5. Evaluation psychologischer Testverfahren

Evaluation von Testverfahren als Grundlage für laufende Verbesserung & Aktualisierung bereits publizierter Tests.

5.1 Richtlinien & Beurteilungssysteme für Tests

Richtlinien	- richten sich unmittelbar an Testentwickler & -Anwender - Keine konkreten Bewertungshinweise
Beurteilungssystem	- Hinweise zur einheitlichen Beurteilung der Verfahren für Rezensenten

Richtlinien: Qualität

„Standards“ for educational and psychological Testung	- AREA & APA & NCME - Testfairness
DIN 33430	Anforderung an Verfahren und deren Einsatz bei berufsbezogener Eignungsbeurteilung
Principles for the validation and use of personell selection procedures	Eignungsdiagnostische Richtlinie
Guidelines for test Translation and Adaption	ITC: Richtlinie zur interkulturellen Übertragung
International guidelines on computer-based and internet delivered testing	ITC: Richtlinie für Spezifikation neuer Technik bei Testanwendung

Richtlinie: Berufsethik & verhalten

Ethical principles of Psychologists and code of conduct	APA: allgemeine Richtlinie
Berufsordnung für Psychologen	BDP: allgemeine Richtlinie
International guidelines for test use	ITC: allgemeine Richtlinie, Anwenderverhalten
Grundsätze für die Anwendung psychologischer Eignungsuntersuchung in Wirtschaft und Verwaltung	BDP: Eignungsdiagnostische Richlinie

Beurteilungssysteme

Buros	Grundlage amerikanischer Testrezensionen
COTAN	Grundlage Niederländisches Testrezensionen
TBS-TK	- Testbeurteilungssystem des Testkuratoriums - Bewertungsschema auf der Grundlage der DIN 33430
DIN Screen	TBS-TK ergänzende Checkliste zur Erfüllung der DIN-Kriterien
Gesamteuropäisches Beurteilungssystem	Review model for the description and evaluation of psychological tests

5.2 Psychometrische Gütekriterien

Objektivität	Durchführungsobjektivität Auswertungsobjektivität Interpretationsobjektivität	
Reliabilität	Reliabilitätsschätzung:	1. Interne Konsistenz 2. Retest-Reliabilität 3. Paralleltestreliabilität 4. Interpret-Reliabilität
Validität		

5.2.1 Objektivität & Reliabilität

Quantitative Maße der Objektivität: z.B. Indizes der Interraterreliabilität
---> Objektivität als Teilaspekt der Reliabilität

Durchführungsobjektivität	Ausmaß, in dem Testergebnisse unabhängig von unterschiedlichen Durchführungsgelegenheiten zustande kommen
Auswertungsobjektivität	Ausmaß, in dem verschiedene Auswerter desgleichen Testbogens zum selben Ergebnis kommen
Interpretationsobjektivität	Ausmaß, in dem „verschiedene“ Interpretationen gleicher Testergebnisse übereinstimmen

Reliablität spielt innerhalb der KTT zentrale Rolle, da diese Theorie des Messfehler ist und die Reliabilität das Ausmaß bezeichnet, in dem die Testergebnisse frei von unsystematischen Testfehlern zustande kommen.

Arten der Reliabilitätsschätzung: Es werden Werte für r_{tt} abgeschätzt/ Korrelationen eines Tests mit sich selbst geschätzt

Interne Konsistenz	- es entstehen miteinander korrelierte Messreihen, indem ein Test 1x erhoben wird und dann so aufgeteilt wird, dass die Testteile miteinander korreliert werden können - Die *Konsistenzkoeffizienten* schätzen das Ausmaß der Gemeinsamkeiten der Testteile - Fehler: Anwendung bei heterogenen Tests; problematisch bei Speedtests; Interpretation der internen Konsistenz als Homogenitätsindex - Einfachste Form: Split-half; Möglich ist auch eine Aufteilung in soviel Teile wie der Test Items hat - Standardmaß der internen Konsistenz: Cronbach-α
Retest-Reliabilität	- ein Test wird von der gleichen Personengruppe zweimal bearbeitet - Technisch ergibt sich Schätzung der Reliabilität ergibt sich als Korrelation der Messreihen - Inhaltlich bezieht sich Retest-Koeffizient auf die zeitliche Stabilität der Messungen - Anwendung auch bei Speedtests & heterogenen Messungen - Wahl des zeitlichen Intervalls?!
Paralleltestreliabilität	- Herstellung zwei unabhängiger, aber gleicher Tests und sie bei der gleichen Stichprobe einzusetzen - Reliabilitätskoeffizient = Korrelation der Messreihen - Anwendung auf heterogene Tests möglich
Interrater-Reliabilität	- Beurteilerübereinstimmung: Beurteilungen des gleichen Verfahrens durch verschiedene Personen sind äquivalent - Es gibt viele Maße zur Beurteilerübereinstimmung, je nach Zweck, Skalenniveau, Zusammensetzung von Beurteilen - --> z.B Intraklassenkorrelation (ICC)

Beziehung zwischen Testlänge & Reliabilität:

- *Verdopplung*: Die Varianz des neuen Tests entspricht der Summer der Varianzen + dem 2-fachen der Kovarianz der beiden alten Tests (in die neue Varianzsumme geht die Fehlervarianz 1-fach ein, die wahre Varianz geht 2-fach ein) --> der neue Test ist reliabler als der Alte
- *Testhalbierung*: die Korrelation zwischen den Hälften muss aufgewertet werden, um den wahren Wert der Reliabilität des gesamten Tests abzuschätzen --> Korrekturformel/ Spearman-Brown-Formel
- *Cronbach-α*: Generalisierung der Testhalbierungsmethode. Stellt im Grunde einen Mittelwert der Konsistenzkoeffizienten über alle denkbaren Aufteilungen des Tests dar. Liefert präzise Schätzung unter der Voraussetzung dass die Items essentiell tau-äquivalent gemessen wurden. Bei tau-kongerischer Messung ist Cronbach-α die Untergrenze der Reliabilität. Steigt die Varianzsumme, so erhöht sich die Reliabilität.

Intraklassenkorrelation:
- unterschiedliche Fälle der ICC ähneln den unterschiedlichen Graden der Äquivalenz
- --> es gibt verschiedene Koeffizienten, falls verschiedene Beurteiler nur gleiche Varianz der Urteile unterstellt wird *(justierte ICC)*, oder zusätzlich auch gleiche Mittelwert *(unjustierte ICC)*

Höhe der Reliabilitätskoeffizienten nach COTAN

Niveau 1	Wichtige Einzelfallentscheidungen: >0.80
Niveau 2	Weniger wichtige Einzelfallentscheidungen: >0.70
Niveau 3	Gruppenuntersuchungen: >0.60

SPSS

Intraklassenkorrelation	- A N A L Y S I E R E N - - > S K A L I E R U N G - - > RELIABILITÄTSANALYSE --> STATISTIKEN --> KORRELATIONSKOEFFIZIENZ IN KLASSEN sowie gewünschte ICC-Variante auswählen mittels MODELL & TYP
Non-paarmetrische Maße der Beurteilerübereinstimmung	- ANALYSIEREN --> DESKRIPTIVE STATISTIKEN --> KREUZTABELLEN --> STATISTIKEN

Reliabilitätsbestimmung in der PTT:
- *spezifische Objektivität:* Unabhängigkeit der Messwerte von der untersuchten Item-/ Personenstichprobe

Verallgemeinerung des Reliabilitätskonzepts der KTT: *Generalisierbarkeitstheorie*
- Varianzanalytischer Ansatz
- Gesamte Testvarianz wird in Komponenten zu Lasen bestimmter Varianzquellen & Kombinationen zerlegt
- Die Varianzkomponenten werden empirisch quantifiziert, indem - je nach Interesse - eine sinnvolle Auswahl möglicher Designs realisiert wird
- Es werden mindestens 2 aufeinander aufbauende Studien benötigt:
    - Explorative Generalisierbarkeitsstudie: Quantifiziert Komponenten, Abgrenzung Merkmalsbereich
    - Vertiefende Entscheidungsstudie

5.2.2 Validität

Zeigt sich durch fortlaufende Akkumulation wissenschaftlicher Evidenz: Es gibt lediglich Hinweise auf Validität die immer wieder neu gesammelt und bewertet werden müssen.

Inhaltsvalidität	- Übereinstimmung der Testinhalte mit zugrunde liegendem Merkmal des Tests - Wie gut repräsentieren einzelne Aufgaben den Merkmalsbereich eines Tests? - Empirische Bestimmung: subjektive Beurteilung
Konstruktvalidität	- Interpretation der Testergebnisse als Indikatoren theoretischer Konstrukte --> misst der Test, was er messen soll? - Empirische Bestimmung: 1) logisch-argumentativ, 2) experimentelle Prüfung, 3) korrelative Analysen
Faktorielle V.	- Bestätigung der intendierten faktoriellen Struktur eines Tests - Hinweise auf Homogenität sind nur dan ein positiver Validitätsindikator, wenn das Merkmal eindimensional sein sollte
Konvergente V.	Korrelation mit konstruktnahe Variablen sollte hoch sein
Diskriminante V.	Korrelation mit konstruktfernen Variablen sollte niedrig sein
Kriterienbezogene Validität	- Gültigkeit der Schlüsse aus einem Testergebnis auf ein praktisch relevantes Kriterium ausserhalb der Testsituation - Empirische Bestimmung: bivariater Korrelationskoeffizient r_{tc}

Abgrenzung des Merkmalsbereichs (Inhaltsvalidiät):

Theoriegeleitete Def.	Ableitung aus einer theoretischen Konstruktdefinition. Für jedes Item muss theoretisch explizit darstellbar sein, wie eine bestimmte Ausprägung des Konstrukts zu einer bestimmten Antwortausprägung führt/ beitragen soll
Operationale Def.	Betrifft besonders kriterienorientierte Leistungstests. Aufgabenpool entspricht einer Stichprobe aus dem Aufgabenuniversum, welches dem Merkmalsbereich zur Gänze entspricht. Das heißt die Verbindung von Merkmal & Testinhalt ist direkt und unmittelbar.

Schrittweise Überprüfung der Regeln im nomologischen Netz, als empirische Bestimmung der Konstruktvalidität (Cronbach & Meehl):
- Alle Aussagen einer idealen Theorie lassen sich durch Axiome zu Gesetzmäßigkeiten über Zusammenhänge zwischen latenten Konstrukten beschreiben
- --> latenter Bereich der Theorie: findet durch semantische Ableitung aus den Konstrukten eine Entsprechung im manifesten Bereich des Beobachtbaren
- Theoretischer & beobachtbarer Bereich & alle Verbindungslinien bilden ein nomologisches Netz

Prüfung konvergenter/ diskriminanter Validität:

Bivariate Korrelationen	- Inferenzstatistische Absicherung durch Prüfung von Null-/ Alternativhypthese. Hierbei werden Grenzwert für diskriminante und konvergente Validität definiert (Maximal: 0-1)
Attentuations-/ Minderungskorrektur	- bei bekannter Reliabilität - Maximal beobachtbare Korrelation zwischen 2 Messfehlerbehafteten Variablen ist um die unkorrelierten Fehleranteile in den Variablen vermindert - einfache oder doppelte Minderungskorrektur - Nur bei Eindimensionalität der Tests sinnvoll, da sonst die Beziehungen auf Konstruktebene durch Überkorrektur überschätzt werden - Bei rein anwendungsbezogenen Fragen sollte Korrektur ausbleiben (oder nur einfache Korrektur) - Fragestellung wie hoch die Variablen auf Konstruktebene zusammenhängen erfordert doppelte Korrektur

Reliabilitäts-Validitäts-Dilemma:
Erhöhung der Trennschärfe führt bei gleichbleibender Validität der einzelnen Items zu einer Verminderung der Validität des gesamten Tests

Multi-Trait-Multi-Methode Matrix (MTMM-Ansatz) von Campbell & Fiske:

		Methode A			**Methode B**			**Methode C**		
		1	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
Methode A	**1** **2** **3**	RA1 r r	 RA2 r	 RA3						
Methode B	**1** **2** **3**	**r** r r	r **r** r	r r **r**	RB1 r r	 RB2 r	 RB3			
Methode C	**1** **2** **3**	**r** r r	r **r** r	r r **r**	**r** r r	r **r** r	r r **r**	RC1 r r	 RC2 r	 RC3

Monomethod-Blöcke (fett, rot umrandet)	Monotrait-Monomethod	Reliabilität (Hauptdiagonale)
	Heterotrait-Monomethod	Diskriminante Validität < MM & MH (Dreiecksmatrizen unter Reliabilität)
Heteromethod-Blöcke (die anderen drei)	Monotrait-Heteromethod (fett, blau)	Konvergente Validität = hoch Diskriminante V. > HM & HH
	Heterotrait-Heteromethod	Diskriminante V. Wenn < MH

Korrelative Analyse der MTMM schöpft Potential nicht aus, es bliebt viel subjektiver Interpretationsspielraum. Auswertung mittels konfirmatorische Faktorenanaylse ist angemessen.

Bivariater Maße der Kriteriumsvalidität

Effektstärke d	- bei Klassifikation von Personen in Gruppen - Gibt Mittelwertsunterschiede in den Testwerten der beiden Gruppen in Einheiten der Standardabweichung an - Abhängige Stichproben: im Nenner steht die Standardabweichung der Differenzen
Beurteilung d	- schwacher Effekt: r= .10 ; d= .20 - mittlerer Effekt: r= .30 ; d= .50 - Starker Effekt: r= .50 ; d= .80

Validität nach dem Zeitpunkt der Erhebung

Retrospektiv	Kriterium wird vor dem Test erhoben (Schulnoten vor IQ-Test)
Konkurrent	Kriterium wird gleichzeitig mit dem Test erhoben
Prädikativ/ prognostisch	Kriterium wird nach dem Test erhoben (Einfluss der Persönlichkeit auf berufliche Leistung)

Möglichkeiten um einer Überanpassung der Stichprobe und damit tendenziellen Überschätzung der Validität entgegen zu wirken:
- statistische Korrekturfaktoren
- Kreuzvalidierung

Verfahren zur multivariaten Validitätsprüfung:
- logistische Regression
- Diskriminanzanalyse

Inkrementelle Validität: Ausmaß indem der neue Test einen Beitrag zur Aufklärung des Kriteriums leistet im Vergleich mit den alten Tests.
- Delta-R: Differenz der Werte für R (R2-R1 = Delta-R)
- Ermittlung im Rahmen einer hierarchischen Regressionsanalyse/ SPSS: ANALYSIEREN --> REGRESSIONSANALYSE --> LINEAR --> UNABHÄNGIGE --> STATISTIKEN: ANPASSUNGSGÜTE DES MODELLS & ÄNDERUNGEN IN R-QUADRAT

5.3 Weitere allgemeine Gütekriterien

DIN 33430 - Anforderungen an die Testdokumentation: nachvollziehbare Dokumentation der Zielsetzung & Anwendungsbereiche/ Empirische Untersuchung/ Konstruktionsschritte/ Gütekriterien des Verfahrens.

5.3.1 Kriterien mit vorwiegend praktischem Anwendungsbezug

Testökonomie	- Beanspruchung der Ressourcen (Kompetenz zur Auswahl/ Durchführung/ Interpretation) - Zeit der Testleiter & Durchführungsdauer - Finanzielles Budget der Anwender
Nutzen der Testverfahren	- Verbesserung des Anteils richtiger Entscheidungen im Vergleich zur Zufallsauswahl - Nutzen ist abhängig von der Basisrate (Anteil geeigneter Personen aus Bewerberpopulation) & Selektionsquote (Anteil der Ausgewählten) --> Monetäre Nutzenanalyse

5.3.2 Kriterien mit Bezug zu Rechten & Reaktionen der Teilnehmer

Gesetzliche Vorschriften	- Psychotherapeutengesetz - Betriebsverfassungsgesetz - Allgemeines Gleichbehandlungsgesetz - Fahrerlaubnisverordnung - Richterrecht
Informed consent	Einverständnis der Teilnehmer auf Grundlage ausreichender Information
Testfairness	- Ausmaß, in dem bestimmte Gruppen von Testteilnehmern durch die Testergebnisse nicht systematisch benachteiligt werden. Abwesenheit von gruppenspezifischer Bias - §1 AGG: Diskriminierungsverbot
Akzeptanz	subjektive, bewertende Einstellungen & Reaktionen
Zumutbarkeit	In Relation zum Nutzen zu betrachtende Vermeidung einer Überbeanspruchung in zeitlicher/ psychischer/ körperlicher Hinsicht
Prozessuale Gerechtigkeit	Die Einhaltung bestimmter Regeln verbessert die Akzeptanz von Testverfahren auch dann, wenn das individuelle Testergebnis mit negativen Konsequenzen verbunden ist.

5.3.3 Kriterien mit Bezug zum Verhalten der Teilnehmer

Unverfälschbarkeit	Unabhängigkeit der Testergebnisse von bewussten/ unbewussten Verzerrungen --> faking good/ faking bad
Sicherstellung der Identität	Lösungsvorschläge laufen auf supervidierte Testung hinaus.

Verfahren zur Kontrolle sozialer Erwünschtheit

Subtile Items	Zielrichtung soll möglichst undurchschaubar sein
forced-choice	Es muss eine Antwortalternative ausgewählt werden
bogus pipeline	Aufdeckung von Verfälschungstendenzen/ Anschluss an simulierten Lügendetektor
bogus items	Frage nach Erfahrungen mit nicht existierenden Gegenständen

5.4 Gütekriterien für die Einzelfalldiagnostik

Vergleichbarkeit	Mit Normwerten
Messgenauigkeit	Der einzelnen Testwerte

5.4.1 Anforderungen an die Normierung

Stichprobengröße (COTAN)	- Niveau 1: N=300 - Niveau 2: N=200 - Niveau 3: N=100
Repräsentativität	- qualitatives Kriterium - Sicherstellung durch Zufallsziehung
Übertragbarkeit	- Der Normwerte auf die aktuelle Fragestellung - Einfluss der Bewerbersituation auf Mittelwerte --> bei berufsrelevanten Persönlichkeitstests liegt dieser im Bereich einer halben bis vollen Standardabweichung --> Entspricht 1-2 STANINE-Punkten --> Normen für Einzellfalldiagnostik unbrauchbar - Mögliche Entgegnung: Erstellung differenzierter Normen
Aktualität der Normwerte	Flynn-Effekt: Verschiebung der Normwert um fast eine volle Standardabweichung im Verlauf von 50 Jahren --> Überprüfung des Tests nach spätestens 8 Jahren

5.4.2 Kennwerte mit Bezug zur Messgenauigkeit

Messgenauigkeit ergibt sich im Rahmen der KTT aus der Reliabilität eines Tests. Messfehler für jedes Individuum folgt einer zufälligen Verteilung um den Wert 0, das heißt die Testwerte werden mit gleicher Wahrscheinlichkeit über-/ unterschätzt

Ermittlung der Messgenauigkeit:

1. Ermittlung des durchschnittlichen Fehlers eines individuellen Testwertes --> Standardabweichung der Fehlerverteilung
   a. Verteilung des Messfehler einer Person bei vielen Messgelegenheiten, oder Verteilung vieler Personen bei einer Messgelegenheit
   b. Messfehler zwischen den Mitgliedern sollen nicht/ nur gering differieren
   c. Annahme: Messfehler ist normalverteilt, Standardschätzfehler: Messerwerte/ FEhler sind bivariat normalverteilt, sowie Unkorreliertheit der Fehler
2. Schätzung des Standardmessfehlers
   a. Äquivalenzhypothese: der beobachtete Testwert nähert sich dem wahren Wert befriedigend an. $SE_x = S_x * \sqrt{1-r_{tt}}$
   b. Regressionshypothese: Wahrer Wert muss aus den Testwerten und Stichprobenwerten noch geschätzt werden: $t_i' = r_{tt} * X_i + M * (1-r_{tt})$
   c. t_i' liegt näher am Stichprobenmittelwert als der beobachtete Wert X_i und zwar umso mehr, je weniger reliabel der Test ist
3. Ermittlung des Standardschätzfehlers: $SE_T = S_x * \sqrt{r_{tt} * (1-r_{tt})}$
4. Berechnung von Konfidenzintervallen
   a. Äquivalenzhypothese: $KI = X_i \pm SE_x * z$
   b. Regressionshypothese: $KI = t_i' \pm SE_x * z$
5. Prüfung des Unterschieds zwischen 2 Testwerten: Berechnung der kritischen Differenz
   a. Äquivalenzhypothese/ gleicher Test: $D = z * S_x * \sqrt{2} * (1-r_{tt})$
   b. Äquivalenzhypothese/ verschiedene Tests: $D = z * S_x * \sqrt{2 - (r_{tt1} + r_{tt2})}$
   c. Regressionshyptohese/ versch. Tests: $D = z * S_x * \sqrt{(1-r^2_{tt})}$