TECHNISCHE UNIVERSITÄT
IN DER KULTURHAUPTSTADT EUROPAS
CHEMNITZ

Professur Psychologie digitaler Lernmedien
Institut für Medienforschung Philosophische Fakultät

Statistik I

Testgütekriterien



Mr Bean Takes An Exam (1991). 20th Century Fox.

Prof. Dr. Günter Daniel Rey

Überblick

- Einleitung
- Objektivität
- Reliabilität
- Validität
- Nebengütekriterien

Einleitung (z. B. Rey, 2020)

- Testgütekriterien als „Beurteilungskriterien“ für eine Messung bzw. einen spezifischen Test
  - Zur Bewertung bzw. zum Vergleich spezieller standardisierter Tests einsetzbar
  - Ebenso zur Bewertung bzw. zum Vergleich verschiedener Datenerhebungsmethoden nutzbar
- Zusammenhang: Gütekriterien hängen miteinander zusammen
- In der Regel gilt: Objektivität > Reliabilität > Validität
- Korrelationen: Viele Testgütekriterien basieren auf Korrelationen

Objektivität (z. B. Rey, 2020)

- Objektivität: Beobachterunabhängigkeit
- Durchführungsobjektivität: Ergebnisse unabhängig vom Testleiter
- Auswertungsobjektivität: Ergebnisse unabhängig vom Testauswerter
- Interpretationsobjektivität: Ergebnisinterpretation unabhängig von der Person, die diese vornimmt

Objektivität



Quelle: Twilight – Biss zum Morgengrauen, Summit Entertainment, Temple Hill Entertainment, Maverick Films.

Objektivität

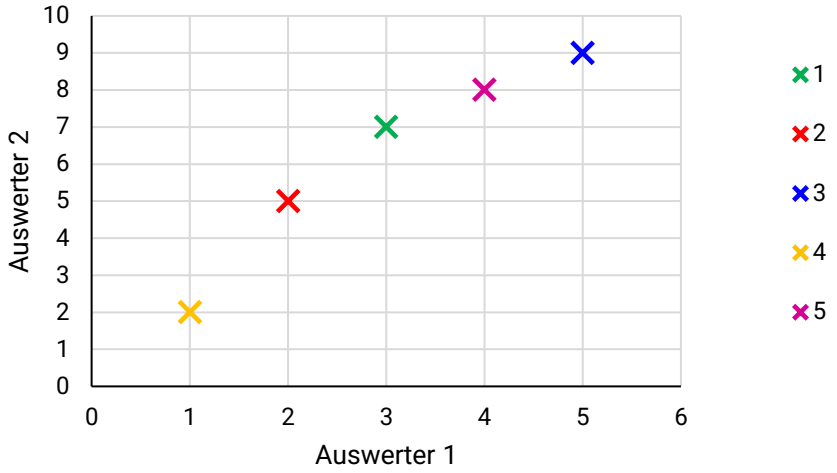
Welche Art der Objektivität wurde in dem Video dargestellt?

- ☐ A: Durchführungsobjektivität
- ☐ B: Auswertungsobjektivität
- ☐ C: Interpretationsobjektivität

Auswertungsobjektivität

- Beispiel zur Bestimmung der Auswertungsobjektivität
- Offene Lerntransferfragen werden von zwei Auswertern bewertet:

VPN	Auswerter 1	Auswerter 2
1	3	7
2	2	5
3	5	9
4	1	2
5	4	8



- Korrelation zwischen den beiden Auswertern: r = .97
- Problem: Korrelationen erfassen keine Unterschiede der Bewertungsstrenge
- Lösung: Verwendung varianzanalytischer Pläne

Reliabilität (z. B. Rey, 2020)

- Reliabilität: Zuverlässigkeit bzw. Genauigkeit einer Messung
- Paralleltestreliabilität: Korrelation zwischen den Ergebnissen zweier ähnlicher Testformen, die zeitnah an derselben Stichprobe erhoben wurden
- Retestreliabilität: Korrelation zwischen zwei Ergebnissen des gleichen Tests, die zu zwei unterschiedlichen Zeitpunkten an derselben Stichprobe erhoben wurden
- Interne Konsistenz: Vergleich der einzelnen Aufgaben bzw. Items eines Tests (Prüfung auf Homogenität)
- Testhalbierungsreliabilität: Korrelation zwischen zwei Hälften des gleichen Tests

Cronbachs Alpha

- Koeffizient Cronbachs α sehr häufig als Maß zur internen Konsistenz eines Messinstrumentes verwendet
- Formel:

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

N = Anzahl an Items bzw. Subskalen
$\bar{r}$ = Durchschnittliche Korrelation zwischen den Items bzw. Subskalen

- Beispiel: Bei einem Test zur kognitiven Belastung mit fünf Items und einer durchschnittlichen Korrelation von r = .3 ergibt sich:

$$\alpha = \frac{5 \cdot 0.3}{1 + (5 - 1) \cdot 0.3} = \frac{1.5}{2.2} \approx 0.68$$

- Gängige Konvention: Bei α > .7 gilt ein Messinstrument als reliabel

Cronbachs Alpha

Wie hoch ist Cronbachs α für eine durchschnittliche Korrelation der Items von r = .2 bei einer Itemanzahl von einmal N = 5 und einmal N = 20?

- ☐ A: Für N = 5 ist α = .56, für N = 20 ist α = .83
- ☐ B: Für N = 5 ist α = .17, für N = 20 ist α = .19
- ☐ C: Für N = 5 ist α = .83, für N = 20 ist α = .56
- ☐ D: Für N = 5 ist α = .19, für N = 20 ist α = .17

Cronbachs Alpha (z. B. Rey, 2020; McNeish, 2018)

- Kritik am Koeffizienten Cronbachs α
  - Itemanzahl: Höhe des Kennwertes stark abhängig von der Itemanzahl des Messinstrumentes
  - Unidimensionalität: Kein Beleg für die Unidimensionalität (Eindimensionalität) eines Messinstrumentes
- Beispiel: Zusammenhang zwischen Itemanzahl und Cronbachs α bei einer durchschnittlichen Korrelation von *r* = .1 zwischen den Items:

Itemanzahl	5	10	15	20	25	30	35	40	45	50
Cronbachs α	.36	.53	.63	.69	.74	.77	.80	.82	.83	.85

- Alternativen: Anstelle des Koeffizienten Cronbachs α sollte andere Maße der internen Konsistenz wie Revelle's omega total oder Greatest Lower Bound (GLB) verwendet werden (McNeish, 2018)

Trennschärfe (Kelava & Moosbrugger, 2020)

- Trennschärfe: Korrelation zwischen einem Item und dem Gesamtwert des Tests
- Formel:

$$r_{it} = r_{(x_{vi}, x_v)}$$

x_{vi} = Itemwert i der Person v
x_v = Gesamtwert der Person v (ggf. abzüglich des jeweiligen Items)

- Differenzierung: Trennschärfe gibt an, wie stark die Differenzierung des jeweiligen Items mit der Differenzierung des Gesamtwertes übereinstimmt
- Gängige Konvention: Bei $0.4 \leq r_{it} \leq .7$ gilt die Trennschärfe als gut
- Trennschärfe ohne vs. mit „part-whole-correction“

Trennschärfe

- Beispiel: Berechnung der Trennschärfe ohne vs. mit „part-whole-correction“
- Trennschärfe ohne „part-whole-correction“: r = .44
- Trennschärfe mit „part-whole-correction“: r = .36

VPN	$IQ_{Einzelitem}$	IQ_{Gesamt}	$IQ_{Gesamt - Item}$
Sheldon	0.3	9.5	9.2
Leonard	0.9	6.5	5.6
Howard	0.4	4.5	4.1
Rajesh	0.6	8.5	7.9
Penny	0.1	1.5	1.4

- Überschätzung: Vor allem bei wenigen Items kann es ohne „part-whole-correction“ zur Überschätzung der Trennschärfe kommen

Validität (z. B. Rey, 2020)

- Validität: Gültigkeit der Messung
  - Misst der Test das, was er messen soll?
  - Nur bei einem validen Test sind die Messergebnisse interpretierbar
- Inhaltliche (oder logische) Validität: Aufgaben des Tests sind inhaltlich identisch mit den Merkmalen, die durch den Test erfasst werden sollen
  - Begründung dieser Validitätsform argumentativ und nicht empirisch-numerisch

Validität (z. B. Rey, 2020)

- Konstruktvalidität: Test erfasst alle Facetten des theoretischen Konstrukts, die erfasst werden sollen
  - Konvergente Validität: (Möglichst hohe) Korrelation zwischen verschiedenen Tests, die dasselbe Konstrukt messen
  - Diskriminante (bzw. divergente) Validität: (Möglichst niedrige) Korrelation zwischen verschiedenen Tests, die verschiedene Konstrukte messen
- Kriterienbezogene Validität: Testergebnis stimmt mit anderen, praktisch relevanten Kriterien (Außenkriterien) überein, die das Merkmal ebenfalls erfassen
  - Konkurrente Validität: Übereinstimmungsvalidität
  - Prognostische (prädiktive) Validität: Vorhersagevalidität

Exkurs: Interne und externe Validität (z. B. Rey, 2020)

- Wichtig: Validität der Testgütekriterien ≠ Validität experimenteller Versuchspläne
- Validität experimenteller Versuchspläne
  - Interne Validität: Veränderungen der abhängigen Variablen lassen sich eindeutig auf Variationen der unabhängigen Variablen zurückführen: UV → AV
  - Externe Validität: Generalisierbarkeit der Ergebnisse auf andere Kontexte (experimentelle Variablenoperationalisierungen, Situationen und Personengruppen)

Validität

Die Angestellten eines Unternehmens verhalten sich bei einem Brand vorbildlich, die in einem Lernprogramm zum Brandschutz gut abgeschnitten haben.

Um welche Form der Validität handelt es sich?

- ☐ A: Inhaltliche Validität
- ☐ B: Konvergente Validität
- ☐ C: Diskriminante Validität
- ☐ D: Konkurrente Validität
- ☐ E: Prognostische Validität
- ☐ F: Interne oder externe Validität

Nebengütekriterien (z. B. Rey, 2020)

- Skalierung: Adäquate Verrechnungsvorschrift
- Normierung: Geeignete, aktuelle Referenzstichprobe
- Testfairness: Keine systematische Diskriminierung
- Ökonomie: Dauer und Kosten der Erhebung gering
- Nützlichkeit: Praktische Relevanz des Merkmals; Beantwortung der Fragestellung möglich
- Zumutbarkeit: Nutzen überwiegt zeitliche, psychische und körperliche Belastung der Testpersonen

Nebengütekriterien (z. B. Rey, 2020)

- Vergleichbarkeit: Existenz von Paralleltestformen oder inhaltsähnlicher Tests
- Unverfälschbarkeit: Beispielsweise durch soziale Erwünschtheit gefährdet
- Transparenz: Verständlichkeit der Instruktion; Übungsitems im Vorfeld; angemessenes Feedback im Anschluss des Tests
- Akzeptanz: Erhebung bzw. Test von Laien akzeptiert
- Äußere Gestaltung: Sprachlich und optisch ansprechend; Anpassung an die Zielgruppe

Zusammenfassung

- Objektivität: „Beobachterunabhängigkeit“ mit Durchführungs-, Auswertungs- und Interpretationsobjektivität
- Reliabilität: Zuverlässigkeit bzw. Genauigkeit der Messung mit Unterteilung in interne Konsistenz, Testhalbierungs-, Paralleltest- und Retestreliabilität
- Cronbachs α: Häufig verwendete Angabe zur internen Konsistenz eines Messinstrumentes, aber aus methodischen Gründen fraglich
- Validität: Gültigkeit der Messung mit den Validitätsformen inhaltliche Validität, Konstruktvalidität und kriterienbezogene Validität
- Zahlreiche weitere Nebengütekriterien

Prüfungsliteratur

- Rey, G. D. (2020). *Methoden der Entwicklungspsychologie. Datenerhebung und Datenauswertung* (3., überarbeitete Auflage). Norderstedt BoD.

(Unter-)Kapitel	Taschenbuch	E-Book (ePUB)	Webseite
Testgütekriterien	S. 61–78	S. 52–65	S. 48–68

- Moosbrugger, H., & Kelava, A. (Hrsg.). (2020). *Testtheorie und Fragebogenkonstruktion* (3. Aufl.). Heidelberg: Springer.
  - Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien) (S. 7–26)

Weiterführende Literatur

- Bühner, M. (2021). *Einführung in die Test- und Fragebogenkonstruktion* (4. Aufl.). München: Pearson.
  - Kapitel 8: Haupt- und Nebengütekriterien
- Schmidt-Atzert, L., Krumm, S. & Amelang, M. (2022). *Psychologische Diagnostik* (6. Aufl.). Berlin: Springer.
  - Testgütekriterien (S. 132–198)
- Sedlmeier, P., & Renkewitz, F. (2024). *Forschungsmethoden und Statistik für Psychologen und Sozialwissenschaftler* (4. Aufl.). München: Pearson.
  - Gütekriterien beim Messen und Testen (S. 79–90)
- McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*, 412–433.