Testtheorie und Testkonstruktion

MSc Klinische Psychologie und Psychotherapie

SoSe 2024

Prof. Dr. Dirk Ostwald

(1) Tests

Tests

Testgütekriterien

Testmodelle

Selbstkontrollfragen

Approbationsordnung für Psychotherapeutinnen und Psychotherapeuten (2020) Anlage 2 (zu § 8 Nummer 2)

Inhalte, die im Masterstudiengang im Rahmen der hochschulischen Lehre zu vermitteln und bei dem Antrag auf Zulassung zur psychotherapeutischen Prüfung nachzuweisen sind.

Vertiefte psychologische Diagnostik und Begutachtung

a) Die studierenden Personen entwickeln und bewerten psychodiagnostische Verfahren nach aktuellen testtheoretischen Modellen.

Zur Vermittlung der Inhalte der vertieften psychologischen Diagnostik und Begutachtung sind bei der Planung der hochschulischen Lehre mindestens 7 ECTS-Punkte vorzusehen und die folgenden Wissensbereiche abzudecken:

a) Diagnostische Modelle und Methoden.

Definition

Definition eines **Tests** nach Moosbrugger und 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 und 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 Variable 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."

Klassifikation

Leistungstests

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

Psychometrische Persönlichkeitstests

- Klinische Tests
- Persönlichkeitsstrukturtests
- Einstellungstests
- Interessentests

Persönlichkeitsentfaltungs-Verfahren

- Formdeuteverfahren
- Verbal-thematische Verfahren
- Zeichnerische- und Gestaltungsverfahren

Items

- Grundbausteine eines psychologischen Tests
- Reize, auf die man eine Reaktion erwartet und zu registrierende Reaktionen
- Standardbeispiel: Frage und Antwortmöglichkeiten
- 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

Beck-Depressionsinventar (BDI-II)

- Selbstbeurteilungsinstrument zur Erfassung der Schwere einer depressiver Symptomatik
- Entwicklung von Aaron T. Beck u. a. (1996)
- Nachfolger des Beck Depressionsinventar A. T. Beck (1961)
- Deutschsprachige Version nach Hautzinger, Keller, und 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
$\geq$ 29	schweres depressives Syndrom

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

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

Tests

Testgütekriterien

Testmodelle

Selbstkontrollfragen

Testgütekriterien

Hauptgütekriterien

1. Objektivität
2. Reliabilität
3. Validität

Nebengütekriterien

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

Moosbrugger und Kelava (2012)

Testgütekriterien

(1) 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.

(2) Reliabilität

Ein Test ist dann *reliabel*, wenn er das Merkmal, das er misst, exakt, d.h. ohne Messfehler misst. Vor dem Hintegrund der Klassischen Testtheorie unterscheidet man die *Retest-Reliabilität*, die *Paralleltest-Reliabilität*, die *Testhalbierungs-Reliabilität (Split-Half-Reliabilität)* und die *Innere Konsistenz* eines Tests.

(3) Validität

Ein Test gilt dann als valide, wenn er das Merkmal, das er messen soll, auch wirklich misst und nicht irgendein anderes. Man unterscheidet die *Inhaltsvalidität* auf Grundlage logischer und fachlicher Überlegungen, die *Augenscheinvalidität* aufgrund der Intuition, die *Kriteriumsvalidität* durch Vergleich mit externen Kriterien und die *Konstruktvalidität*, meist aufgrund testtheoretischer Annahmen und Modelle wie zum Beispiel Faktorenanalysen.

Moosbrugger und Kelava (2012)

Beck-Depressionsinventar (BDI-II)

Durchführungsobjektivität

Schriftliche Instruktion des BDI-II Manuals

"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 gleichermaBen 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 O 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 O und 63 Punkten annehmen."

Ergebnisinterpretation

- Alle Selbstbeurteilungen weisen eine Tendenz zur Ergebnisverzerrung auf.
- Klinisch relevant ist es, auf spezifische Iteminhalte (Suizidalität) zu achten.

Hautzinger, Keller, und Kühner (2006)

Nebengütekriterien

(4) 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.

(5) Normierung (Eichung)

Unter der Normierung (Eichung) eines Tests versteht man das Erstellen eines Bezugssystems, mit dessen Hilfe die Ergebnisse einer Testperson im Vergleich zu den Merkmalsausprägungen anderer Personen eindeutig eingeordnet und interpretiert werden können. Meist nutzt man für die Bestimmung von Normwerten eine sogenannte Eichstichprobe.

(6) Testökonomie

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

Moosbrugger und Kelava (2012)

Nebengütekriterien

(7) 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.

(8) 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.

(9) 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.

(10) 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 geschlechtspezifischen Gruppen führen.

Moosbrugger und Kelava (2012)

Tests

Testgütekriterien

Testmodelle

Selbstkontrollfragen

Klassische Testtheorie

- Probabilistisches Modell von Item- und Summenwerten eines Tests
- Zentrale Beiträge von Gulliksen (1950) und F. Lord und Novick (1968)
- Messfehlermodell der Form

$$\text{Observed Score} = \text{True Score} + \text{Error Score} \tag{1}$$

- Modellierung von intra- und interindividueller Variabilität
- Grundlage für die quantitative Reliabiliätsbeurteilung

$$\Rightarrow \text{Paralleltestreliabiliät, Spearman-Brown-Formel, Cronbach's } \alpha$$

Faktorenanalyse

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

$$\begin{aligned} \text{Latent Variable} &= \text{Zero} + \text{Latent Error} \\ \text{Observed Variable} &= \text{Loadings} \cdot \text{Latent Variable} + \text{Observation Error} \end{aligned} \tag{2}$$

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

$\Rightarrow$ Bestimmung der faktoriellen Validität

Item-Response-Theorie

- Probabilistisches Modell von Itemantworten als Funktion von Fähigkeiten/Zustand
- Zentrale Beiträge von Rasch (1960), Novick (1966), F. M. Lord (1980)
- Verschiedene mathematische Formen der Logistischen Regression
- Einsatz und Entwicklung insbesondere bei Fähigkeitstests
- Bezüge zur Messtheorie nach Stevens und zum Rasch-Modell
- Zur Bedeutung im Klinischen Kontext, siehe Reise und Waller (2009)

$\Rightarrow$ Kein Thema des Seminars im SoSe 2024 und im BDI-II Manual

Tests

Testgütekriterien

Testmodelle

Selbstkontrollfragen

Selbstkontrollfragen

1. Erläutern Sie den Begriff des psychologischen Tests.
2. Nennen und erläutern Sie die Hauptklassen psychologischer Tests.
3. Erläutern Sie den Begriff des Items.
4. Skizzieren Sie die Modelle der Klassischen Testheorie, der Faktoranalyse und der Item-Response-Theorie.
5. Nennen und erläutern Sie die drei Hauptgütekriterien psychologischer Tests.
6. Nennen und erläutern Sie sieben Nebengütekriterien psychologischer Tests.

Referenzen I

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, und 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.

Gulliksen, Harold. 1950. *Theory of Mental Tests*. Hoboken: John Wiley & Sons Inc. https://doi.org/10.1037/13240-000.

Hautzinger, M, F. Keller, und 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.

Krauth, Joachim. 1995. *Testkonstruktion und Testtheorie*. Weinheim: Beltz, Psychologie Verl.-Union.

Lawley, N. D. 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., und Ulrich Raatz. 1998. *Testaufbau und Testanalyse*. 6. Auflage. Weinheim: Beltz, Psychologie Verlags Union.

Lord, F. M. 1980. *Applications of Item Response Theory To Practical Testing Problems*. 0. Aufl. Routledge. https://doi.org/10.4324/9780203056615.

Referenzen II

Lord, F., und 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, und Augustin Kelava, Hrsg. 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.

Rasch, G. 1960. *Probabilistic Models for Some Intelligence and Attainment Tests*. Expanded ed. Chicago: University of Chicago Press.

Reise, Steven P., und 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.

Rost, Jürgen. 2004. *Lehrbuch Testtheorie, Testkonstruktion*. 2., vollständig überarbeitete und erweiterte Aufl. Aus dem Programm Huber. Bern: H. Huber.

Spearman, C. 1904. „"General Intelligence," Objectively Determined and Measured“. *The American Journal of Psychology* 15 (2): 201. https://doi.org/10.2307/1412107.