprägung, es können aber auch qualitative (kategoriale) Aussagen von Interesse sein

Tests im engeren und weiteren Sinn nach Rost (2004)

Tests im engeren Sinn

- Testverfahren, die die Testpersonen nicht willentlich in eine gewünschte Richtung verfälschen kann

Tests im weiteren Sinn

- Fragebögen, standardisierte Interviews und standardisierte Beobachtungen

Testbestandteile

Materialbestandteile

- Testmanual
- Testmaterial (Items)
- Auswertungshilfen

Durchführungsbestandteile

- Testanweisung
- Testdurchführung
- Testauswertung

Lienert and Raatz (1998)

Definition

Item

Kleinste Beobachtungseinheit und damit Grundbausteine eines psychologischen Tests

Item = Itemstamm + Antwortformat

Itemstamm

- Reiz, auf den eine Reaktion erwartet wird
- Beispiele: Frage, Aussage, Bild, Geschichte, Rechenaufgabe, ...

Antwortformat

- Art und Weise der Reaktionsregistrierung
- Beispiele: Kreuz auf einer Skala, leere Zeile, Zeichenpapier

Oft besteht ein Test aus mehreren Untertests, die aus mehreren Items bestehen

Prinzipiell können Untertests und Tests aus nur einem Item bestehen

Jedes Item kann also auch als ein Test angesehen werden

Lord and Novick (1968), Rost (2004)

Test vs. Experiment

Psychologische Tests sind spezielle Experimente

Unabhängige Variable

- Test mit den Leveln “Items”

Abhängige Variable

- Itemantworten oder Gesamtsummenscore

Experimentelle Einheiten

- Testpersonen mit Messwiederholung

Konzeptualisierung als Einstichprobendesigns oder Testpersonen $\times$ Items Designs

Rost (2004)

Definition

Testarten

Antwortformate

Testtheorien

Selbstkontrollfragen

Klassifikation

Psychometrische Persönlichkeitstests

- Klinische Tests
- Einstellungstests
- Persönlichkeitsstrukturtests
- Interessentests

Leistungstests

- Entwicklungstests
- Intelligenztests
- Allgemeine Leistungstests
- Schultests
- Spezielle Funktionsprüfungs- und Eignungstests

Persönlichkeitsentfaltungungsverfahren

- Formdeutungsverfahren
- Verbal-thematische Verfahren
- Zeichnerische- und Gestaltungsverfahren

Charakteristika

Psychometrische Persönlichkeitstests

- Messen einer akuten Symptomatik (state) oder stabilen Eigenschaft (trait)
- Selbstauskunftsfragebögen ohne richtige oder falsche Antworten
- Voraussetzungen: Metakognition und Offenbarungsbereitschaft
- Verfälschung und Durchschaubarkeit in der Regel leicht

Leistungstests

- Messen einer Leistung
- Leistung kann als richtig oder falsch klassifiziert werden
- Schnelligkeitstest (Speedtests) und Niveautests (Powertests)
- Verfälschung nach unten möglich

Persönlichkeitsentfaltungsverfahren

- Projektive Verfahren
- Projektion als psychoanalytischer Abwehrmechanismus
- ICH projiziert ÜBER-ICH-konfliktträchtige Triebregungen nach außen
- Üblicherweise unstrukturiert und qualitativ

Klinische Tests

Fragebogenverfahren zur ICD/DSM-basierten Diagnose psychischer Störungen

Störungsübergreifende und störungsspezifische Verfahren

Beispiel: Beck-Depressionsinventar (BDI-II)

- Selbstbeurteilungsinstrument zur Erfassung der Schwere einer depressiven Symptomatik
- Entwicklung von Aaron T. Beck et al. (1996)
- Nachfolger des Beck Depressionsinventar A. T. Beck (1961)
- Deutschsprachige Version nach Hautzinger, Keller, and Kühner (2006)
- Sprachliche Überarbeitung und Anpassung an aktuelle Diagnosemanuale
- 21 Items zur Erfassung des Schweregrads von 0 bis 3 und summative Gesamtauswertung
- Grenzwerte der [Nationalen Versorgungsleitlinie Unipolare Depression](#)

0–13 keine Depression bzw. klinisch unauffällig oder remittiert

14–19 leichtes depressives Syndrom

20–28 mittelgradiges depressives Syndrom

≥ 29 schweres depressives Syndrom

Klinische Tests

Beispiel: Beck-Depressionsinventar (BDI-II)

Beispielitem Traurigkeit

- (0) Ich bin nicht traurig
- (1) Ich bin oft traurig
- (2) Ich bin ständig traurig
- (3) Ich bin so traurig oder unglücklich, dass ich es nicht aushalte

Klinische Tests

Beispiel: Beck-Depressionsinventar (BDI-II)

Items

1. Traurigkeit
2. Pessimismus
3. Versagensgefühle
4. Verlust an Freude
5. Schuldgefühle
6. Bestrafungsgefühle
7. Selbstablehnung
8. Selbstkritik
9. Suizidgedanken
10. Weinen
11. Unruhe
12. Interessensverlust
13. Entschlussunfähigkeit
14. Wertlosigkeit
15. Energieverlust
16. Veränderungen der Schlafgewohnheiten
17. Reizbarkeit
18. Appetitveränderung
19. Konzentrationsschwierigkeiten
20. Müdigkeit
21. Verlust an sexuellem Interesse

Einstellungstests

Psychologische Einstellung

- Relativ stabile, aber veränderbare innere Haltung einer Person
- Haltung gegenüber Objekten, Personen, Gruppen, Ideen oder Situationen
- Kognitive, affektive und konative (handlungsbereitschaftsorientierte) Komponenten

Einstellungstest

- Psychologisches Messinstrument zur systematischen Erfassung von Einstellungen
- Meist durch Fragebögen oder Skalen realisiert
- Testperson bewertet Aussagen in unterschiedlichem Maß

Ziele eines Einstellungstests

- Erhebung von Wertorientierungen, Vorurteilen oder politischen/religiösen Überzeugungen
- Messung von arbeits- oder organisationsbezogenen Einstellungen
- Forschung zu sozialen Einstellungen (z. B. Minderheiten, Umweltschutz, Gesundheit)

Einstellungstests

Beispiel: General Ecological Behavior (GEB) Scale

Ursprünglich: Messinstrument für umweltschützendes Verhalten (F. G. Kaiser (1998))

Später: Messinstrument für individuelle Umwelteinstellung (F. Kaiser, Byrka, and Hartig (2010))

Theoretische Grundlage: Campbell-Paradigma (Campbell (1963))

- Grundidee: Stärke der Einstellung über erbrachte *Verhaltenskosten* messbar
- Einstellung: Ziel (z. B. "Umwelt schützen")
- Verhaltenskosten: Eingesetzte Mittel zur Zielerreichung

Verhaltenskosten

- Unterschiedliche Kosten für unterschiedliche Verhaltensmittel
- Bsp. Recycling (gering) vs. Mitgliedschaft in Umweltschutzgruppe (hoch)
- Kostenarten: Aufwand, Zeit, Geld, soziale Normen, kulturelle Praktiken
- Je mehr und je höhere Kosten eine Person trägt, desto stärker ausgeprägte Umwelteinstellung

Testarten

Einstellungstests

Beispiel: General Ecological Behavior (GEB) Scale

Testaufbau

- 50 Selbstberichte von Verhaltensweisen, mit deren Hilfe Menschen die Umwelt schützen
- 32 Selbstberichte erfordern Auskunft über die Häufigkeit ihres Auftretens
- Antwortmöglichkeiten: "nie", "selten", "gelegentlich,"oft" "sehr oft"
- 18 Selbstberichte erfordern eine Ja/Nein-Aussage.
- "keine Angabe" bei allen Selbstberichten möglich

Itembeispiele

"Ich kaufe Lebensmittel aus kontrolliert biologischem Anbau."

☐ nie ☐ selten ☐ gelegentlich ☐ oft ☐ sehr oft ☐ keine Angabe

"Altglas bringe ich zum Sammelcontainer."

☐ nie ☐ selten ☐ gelegentlich ☐ oft ☐ sehr oft ☐ keine Angabe

"Ich bin Mitglied in einer Umweltschutzorganisation."

☐ ja ☐ nein ☐ keine Angabe

Testarten

Leistungstests

Beispiel: Konzentrationstest für Kinder (KoKi)

Objektive und kindgerechte Erfassung der Konzentrationsfähigkeit

Testaufbau und Aufgabenstellung

- 8 Testseiten mit je 100 Bildern, jede Seite mit 10 Zeilen $\times$ 10 Bilder (Hunde und Pferde)
- Kinder sollen Hunde durchstreichen, die in Richtung einer Wurst am Zeilenanfang/ende blicken
- Alle anderen Bilder = Distraktoren

Spezielle Distraktoren

- Auf der Hälfte der Testseiten: "verzauberte" Pferde (z. B. mit Flügeln, Geweih, 3 Beinen)
- Diese erregen besonders viel Aufmerksamkeit $\Rightarrow$ erschwerte Leistungskontrolle
- Erwartbar: geringere Testergebnisse auf Seiten mit "verzauberten" Pferden

Erhobene Kennwerte

- Konzentrationsfähigkeit, Arbeitstempo, Genauigkeit
- Unablenkbarkeit (Ablenkbarkeit durch "verzauberte" Pferde)

Diagnostische Relevanz

- Kombination der vier Kennwerte liefert Hinweise für ADHS-Diagnostik

Leistungstests

Beispiel: Konzentrationstest für Kinder (KoKi)

Beispielitem



Schmidt-Atzert and Funsch (2023)

Projektive Verfahren

Rorschach Test



Thematischer Apperzeptionstest



Rorschach (1921), Revers and Taeber (1968)

Definition

Testarten

Antwortformate

Testgütekriterien

Testtheorien

Selbstkontrollfragen

Antwortformat eines psychologischen Tests

Die Art und Weise, wie Testpersonen auf die gestellten Aufgaben reagieren sollen

Wahl des Antwortformates ist abhängig von Testart und Testaufgabe

Freies/offenes Antwortformat

- Kurzaufsatzaufgaben
- Ergänzungsaufgaben

Gebundenes/geschlossenes Antwortformat

- Ordnungsaufgaben
- Auswahlaufgaben
- Beurteilungsaufgaben

Wir folgen Pospeschill (2022), Kapitel 2

Antwortformate

Offene Antwortformate

- Die Testperson formuliert die Antwort auf Grundlage einer strukturellen Vorgabe eigenständig
- Einsatz bei projektiven Verfahren und Abfragen komplexer Wissensinhalte
- Man unterscheidet Kurzaufsatzaufgaben und Ergänzungsaufgaben

Kurzaufsatzaufgaben

- Kurze schriftliche Abhandlung von einem Wort bis zu mehreren Sätzen
- Begrenzung der Wortzahl kann sinnvoll sein

Ergänzungsaufgaben

- Offene Fragen mit einem Schlüsselwort am Satzende
- Lückentext mit zu ergänzenden Schlüsselwörtern im laufenden Text

Vorteile offener Antwortformate

- Bei Leistungstests: Eigenständige Leistung, keine Zufallstreffer möglich
- Testen von Sprachverständnis, Motivation, Interesse, Kreativität, Reproduktion, Transfer

Nachteile offener Antwortformate

- Hoher Bearbeitungsaufwand und kognitive Belastung für Testpersonen
- Hoher Auswertungsaufwand und erschwerte Auswertungsobjektivität für Testende

Pospeschill (2022)

Offene Antwortformate

Kurzaufsatz:

Definieren Sie den Begriff „Gesundheit“ nach dem Verständnis der WHO.

(Lösung: Gesundheit ist ein Zustand des völligen körperlichen, seelischen und sozialen Wohlbefindens und nicht nur das Freisein von Krankheit und Gebrechen.)

Ergänzungsaufgabe: Offene Fragen

[Einfachantwort] (1) Nennen Sie das Hauptalkaloid des Tabakblattes?

[Reihenantwort] (2) Nennen Sie vier Rechte eines Käufers bei aufgetretenem Sachmangel?

[Analogieantwort] (3) Dienstag steht zu Mittwoch wie April zu ...?

[Folgeantwort] (4) 4 16 256

(Lösung: 1. Nikotin; 2. Nacherfüllung, vom Vertrag zurücktreten, Kaufpreisminderung, Schadenersatz; 3. Mai; 4. 65.536)

Ergänzungsaufgabe: Lückentext

Ergänzen Sie bei dem folgenden Text die fehlenden Wortteile:

La__ Saturday was Mr. Miller's 70th birth__. His son and h__ wife had a go__ idea. In the mor__ they and their chil__ prepared a nice brea__ with ham and eg__.

(Lösung: Last; birthday; his; good; morning; children; breakfast; eggs)

Pospeschill (2022)

Ordnungsaufgaben

Umordnungsaufgaben und Zuordnungsaufgaben

Umordnungsaufgaben

- Testen schlussfolgernden Denkens oder des Ursache-Wirkungs-Verständnisses
- Nicht-verbale Aufgaben schließen Beeinträchtigung durch Lesefähigkeit aus
- Bildliche Aufgaben haben einen hohen Materialentwicklungsaufwand

Zuordnungsaufgaben

- Wissensprüfung durch Wiedererkennung
- Einfache Konstruktion, hohe Ökonomie und Auswertungsobjektivität
- Keine Reproduktion, Wiedererkennung reicht aus
- Zufallstreffer sind bei Fehlen von nicht-zuordbaren Antworten möglich

Pospeschill (2022)

Ordnungsaufgaben

Umordnungsaufgabe:

Bringen Sie die Worte in eine sinnvolle Reihenfolge
Herrn seinen ein rettete treuer Hund

(Lösung: Ein treuer Hund rettete seinen Herrn.)

Zuordnungsaufgabe:

Ordnen Sie dem jeweiligen Fachbegriff sein Synonym zu:

- | | |
|------------------|------------------------|
| (a) Reliabilität | (1) Eichung |
| (b) Validität | (2) Zuverlässigkeit |
| (c) Objektivität | (3) Nützlichkeit |
| (d) Utilität | (4) Gültigkeit |
| (e) Normierung | (5) Intersubjektivität |

(Lösung: a-2; b-4; c-5; d-3; e-1)

Pospeschill (2022)

Auswahlaufgaben

- Auswahl zutreffender Antworten bei Vorgabe mehrerer Alternativen
- Dichotom = Zwei mögliche Antworten, Multiple Choice = Mehrere mögliche Antworten
- *Distraktoren* (Ablenkungsantworten) dienen der Schwierigkeit bei Leistungstests
- *Disjunktion* (Trennung) zwischen richtigen und falschen Aussagen wichtig bei Leistungstests
- *Exhaustivität* (Vollständigkeit) wichtig bei Persönlichkeitstests
- Fehlende Exhaustivität führt zu *Forced-Choice-Antworten*

Vorteile von Auswahlaufgaben

- Einfach, ökonomisch, objektiv

Nachteile von Auswahlaufgaben

- Hohe Ratewahrscheinlichkeit
- Reine Rekognitionsleistung, keine Problemlösefähigkeiten- oder Kreativitätsleistung
- Distraktionsentwicklung kann komplex sein

Pospeschill (2022)

Auswahlaufgaben

Einfachwahlaufgabe:

Welche der nachfolgenden Zahlengruppen folgt nicht der Logik der übrigen:

(a) 2468 (b) 1369 (c) 1234 (d) 3546 (e) 6789

(Lösung: d)

Mehrfachwahlaufgabe:

Wählen Sie aus den nachfolgenden Begriffen diejenigen aus, die zu einer gemeinsamen Kategorie gehören:

Messer Butter Zeitung Brot Zigarre Armband

(Lösung: Butter, Brot mit der gemeinsamen Kategorie „Lebensmittel“)

Pospeschill (2022)

Auswahlaufgaben

Nicht disjunkte Antwortalternativen:

Welches ist der Oberbegriff für Birne?

- (a) *Strauchfrüchte* (b) *Obst* (c) *Gemüse* (d) *Baumfrüchte* (e) *Erdfrüchte*

(Lösung: b und d)

Nicht exhaustive Antwortalternativen:

Hund steht zu Welpen wie Schwein zu ...

- (a) *Nutztier* (b) *Lamm* (c) *Sau* (d) *Ferkel* (e) *Eber*

(Lösung: d)

Forced-Choice-Antwortalternativen:

Welche der folgenden Freizeitbeschäftigungen sagt Ihnen am meisten zu?

Wählen Sie bitte die Antwortalternative, die am ehesten auf Sie zutrifft.

- (a) *zu Hause bleiben und fernsehen*
(b) *einen Spaziergang unternehmen*
(c) *sich mit Freunden treffen*
(d) *in die Stadt gehen und einen Einkaufsbummel machen*

Pospeschill (2022)

Beurteilungsaufgaben

- Anwendung vorzugsweise im Rahmen von klinischen Fragebögen und Persönlichkeitstests
- Einschätzung eines *Statements* mit *diskreter Ratingskala* oder *kontinuierlicher Analogskala*
- Aus den (gewichteten) Punktwerten wird meist ein Gesamtpunktwert (*score*) errechnet

Wichtige Aspekte von Beurteilungsaufgaben

1. Anzahl der Skalenstufen
2. Forciertheit des Ratings
3. Ausweichkategorien
4. Polarität der Antwortskala
5. Balance der Antwortskala
6. Verankerung der Antwortskala
7. Symbole und Marker
8. Kombinierte Antwortskalen
9. Skalenniveau im Sinne der Repräsentationstheorie des Messens

Pospeschill (2022)

Beurteilungsaufgaben

Anzahl von Skalenstufen

- Abhängig von und determinierend für die Urteilsdifferenzierbarkeit
- Meist 9 ± 1 Skaleneinheiten bei einzelner Ratingskala (Test mit einem Item)
- Meist 5 ± 2 Skaleneinheiten bei Itembatterien (Test mit vielen Items)
- Entfällt für grafische Analogskalen

Unforciertes und forciertes Rating

- Ungerade Anzahl an Kategorien enthalten Mittelpunkt als mittleres Rating/Ausweichoption
- Mittelpunkt kann uneindeutig interpretiert werden = unforciertes Rating
- Gerade Anzahl an Kategorien enthalten keinen Mittelpunkt
- Testperson ist zu einem Urteil gezwungen = forciertes Rating

Ausweichkategorie

- "Weiß nicht", "nicht anwendbar", "keine Meinung"
- Valide Information über Nichtbearbeitung bzw. Nichtbeantwortung der Frage

Pospeschill (2022)

Beurteilungsaufgaben

Verbale Ratingskalen (unforciert)

... zur Skalierung einer Häufigkeit



... zur Skalierung einer Intensität



... zur Skalierung einer Wahrscheinlichkeit



... zur Skalierung einer Bewertung



... zur Skalierung einer Einstellung

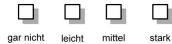


Pospeschill (2022)

Beurteilungsaufgaben

Verbale Ratingskalen (forciert)

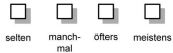
... zur Skalierung einer Intensität



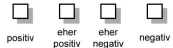
... zur Skalierung einer Bewertung



... zur Skalierung einer Häufigkeit



... zur Skalierung einer Einstellung

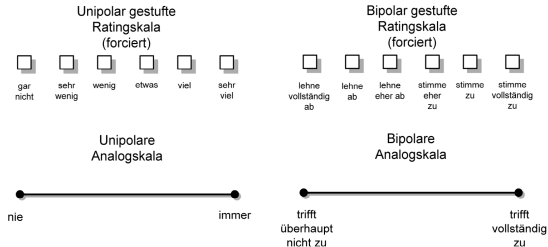


Pospeschill (2022)

Beurteilungsaufgaben

Polarität der Antwortskala

- *Bipolare Skalen* mit positivem (starke Zustimmung) und negativem (starke Ablehnung) Pol
- *Unipolare Skalen* mit Bezugspunkt geringster Intensität und Pol größter Intensität
- Abhängig vom erfragten Gegenstandsbereich und Iteminhalt

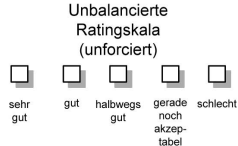
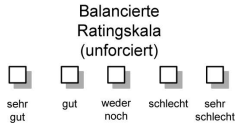


Pospeschill (2022)

Beurteilungsaufgaben

Balance der Antwortskala

- Balancierte Skala: Gleiche Anzahl an positiven und negativen Kategorien
- Unbalancierte Skala: Ungleiche Anzahl an positiven und negativen Kategorien
- Allgemein werden balancierte Skalen vorgezogen
- Unbalancierte Skalen bei Annahmen über vorwiegenden Antwortbereich sinnvoll



Pospeschill (2022)

Beurteilungsaufgaben

Verankerung der Antwortskala

- Definition der Skalenendpunkte sowie der einzelnen Skalenstufen
- *Numerische Verankerung*: Bezeichnung der Kategorien mit Zahlen
- *Verbale Verankerung*: Bezeichnung der Kategorien mit Worten oder Sätzen
- Allgemein kann Äquidistanz der Kategorienurteile nicht impliziert werden

Symbole und Marker

- Verwendbarkeit bei mit Zahlen oder semantischen Kategorien ungeübten Testpersonen

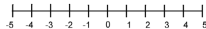
Kombinierte Skalen

- Multiple Skalierung durch sowohl numerische, verbale oder symbolische Verankerung

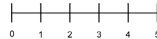
Pospeschill (2022)

Beurteilungsaufgaben

Numerische
Ratingskala
(unforciert)



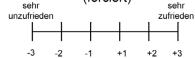
Numerische
Ratingskala
(forciert)



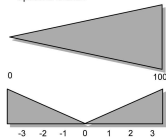
Kombinierte
Ratingskala
(unforciert)



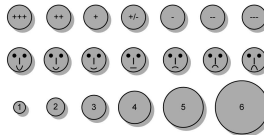
Kombinierte
Ratingskala
(forciert)



Optische Skalen



Symbolische Marker



Pospeschill (2022)

Beurteilungsaufgaben

Skalenniveau im Sinne der Repräsentationstheorie des Messens

- Aus Äquidistanz der Ratingkategorien kann nicht auf Äquidistanz der Urteile geschlossen werden
- Ratingtechnik ist eine Einzelreizmethode, kein Additive Conjoint Measurement
- Das repräsentationstheoretische Skalenniveau einer Ratingskala ist i. Allg. unbestimmt
- Bei jeder Messung interagieren das zu messende Merkmal, das Messinstrument, die Testperson

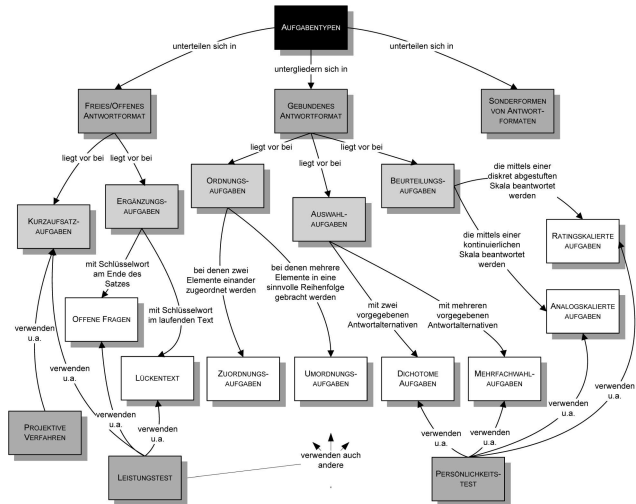
Vorteile von Beurteilungsaufgaben

- Leichte Handhabung, Ökonomie, allgemeine Akzeptanz durch Testpersonen

Nachteile von Beurteilungsaufgaben

- Eher Ordinal- denn Intervallskalencharakter
- Subjektive Interpretation der Antwortkategorien (z.B. "selten", "manchmal", "oft")

Pospeschill (2022)



Definition

Testarten

Antwortformate

Testgütekriterien

Testtheorien

Selbstkontrollfragen

Testgütekriterien

Qualitätseigenschaften von psychologischen Tests

Grundlage der Argumentation für die Wissenschaftlichkeit eines Testverfahrens

Qualitative Konzepte die (nur) zum Teil quantitativ erfasst werden können

Hauptgütekriterien

1. Objektivität
2. Reliabilität
3. Validität

Nebengütekriterien

1. Normierung
2. Skalierung
3. Testökonomie
4. Nützlichkeit
5. Zumutbarkeit
6. Unverfälschbarkeit
7. Fairness

Wir folgen Kapitel 2 in Moosbrugger and Kelava (2012)

Objektivität

“Ein Test ist dann *objektiv*, wenn er dasjenige Merkmal, das er misst, unabhängig von Testleitenden und Testauswertenden misst. Außerdem müssen klare und anwenderunabhängige Regeln für die Ergebnisinterpretation vorliegen.”

Durchführungsobjektivität

- Unabhängigkeit der Testpersonantwort von den Testleitenden und der Testsituation
- Standardisierte Testausführung anhand des Testmanuals
- Höchstmögliche Kontrolle der Testsituation

Auswertungsobjektivität

- Unabhängigkeit der Testauswertung von den Testauswertenden
- Bei gebundenem Antwortformat und automatisierter Auswertung üblicherweise gegeben
- Bei offenen Antwortformaten aufwändiger herzustellen

Interpretationsobjektivität

- Unabhängigkeit der Testergebnisschlussfolgerungen von Testanwendern
- Meist durch Normwertklassifikationen gegeben

Moosbrugger and Kelava (2012)

Beispiel BDI-II

Durchführungsobjektivität durch schriftliche Durchführungsinstruktion

Dieser Fragebogen enthält 21 Gruppen von Aussagen. Bitte lesen Sie jede dieser Gruppen von Aussagen sorgfältig durch und suchen Sie sich dann in jeder Gruppe eine Aussage heraus, die am besten beschreibt, wie Sie sich in den letzten zwei Wochen, einschließlich heute, gefühlt haben. Kreuzen Sie die Zahl neben der Aussage an, die Sie sich herausgesucht haben (0,1,2 oder 3). Falls in einer Gruppe mehrere Aussagen gleichermaßen auf Sie zutreffen, kreuzen Sie die Aussage mit der höheren Zahl an. Achten Sie bitte darauf, dass Sie in jeder Gruppe nicht mehr als eine Aussage ankreuzen, das gilt auch für Gruppe 16 (Veränderungen der Schlafgewohnheiten) oder Gruppe 18 (Veränderungen des Appetits).

Auswertungsobjektivität

Der Fragebogen wird durch die einfache Addition der angekreuzten Aussagen ausgewertet. Jedes Item wird auf einer 4-Punkt-Skala bewertet, die von 0 bis 3 reicht. Kreuzt ein Proband bei einem Item mehrere Aussagen an, so geht nur die Aussage mit der höchsten Ziffer in den Summenwert ein. Der Gesamtwert des BDI-II kann Werte zwischen 0 und 63 Punkten annehmen.

Intepretationsobjektivität

0 - 8 Keine, 9 - 13 minimale, 14 - 19 leichte, 20 - 28 mittelschwere, 29 - 63 schwere Depression

Hautzinger, Keller, and Kühner (2006)

Reliabilität

“Ein Test ist dann *reliabel*, wenn er das Merkmal, das er misst, exakt, d.h. ohne Messfehler misst.”

- Die klassische Testtheorie bettet diese Definition in ein probabilistisches Modell
- Die klassische Testtheorie ist in erster Linie eine quantitative Reliabilitätstheorie

Retest-Reliabilität

- Korrelation der Ergebnisse des identischen Tests zwischen zwei Testmessungen
- Abhängig von Zeitintervall-, Erinnerungs-, und Lerneffekten

Paralleltest-Reliabilität

- Korrelation der Ergebnisse eines Tests mit dem Ergebnis eines Paralleltests
- Paralleltest als Test aus möglichst ähnlichen Items (Itemzwillinge)

Testhalbierungs-Reliabilität (Split-Half-Reliabilität)

- Unterteilung eines Tests in zwei möglichst parallele Testhälften
- Korrekturfaktor zur Hochrechnung auf ursprüngliche Itemanzahl (Spearman-Brown-Formel)

Interne Konsistenz

- Verallgemeinerung der Testhalbierungs-Reliabilität auf Einzelitemebene
- Cronbach's α der klassischen Testtheorie

Moosbrugger and Kelava (2012)

Testgütekriterien

Beispiel BDI-II

In der Studie von Beck et al. (1996) resultierte bei 26 Patienten, die im Abstand von einer Woche untersucht wurden, ein Retestkoeffizient von $r_{tt} = .93$. Aus Studententstichproben werden für ein- bis zweiwöchige Zeiträume Koeffizienten von $r_{tt} = .74 - .96$ berichtet (Leigh & Antony-Tolbert, 2001; Al-Musawi, 2001; Sprinkle, Lurie, Insko et al., 2002; Ghassemzadeh, Mojtabei, Karamghadiri & Ebrahimkhani, 2005). In Tabelle 2 sind bisherige Ergebnisse zur internen Konsistenz (Cronbach's α) des BDI II aufgelistet. In psychiatrischen Stichproben lag die interne Konsistenz im Bereich von $.89 \leq \alpha \leq .94$, in nichtklinischen Stichproben im Bereich von $.84 \leq \alpha \leq .91$.

Hautzinger, Keller, and Kühner (2006)

Beispiel GEB

Die gefundene Test-Retest-Reliabilität liegt bei $r_{tt} = .76$ (Kaiser & Wilson, 2000) und $r_{tt} = .83$ (Kaiser, Frick & Stoll-Kleemann, 2001) und sogar bei $r_{tt} = .96$ (Kaiser & Byrka, 2011) und $r_{tt} = .99$ (Kaiser, Brügger, Hartig, Bogner & Gutscher, 2014), wenn korrigiert um die Messfehlerausdünnung.

F. G. Kaiser (2020)

Beispiel KoKi

Interne Konsistenz (Cronbachs Alpha) für Konzentrationsleistung = .96, Arbeitstempo = .94 und Genauigkeit = .83. Retestreliabilität (Intervall von 6 bis 12 Wochen) für Konzentrationsleistung = .84, Arbeitstempo = .85, Genauigkeit = .53 sowie Unablenkbarkeit = .47.

Schmidt-Atzert and Funsch (2023)

Validität

“Ein Test gilt dann als valide, wenn er das Merkmal, das er messen soll, auch wirklich misst und nicht irgendein anderes.”

- Es existiert keine allgemeine quantitative Validitätstheorie
- Je nach Testart werden unterschiedliche Maße zur Validitätsargumentation genutzt

Augenscheinvalidität

- Validitätseinschätzung durch Laien
- Qualitativer Grad der intuitiv-repräsentativen Erfassung des Merkmals

Inhaltsvalidität

- Validitätseinschätzung durch Testexpert:innen
- Qualitativer Grad der fachlich-repräsentativen Erfassung des Merkmals

Kriteriumsvalidität

- Quantitativer Grad des Zusammenhangs eines Testergebnisses mit einem externen Kriterium
- Beispiel: Varianzaufklärung von Psychologiestudienleistungen durch Studieneingangstest

Validität

Konstruktvalidität

Grad der wissenschaftlichen Fundierung des Schlusses von Testwerten auf Merkmale

Zur Erinnerung: [Validität in Einführung in die psychologischen Forschungsmethoden](#)

Struktursuchendes Vorgehen

- Exploratorische Faktoranalyse zur Inspiration einer Faktorentdeckung
- Konvergente Validität (Übereinstimmung mit Messung ähnlicher Merkmale)
- Diskriminatorische Validität (Abgrenzung zur Messung anderer Merkmale)

Strukturprüfendes Vorgehen

- Faktorielle Validität mithilfe konfirmatorischer Faktoranalysen
- Homogenitätsprüfungen mithilfe von Item-Response-Theorie-Modellen
- Strukturgleichungsmodellevaluation

Moosbrugger and Kelava (2012)

Beispiel BDI-II

Konvergente und diskriminante Validität

In Tabelle 3 sind Studien zum Zusammenhang des BDI II mit anderen Selbst- und Fremdbeurteilungsskalen aufgeführt. Wie daraus hervorgeht, korreliert das Verfahren mit anderen Depressionsskalen zwischen $r = .68$ bis $r = .89$, etwas geringer mit selbst und fremdbeurteilter Angst/Ängstlichkeit ($.37 \leq r \leq .60$) sowie im ähnlichen Bereich mit konstrukt-nahen kognitiven Skalen ($.37 \leq r \leq .68$). Die Differenzierungsfähigkeit bezüglich Depression und Angst anhand des BDI II ist eingeschränkt und abhängig von der Methode: fremdbeurteilte Angst anhand der HRSA korreliert niedriger mit dem BDI II als selbstbeurteilte Angst anhand des BAI (vgl. Tabelle 3). Bislang fehlen Studien, die Zusammenhänge zwischen BDI II und konstruktfernen Verfahren (z.B. symptomferne Persönlichkeitstraits) erfassen.

Faktorielle Validität

In der Untersuchung von Beck et al. (1996) legten Scree-Tests aus explorativen Faktorenanalysen sowohl für die Patientenstichprobe als auch für die Studentenstichprobe die Extraktion zweier Faktoren nahe. Nach parallel durchgeführten Hauptachsenanalysen mit anschließender obliquer (Promax-)Rotation resultierte für die Patientenstichprobe ein Faktor, den die Autoren als "somatisch-affektiven" Faktor bezeichneten, auf einem zweiten Faktor luden v.a. kognitive Items des BDI-II, er wurde demnach als "kognitiver" Faktor bezeichnet. Die beiden obliquen Faktoren korrelierten miteinander zu $r = 0.66$. In der Studentenstichprobe repräsentierte dagegen der erste Faktor die "kognitiv-affektive", der zweite Faktor die "somatische" Dimension. Hier betrug die Korrelation der beiden Faktoren $r = 0.62$.

Beispiel GEB

Validität

Obwohl sich das mehrdimensionale Rasch-Modell dem eindimensionalen Rasch-Modell bei der mittels GEB erfassten Neigung, sich umweltschützend zu verhalten, als formal überlegen erwies, ist der Informationsverlust des eindimensionalen Modells vernachlässigbar gering. So lagen die messfehlerkorrigierten Korrelationen zwischen den Verhaltensneigungen in den verschiedenen Verhaltensdomänen in der Regel jenseits von .50 und im Bereich der Reliabilitäten der einzelnen Skalen. Entsprechend erscheint es nicht nur machbar, sondern aus Sparsamkeitsgründen auch ratsam, Umweltschutzneigung, wie beabsichtigt, als eine mehr oder weniger konsistente, allgemeine Verhaltensneigung zu erfassen (Kaiser & Wilson, 2004).

Beispiel KoKi

Validität

Konvergente Konstruktvalidität: Korrelationen der "Konzentrationsleistung" im KoKi mit dem d2-R (.69), der Skala "Aufmerksamkeit-selektiv" der IDS (.58), der "Verarbeitungsgeschwindigkeit" der WISC-IV (.55), des "Wahrnehmungstempus" im CFT 1-R (.57). Korrelationen der "Genauigkeit" im KoKi mit Fehlerwerten im d2-R (.48) und der IDS (.40). Zusammenhänge des KoKi mit der Lehrerfremdbeurteilung ADHS-FBB im DISYPS II zwischen .22 und .33 (Korrelationsbeträge). Erwartungsgemäßer Leistungsunterschied im KoKi zwischen Kindern mit und ohne ADHS (Δ ca. 0.5 SD).

Beziehungen zwischen den drei Hauptgütekriterien

Objektivität $\Rightarrow$ Reliabilität $\Rightarrow$ Validität

- Ein Test, der in anderer Testsituation andere Resultate ergibt, kann nicht reliabel sein
- Ein Test, der ein Konstrukt zu unreliabel misst, kann nicht sehr valide sein

Reliabilitäts-Kriteriumsvaliditäts-Dilemma der klassischen Testtheorie

- Abnahme der Kriteriumsvalidität bei Erhöhung der Reliabilität
- Einstieg in Rost (2004), Kapitel 6.4.3

Rost (2004)

Nebengütekriterien

Skalierung

Ein Test erfüllt das Gütekriterium der Skalierung, wenn die laut Verrechnungsregel resultierenden Testwerte die qualitativen Merkmalsrelationen adäquat abbilden. Formale Skalierungseigenschaften werden meist (nur) im Rahmen der Item-Response-Theorie überprüft.

Testökonomie

Ein Test erfüllt das Gütekriterium der Ökonomie, wenn er, gemessen am diagnostischen Erkenntnisgewinn, relativ wenig finanzielle und zeitliche Ressource beansprucht.

Nützlichkeit

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

Moosbrugger and Kelava (2012)

Nebengütekriterien

Zumutbarkeit

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

Unverfälschbarkeit

Ein Testverfahren erfüllt das Gütekriterium der Unverfälschbarkeit, wenn das Verfahren derart konstruiert ist, dass die zu testende Person durch gezieltes Testverhalten die konkreten Ausprägungen ihrer Testwerte nicht steuern bzw. verzerren kann.

Fairness

Ein Test erfüllt das Gütekriterium der Fairness, wenn die resultierenden Testwerte zu keiner systematischen Benachteiligung bestimmter Personen aufgrund ihrer Zugehörigkeit zu ethnischen, soziokulturellen oder geschlechtsspezifischen Gruppen führen.

Moosbrugger and Kelava (2012)

Definition

Testarten

Antwortformate

Testgütekriterien

Testtheorien

Selbstkontrollfragen

Testtheorien

- Probabilistische Modelle zum Zusammenhang von zu erfassendem Konstrukt und Testverhalten
- Grundlage der Fragebogendatenanalyse

Hauptrichtungen

- Klassische Testtheorie
- Faktoranalyse
- Item-Response-Theorien

Besonderheiten

- Klassische Testtheorie und Faktoranalyse sind Modelle mit latenten Zufallsvariablen
- Item-Response-Theorien können latente Zufallsvariablen enthalten

Standarddatenstruktur der Testtheorie und Fragebogendatenanalyse

Beobachtung von Werten von m Items von n Testpersonen

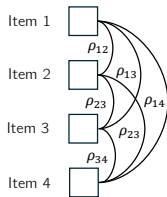
Table 1: 20 Testpersonen $\times$ 10 Items (Antwortskala 1–5)

Testperson	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10
1	1	5	3	4	2	1	1	4	3	2
2	4	2	2	4	2	3	4	2	4	5
3	1	2	2	1	5	3	1	2	5	5
4	2	1	5	5	2	3	4	4	4	4
5	5	4	2	5	5	3	2	5	1	4
6	3	1	1	1	4	4	5	1	1	1
7	2	4	3	1	5	1	2	5	5	2
8	3	3	3	3	4	1	2	5	5	4
9	3	2	4	2	1	4	2	5	1	1
10	1	2	3	2	3	2	3	4	5	1
11	5	4	1	3	2	1	1	4	4	3
12	5	4	4	2	3	2	2	4	2	1
13	2	4	5	4	3	3	3	2	2	1
14	2	2	1	3	1	4	3	1	1	4
15	1	4	1	5	5	1	3	3	5	3
16	5	1	4	2	4	3	3	1	5	5
17	5	1	5	2	4	5	5	1	2	3
18	1	4	5	1	1	3	3	4	5	5
19	1	1	4	3	5	4	1	5	2	1
20	5	2	5	3	5	2	2	1	4	1

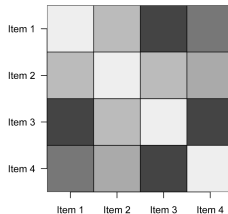
Standardproblem der Testtheorie und Fragebogendatenanalyse

Erklärung der beobachteten Itemkorrelationsstruktur durch latente Variablen

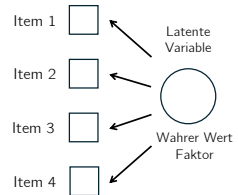
Empirische Perspektive



Kovarianzperspektive



Modellbasierte Perspektive



Klassische Testtheorie

- Probabilistisches Modell von Item- und Summenwerten eines Tests
- Zentrale Beiträge von Gulliksen (1950) und Lord and Novick (1968)
- Latentes Variablenmodell der Form

$$\text{Beobachteter Wert} = \text{Wahrer Wert} + \text{Fehler} \quad (1)$$

- Meist als “Messfehlermodell” beschrieben
- Erklärung von Antwortmustern durch einen nicht-direkt observierbaren True Score
- Modellierung von intra- und interindividueller Variabilität
- Grundlage der quantitativen Reliabilitätsbeurteilung

⇒ Paralleltest-Reliabilität, Cronbach's α , Spearman-Brown-Formel

Faktoranalysen

- Probabilistisches Modell der Kovarianzeigenschaften von Items eines Tests
- Zentrale Beiträge von Spearman (1904), Hotelling (1933), Lawley (1940)
- Latentes Variablenmodell der Form

$$\text{Latenter Faktorwerte} = \text{Latenter Fehler} \quad (2)$$

$$\text{Beobachteter Werte} = \text{Ladungen} \cdot \text{Latenter Faktorwerte} + \text{Fehler}$$

- Erklärung von Antwortmustern durch wenige nicht-direkt observierbare Faktoren
- Exploratorische und konfirmatorische Varianten

⇒ Bestimmung der faktoriellen Validität

Item-Response-Theorien

- Probabilistisches Modell von Itemantworten als Funktion von latenter Zustände
- Zentrale Beiträge von Rasch (1960), Novick (1966), Lord (1980)
- Verschiedene mathematische Formen der logistischen Regression

$$\mathbb{P}(\text{Itemantwort}) = f(\text{Testparameter}, \text{Personenparameter}) \quad (3)$$

- Einsatz und Entwicklung insbesondere bei Leistungstests
- Zur Bedeutung im klinischen Kontext, siehe Reise and Waller (2009)

⇒ Anwendung insbesondere im Bereich des adaptiven Leistungstestens

Definition

Testarten

Antwortformate

Testgütekriterien

Testtheorien

Selbstkontrollfragen

Selbstkontrollfragen I

1. Erläutern Sie den Begriff des psychologischen Tests nach Lienert (1968) und kommentiert von Rost (2004).
2. Nennen Sie die typischen Material- und Durchführungsbestandteile eines Tests.
3. Erläutern Sie die Begriffe Item, Itemstamm, Antwortformat.
4. Erläutern Sie den Begriff des psychometrischen Persönlichkeitstests.
5. Nennen Sie vier Formen psychometrischer Persönlichkeitstests.
6. Erläutern Sie den Begriff des Leistungstests.
7. Nennen Sie vier Formen von Leistungstests.
8. Erläutern Sie den Begriff des Persönlichkeitsentfaltungsverfahrens.
9. Nennen Sie drei Formen von Persönlichkeitsentfaltungsverfahren.
10. Erläutern Sie den Begriff des klinischen Tests.
11. Erläutern Sie den Begriff des Einstellungstests.
12. Erläutern Sie den Begriff des freien Antwortformats und nennen Sie zwei Beispiele.
13. Erläutern Sie Vor- und Nachteile freier Antwortformate.
14. Erläutern Sie den Begriff des gebundenen Antwortformats und nennen Sie drei Beispiele.
15. Erläutern Sie den Begriff der Ordnungsaufgabe und nennen Sie zwei Beispiele.
16. Erläutern Sie den Begriff der Auswahlaufgabe.
17. Erläutern Sie die Begriffe der Distraktoren, der Disjunktion und der Exhaustivität bei Auswahlaufgaben.
18. Erläutern Sie Vor- und Nachteile von Auswahlaufgaben.
19. Erläutern Sie den Begriff der Forced-Choice-Auswahlaufgabe.
20. Erläutern Sie den Begriff der Beurteilungsaufgabe.
21. Erläutern Sie die Begriffe des forcierten und unforcierten Ratings bei Beurteilungsaufgaben.
22. Erläutern Sie den Begriff der Ausweichkategorie bei Beurteilungsaufgaben.
23. Erläutern Sie den Begriff der Antwortskalenpolarität bei Beurteilungsaufgaben.

24. Erläutern Sie den Begriff der Antwortskalenbalance bei Beurteilungsaufgaben.
25. Erläutern Sie den Begriff der Antwortskalenverankerung bei Beurteilungsaufgaben.
26. Erläutern Sie Vor- und Nachteile von Beurteilungsaufgaben.
27. Erläutern Sie den Begriff des Testgütekriteriums.
28. Nennen Sie die drei Hauptgütekriterien für psychologische Tests.
29. Nennen Sie sieben Nebengütekriterien für psychologische Tests.
30. Geben Sie die Definition der Objektivität nach Moosbrugger and Kelava (2012) wieder.
31. Erläutern Sie die Begriffe Durchführungsobjektivität, Auswertungsobjektivität, Interpretationsobjektivität
32. Geben Sie die Definition der Reliabilität nach Moosbrugger and Kelava (2012) wieder.
33. Erläutern Sie die Begriffe der Retest-, Paralleltest- und Split-Half-Reliabilität
34. Erläutern Sie den Begriff der internen Konsistenz eines psychologischen Tests.
35. Geben Sie die Definition der Validität nach Moosbrugger and Kelava (2012) wieder.
36. Erläutern Sie die Begriffe der Augenschein-, Inhalts- und Kriteriumsvalidität.
37. Erläutern Sie den Begriff der Konstruktvalidität eines psychologischen Tests.
38. Nennen Sie die drei Hauptrichtungen der Testtheorie.
39. Erläutern Sie die Standarddatenstruktur der Testtheorie und Fragebogendatenanalyse.
40. Erläutern Sie das Standardproblem der Testtheorie und Fragebogendatenanalyse.
41. Erläutern Sie den Begriff der Klassischen Testtheorie.
42. Erläutern Sie den Begriff der Faktoranalyse.
43. Erläutern Sie den Begriff der Item-Response-Theorien.

- Beck, A. T. 1961. "An Inventory for Measuring Depression." *Archives of General Psychiatry* 4 (6): 561. <https://doi.org/10.1001/archpsyc.1961.01710120031004>.
- Beck, Aaron T., Robert A. Steer, Roberta Ball, and William F. Ranieri. 1996. "Comparison of Beck Depression Inventories-IA and-II in Psychiatric Outpatients." *Journal of Personality Assessment* 67 (3): 588–97. https://doi.org/10.1207/s15327752jpa6703_13.
- Bühner, Markus. 2010. *Einführung in die Test- und Fragebogenkonstruktion*. 2., aktualisierte und erw. Aufl., [Nachdr.]. Pearson Studium Psychologie. München: Pearson Studium.
- Campbell, Donald T. 1963. "Social Attitudes and Other Acquired Behavioral Dispositions." In *Psychology: A Study of a Science. Study II. Empirical Substructure and Relations with Other Sciences. Volume 6. Investigations of Man as Socius: Their Place in Psychology and the Social Sciences*, 94–172. New York, NY, US: McGraw-Hill. <https://doi.org/10.1037/10590-003>.
- Gulliksen, Harold. 1950. *Theory of Mental Tests*. Hoboken: John Wiley & Sons Inc. <https://doi.org/10.1037/13240-000>.
- Hautzinger, M, F. Keller, and C. Kühner. 2006. *BDI-II Beck Depressions-Inventar*. Pearson.
- Hotelling, Harold. 1933. "Analysis of Complex Variables into Principal Components." *Journal of Educational Psychology* 24: 417-441 and 498-520.
- Kaiser, F. G. 1998. "A General Measure of Ecological Behavior¹." *Journal of Applied Social Psychology* 28 (5): 395–422. <https://doi.org/10.1111/j.1559-1816.1998.tb01712.x>.
- . 2020. "GEB-50. General Ecological Behavior Scale." ZPID (Leibniz Institute for Psychology Information) – Open Test Archive. <https://doi.org/10.23668/PSYCHARCHIVES.3453>.

Referenzen II

- Kaiser, Florian, Katarzyna Byrka, and Terry Hartig. 2010. "Reviving Campbell's Paradigm for Attitude Research." *Personality and Social Psychology Review* 14 (4): 351–67. <https://doi.org/10.1177/1088868310366452>.
- Krauth, Joachim. 1995. *Testkonstruktion und Testtheorie*. Weinheim: Beltz, Psychologie Verl.-Union.
- Lawley, Derrick. 1940. "The Estimation of Factor Loadings by the Method of Maximum Likelihood." *Proceedings of the Royal Society of Edinburgh. Section B: Biological Sciences*.
- Lienert, Gustav A. 1968. *Testaufbau und Testanalyse*. 4th ed. München Weinheim: Psychologie Verlags Union.
- Lienert, Gustav A., and Ulrich Raatz. 1998. *Testaufbau und Testanalyse*. 6. Auflage. Weinheim: Beltz, Psychologie Verlags Union.
- Lord, Frederic M. 1980. *Applications of Item Response Theory To Practical Testing Problems*. 0th ed. Routledge. <https://doi.org/10.4324/9780203056615>.
- Lord, Frederic M., and Melvin R. Novick. 1968. *Statistical Theories of Mental Test Scores*. Nachdr. der Ausg. Reading, Mass. [u.a.], 1968. The Addison-Wesley Series in Behavioral Science: Quantitative Methods. Charlotte, NC: Information Age Publ.
- Moosbrugger, Helfried, and Augustin Kelava, eds. 2012. *Testtheorie und Fragebogenkonstruktion: mit 66 Abbildungen und 41 Tabellen*. 2., aktualisierte und überarbeitete Auflage. Springer-Lehrbuch. Berlin Heidelberg: Springer.
- Novick, Melvin R. 1966. "The Axioms and Principal Results of Classical Test Theory." *Journal of Mathematical Psychology* 3 (1): 1–18. [https://doi.org/10.1016/0022-2496\(66\)90002-2](https://doi.org/10.1016/0022-2496(66)90002-2).
- Pospeschill, Markus. 2022. *Testtheorie, Testkonstruktion, Testevaluation*. 2nd ed. München: Ernst Reihnhardt Verlag.

- Rasch, G. 1960. *Probabilistic Models for Some Intelligence and Attainment Tests*. Expanded ed. Chicago: University of Chicago Press.
- Reise, Steven P., and Niels G. Waller. 2009. "Item Response Theory and Clinical Measurement." *Annual Review of Clinical Psychology* 5 (1): 27–48. <https://doi.org/10.1146/annurev.clinpsy.032408.153553>.
- Revers, W. J., and K Taeber. 1968. *Der Thematische Apperzeptionstest*. Bern: Huber.
- Rorschach, Hermann. 1921. *Psychodiagnostik: Methodik und Ergebnisse eines wahrnehmungsdiagnostischen Experiments*. Arbeiten zur angewandten Psychiatrie. Bern: Bircher.
- Rost, Jürgen. 2004. *Lehrbuch Testtheorie, Testkonstruktion*. 2., vollständig überarbeitete und erweiterte Aufl. Aus dem Programm Huber. Bern: H. Huber.
- Schmidt-Atzert, Lothar, and Karin Funsch. 2023. *KoKi - Konzentrationstest für Kinder*. 1st ed. Hogrefe Verlag.
- Spearman, C. 1904. ""General Intelligence," Objectively Determined and Measured." *The American Journal of Psychology* 15 (2): 201. <https://doi.org/10.2307/1412107>.
- Warren, Howard C. 1934. *Dictionary of Psychology*. Routledge Revivals Ser. Milton: Routledge.