

Inhaltsverzeichnis

1. Kapitel: Klassische Testtheorie. Grundlagen und Erweiterungen für heterogene Tests und Mehrfacettenmodelle Von Wolfgang A. Rauch und Helfried Moosbrugger

1	Konzeptuelle Grundlagen und Definitionen	1
1.1	Einführung	1
1.2	Testtheorien und Voraussetzungen für deren Anwendung	2
1.3	Gütekriterien als Anforderungen an die Qualität psychologischer Diagnostik	4
1.3.1	Reliabilität als normative Anforderung	5
1.3.2	Validität als normative Anforderung	7
1.3.3	Testwert, Subtests, Testlets und Testitems	10
2	True Scores und Fehlerwerte	11
2.1	Definition der KTT	11
2.2	Einführung in das Problem der Messwiederholung	12
2.2.1	Das Uhrenbeispiel	12
2.2.2	Zur Wahrheit von „wahren“ Werten	13
2.3	Modellvorstellungen zur Entstehung der Messwertvariabilität	13
2.3.1	Zufallsziehung von Individuen	13
2.3.2	Modell der intraindividuellen Verteilung und des „stochastischen Individuums“	14
2.3.3	Zwei Dimensionen der Testwertvariabilität	15
2.4	Konstruktion des „klassischen“ Messfehlermodells	16
2.5	Experimentelle und lineare Unabhängigkeit von Testwerten	18
3	„Klassische“ Ansätze zur Schätzung von Reliabilitätskoeffizienten	20
3.1	Ausgangspunkt: Parallele Tests	20
3.2	Paralleltest-Reliabilitätsschätzung	21
3.3	Retest-Reliabilitätsschätzung	24
3.4	Reliabilitätsschätzung über Testzerlegung	24
3.4.1	Stufenweise Abschwächungen der Parallelitätsannahme	25
3.4.1.1	Parallelität	25
3.4.1.2	Tau-Äquivalenz und essenzielle Tau-Äquivalenz	25
3.4.1.3	Kongenerische Testwertvariablen	25
3.4.1.4	Nominelle Parallelität	26
3.4.2	Reliabilitätsschätzung über Testzerlegung in zwei oder drei Teile	27
3.4.2.1	Die Spearman-Brown-Formel für zwei Testhälften	27

3.4.2.2	Split-Half-Reliabilität unter (essenzieller) Tau-Äquivalenz	27
3.4.2.3	Kristofs Reliabilitätskoeffizient für drei Testteile	28
3.4.3	Reliabilitätsschätzung über Testzerlegungen in beliebig viele Teile	28
3.4.3.1	Generalisierter Spearman-Brown-Koeffizient	28
3.4.3.2	Cronbachs Alpha	29
3.4.3.3	Reliabilität bei kongenerischen Tests	29
3.4.3.4	Bias von Cronbachs Alpha bei kongenerischen Tests	31
3.4.3.5	Guttmans Koeffizient	32
3.4.3.6	McDonalds Koeffizient Ω	32
3.4.4	Schlussfolgerungen für die Reliabilitätsschätzung über beliebige Testzerlegungen	34
4	Erweiterung der KTT: Heterogene Tests	35
4.1	Definition von Heterogenität	35
4.2	Ursachen von Heterogenität	35
4.3	Ansätze zur Modellierung von heterogenen Tests	36
4.3.1	Zur Eindeutigkeitsproblematik bei heterogenen Testmodellen	36
4.3.1.1	Zur Uneindeutigkeit unterschiedlicher mehrfaktorieller Modelle	36
4.3.1.2	Zur Uneindeutigkeit von mehrfaktoriellen Modellen und Modellen mit korrelierten Fehlern	37
4.3.2	Mehrdimensionalität oder korrelierte Fehler?	38
5	Zur Rolle der konfirmatorischen Faktorenanalyse in der erweiterten KTT	39
5.1	Modelle mit latenten Variablen	39
5.2	Konditionale Unabhängigkeit	41
5.3	Latent-Trait-Modelle	42
5.4	Zur Unterscheidung von starken True-Score-Modellen und Latent-Trait-Modellen	42
5.5	Das konfirmatorische Faktorenmodell als Latent-Trait-Modell	43
5.5.1	Modellannahmen	43
5.5.2	Maximum-Likelihood-Schätzung und Modellfit	44
5.5.3	Relaxierte Annahmen bei konfirmatorischen Faktorenanalysen	44
5.5.4	Pragmatische Aspekte der konfirmatorischen Faktorenanalyse im Rahmen der erweiterten KTT	44
6	Reliabilitätsschätzung bei heterogenen Tests mit konfirmatorischen Faktorenanalysen	46
6.1	Reliabilitätsschätzung bei einfaktoriellen Modellen mit korrelierten Fehlern	47
6.1.1	„Klassische“ Reliabilität bei korrelierten Fehlern	47
6.1.2	Cronbachs Alpha bei korrelierten Fehlern	48
6.1.3	Koeffizient ω und korrelierte Fehler	48
6.2	Reliabilitätsschätzung mit mehrfaktoriellen Modellen	49
6.2.1	Das hierarchische Faktormodell	50
6.2.2	McDonalds ω und Koeffizient ω_H	52
6.2.3	Zum Einfluss der Gruppenfaktoren	52

6.3 Welcher Koeffizient sollte bei heterogenen Tests zur Reliabilitäts-schätzung gewählt werden?	55
7 Reliabilitätsschätzung bei heterogenen Tests mit explorativen Faktorenanalysen	57
7.1 Einführung	57
7.2 Die „größte untere Schranke“ der Reliabilität	58
7.3 Bentlers (2004) maximale eindimensionale Reliabilität	59
8 Stichprobentheoretische Erwägungen	60
8.1 Zur „Populationsabhängigkeit“ der KTT	60
8.2 Zu Schätzungen von Reliabilitätskoeffizienten aus Stichprobendaten ..	61
9 Mehrfacettenmodelle	62
9.1 Überblick	62
9.2 Multi-Trait-Multi-Method-Analyse	63
9.2.1 Konvergente und diskriminante Validität	63
9.2.2 Messmethoden, Methodeneffekte und Validität	64
9.2.3 Faktorenanalytische MTMM-Ansätze	66
9.3 Latent-State-Trait-Theorie	68
9.3.1 Konsistenz und (situationale) Spezifität	68
9.3.2 LST-Modell und LST-Koeffizienten	69
9.3.3 Zur Unterscheidung von Situationen und Messgelegenheiten ..	70
9.3.3.1 Bekannte oder unbekannte Situationen?	71
9.3.3.2 Mehrebenenmodelle für geschachtelte Situations-effekte	73
9.4 Generalisierbarkeitstheorie	74
9.4.1 Grundbegriffe der G-Theorie	74
9.4.2 Varianzkomponenten	75
9.4.3 Generalisierbarkeitskoeffizienten	77
9.4.4 Geschachtelte Messprozeduren	78
9.4.5 Entscheidungsstudien	79
9.4.6 Zusammenfassende Betrachtungen zur Generalisierbarkeits-theorie	79
10 Schlussfolgerungen für die Gesellschaft	80
10.1 Weiterentwicklungen der KTT	80
10.2 Konsequenzen für Lehre und Forschung	81
Literatur	82

2. Kapitel: Psychometrische Grundlagen von Large Scale Assessments

Von Oliver Walter und Jürgen Rost

1 Testkonstruktion	88
1.1 Entwicklung der Rahmenkonzeptionen	89
1.2 Entwicklung der Testaufgaben	90

1.3	Revision der Testaufgaben, Präpilotstudien und Reviewprozesse	92
1.4	Übersetzungen der Testaufgaben	93
1.5	Signierung und Kodierung von Aufgaben mit offenem Antwortformat	94
1.6	Durchführung einer Voruntersuchung	95
2	Testdesign	96
3	Stichprobendesign	100
3.1	Definition der Zielpopulation	100
3.2	Die Entwicklung des Stichprobendesigns	102
3.3	Stichprobenfehler und Stichprobengröße	104
3.4	Ziehung der Stichprobe und Berechnung von Stichprobengewichten . .	107
4	Skalierung der Leistungsdaten	108
4.1	Anforderungen an die Skalierung	108
4.2	Für Large Scale Assessments geeignete Testmodelle	109
4.2.1	Logistische Testmodelle als Basismodelle in Large Scale Assessments	110
4.2.2	Verallgemeinerungen des Rasch-Modells	114
4.3	Parameterschätzung in Large Scale Assessments	118
4.3.1	Grundprinzip der Maximum-Likelihood-Methode	118
4.3.2	Itemparameterschätzung	119
4.3.3	Hintergrundmodell und latente Regression	125
4.3.4	Plausible Values und Personenparameterschätzer	130
4.3.5	Linking von Skalen verschiedener Erhebungen	133
5	Reliabilität und Modellgeltung	134
5.1	Die Schätzung der Reliabilität	134
5.2	Aspekte der Modellgeltung	138
5.2.1	Globale Modellgeltungstests	139
5.2.2	Spezielle Modellgeltungstests	140
5.2.3	Itemfitmaße	141
6	Schlussfolgerungen	144
	Literatur	145

3. Kapitel: Methoden der Item- und Skalenkonstruktion Von Safir Yousfi

1	Messtheoretische Grundlagen der Testkonstruktion	151
1.1	Grundriss der psychometrischen Testtheorie	153
1.1.1	Klassische und Probabilistische Messmodelle	155
1.1.2	Messmodelle der psychometrischen Testtheorie	156
1.2	Validität	159
2	Strategien der Testkonstruktion	164
2.1	Deduktive Methode	165
2.2	Induktive Methode	167

2.3	Externale Methode	170
2.4	Vergleich der Testkonstruktionsstrategien	171
2.5	Unterscheidungsmerkmale von psychologischen Testverfahren	172
3	Generierung von Items	175
3.1	Verhaltensstichproben, Simulationen und situative Fragen	175
3.2	Prototypenansatz	177
3.3	Lexikalischer Ansatz	178
3.4	Facettentheoretische Ansätze	179
3.5	Rationale Itemkonstruktion	180
3.6	Empfehlungen für die Itemkonstruktion bei Selbstberichtsdaten	182
4	Aggregation	185
4.1	Messung anhand von einzelnen Items	186
4.2	Aggregation durch Addition oder Mittelwertberechnung	187
4.3	Aggregation durch Linearkombination	188
4.4	Weitere statistische Methoden	189
5	Selektion von Items	190
5.1	Wissenschaftliche Aspekte der Itemselektion	190
5.2	Auswahl nach Itemkennwerten	192
5.2.1	Externe Validität	192
5.2.2	Faktorielle Validität	193
5.2.3	Interne Validität (Itemfit)	194
5.2.4	Klassische Trennschärfe- und Itemschwierigkeitskoeffizienten ..	195
5.2.5	Probabilistische Itemparameter	197
5.2.5.1	Trennschärfekonzepte der probabilistischen Testtheorie	197
5.2.5.2	Trennschärfe und Gütekriterien in der probabilis-	
	tischen Testtheorie	198
5.2.5.3	Adaptives Testen	200
5.2.6	Inhaltliche Kriterien	201
5.3	Auswahl nach Skalenskennwerten	202
5.3.1	Algorithmen	203
5.3.2	Zielvariablen	204
5.3.3	Empirie	206
5.4	Optimal Test Design	206
6	Fazit	208
	Literatur	209

4. Kapitel: Automatisierte Itemgenerierung: Aktuelle Ansätze, Anwendungen und Forschungen
Von Martin Arendasy und Markus Sommer

1	Einleitung	215
1.1	Testsicherheit	215
1.2	Bedeutung unterschiedlicher Aspekte der Validität	217

2	Automatisierte Itemgenerierung	219
2.1	Klassifikation der Ansätze zur Automatisierten Itemgenerierung	219
2.1.1	Grad der inhaltlich-theoretischen Fundierung	219
2.1.2	Grad der freien Variierbarkeit der einzelnen Bauelemente	221
2.1.3	Einbeziehung einer Qualitätssicherungskomponente	223
2.1.4	Grad der Automatisierung	225
2.2	Einordnung der unterschiedlichen Ansätze der AIG in das Klassifikationsschema	226
3	Konstruktionsphasen eines Zwei-Komponenten-Itemgenerators	232
3.1	Beschreibung des zu messenden latenten Traits	232
3.2	Ableiten der lösungsrelevanten kognitiven Prozesse und Wissensstrukturen	232
3.3	Ableitung von Radicals	234
3.4	Ableitung der funktionalen Einschränkungen	234
3.5	Formulierung und empirische Untersuchung der zu überprüfenden Meilensteine	235
4	Aktuelle Anwendung eines Zwei-Komponenten-Itemgenerators	236
4.1	Definition des zu messenden latenten Traits	236
4.2	Ableiten der lösungsrelevanten kognitiven Prozesse und Wissensstrukturen	237
4.3	Beschreibung des Aufgabenmaterials	240
4.4	Ableitung der funktionalen Einschränkungen und der Radicals	241
4.5	Beschreibung des Itemgenerators	245
4.6	Formulierung der zu überprüfenden Meilensteine und aktuelle empirische Befunde	246
4.6.1	Überprüfung der Notwendigkeit der funktionalen Einschränkungen	246
4.6.2	Überprüfung der Konstruktrepräsentation der Endlosschleifen	249
4.6.3	Weiterführende Ergebnisse zur Konstruktrepräsentation	252
4.6.4	Überprüfung der nomothetischen Spanne der Endlosschleifen	254
5	Weiterführende Forschungsfragestellungen	258
5.1	Nutzen in der Phase der Itemkonstruktion	258
5.2	Nutzen bei der Kalibrierung automatisch generierter Items	259
5.2.1	Möglichkeiten der Reduktion von Kalibrierungskosten im schemabasierten Ansatz	259
5.2.2	Möglichkeiten der Reduktion von Kalibrierungskosten im elementbasierten Ansatz	263
6	Diskussion und Ausblick	267
	Literatur	270

5. Kapitel: Kriteriumsorientierte Diagnostik

Von Philipp Yorck Herzberg und Andreas Frey

1	Einleitung	281
2	Vergleich von kriteriumsorientierten und normorientierten Tests	282
3	Konstruktion kriteriumsorientierter Tests	285
4	Modelle kriteriumsorientierter Tests	286
	4.1 Klassische Testtheorie	287
	4.2 Binomialmodell	287
	4.3 Item-Response-Modelle	289
	4.3.1 Annahmen	289
	4.3.2 Modellansatz	290
	4.3.3 Parameterschätzung	291
	4.3.4 Schluss auf das Kriterium	291
	4.3.5 Auswahl optimaler Items	296
	4.3.6 Fazit	297
5	Setzen von Standards	297
	5.1 Setzen von Standards durch Expertenbeurteilung	298
	5.1.1 Beurteilung von Testitems	298
	5.1.2 Beurteilung von Testpersonen	301
	5.1.3 Expertenurteilsmethoden im Vergleich	301
	5.2 Setzen von Standards als Angabe von Normen und Quoten	302
	5.3 Setzen von Standards in Relation zu einem Außenkriterium	303
	5.4 Beurteilung klassischer Methoden der Standardsetzung	303
	5.5 Neue Ansätze zum Setzen von Standards	304
	5.5.1 Ein Rahmenmodell der Standardsetzung: Generic Eclectic Method	304
	5.5.2 Einbezug statistischer Methoden bei der Standardsetzung	305
6	Validität kriteriumsorientierter Tests	306
	6.1 Kontentvalidität	307
	6.2 Konstruktvalidität	309
	6.3 Kriteriumsvalidität	311
7	Reliabilität	312
	7.1 Reliabilität von Klassifikationsentscheidungen	313
	7.1.1 Schwellenverlust-Indizes	313
	7.1.2 Quadrierte Fehlerfunktionen	315
	7.1.3 Domänenwerte	316
	7.1.4 Fazit und praktische Hinweise	316
8	Fazit	317
	Literatur	318

6. Kapitel: Verhaltensbeobachtung

Von Frank M. Spinath und Nicolas Becker

1	Definition und Einteilung	326
1.1	Definition der wissenschaftlichen Verhaltensbeobachtung	326
1.2	Einteilungsgesichtspunkte	327
1.2.1	Unterteilung nach Systematik der Verhaltensbeobachtung	328
1.2.2	Unterteilung nach Segmentierung des Verhaltensstroms	328
1.2.3	Unterteilung nach Art der Abbildung	330
1.2.4	Unterteilung nach Art der Verhaltensstichprobe	335
1.2.5	Rahmen- und Durchführungsbedingungen	337
2	Beobachtungsfehler, Beobachtertraining und Gütekriterien	339
2.1	Fehlerquellen bei der Verhaltensbeobachtung	339
2.1.1	Fehler zu Lasten des Beobachtungsumfeldes	341
2.1.2	Fehler zu Lasten des Beobachtungssystems	341
2.1.3	Fehler zu Lasten des Beobachters	341
2.1.3.1	Wahrnehmungsfehler	341
2.1.3.2	Interpretationsfehler	343
2.1.3.3	Erinnerungsfehler	344
2.1.3.4	Wiedergabefehler	345
2.2	Beobachtertraining	345
2.2.1	Beobachterfehlertraining	345
2.2.2	Beobachtungsdimensionstraining	346
2.2.3	Bezugsrahmenstraining	346
2.2.4	Verhaltensbeobachtungstraining	346
2.2.5	Gegenüberstellung der Effektivität der Trainingsansätze	347
2.3	Gütekriterien	347
2.3.1	Objektivität und Reliabilität	347
2.3.1.1	Verfahren für Nominaldaten	349
2.3.1.2	Verfahren für Intervalldaten	351
2.3.2	Validität	353
2.3.2.1	Inhaltsvalidität	353
2.3.2.2	Kriteriumsvalidität	354
2.3.2.3	Konstruktvalidität	354
3	Differenzielle Perspektive	356
3.1	Nutzen von Verhaltensbeobachtungen in der differentiellen Psychologie	356
3.2	Exkurs: Ökologische Validität des Dispositionsbegriffes	356
3.3	Dispositiondiagnostik durch Verhaltensbeobachtungen	359
3.3.1	Dispositiondiagnostik durch Beobachtungsaggregation	359
3.3.2	Dispositiondiagnostik durch Kontrolle von Störeinflüssen	362
4	Schlussfolgerungen für die Gesellschaft	365
	Literatur	367

7. Kapitel: Interview

Von Karl Westhoff und Anja Strobel

1	Begriffsbestimmung	371
2	Einflüsse auf die Validität von Interviews	373
2.1	Strukturierungsmerkmale im Interviewprozess	374
2.1.1	Planung	374
2.1.2	Durchführung	375
2.1.3	Auswertung	376
2.2	Gestaltungsmöglichkeiten im Interviewprozess	377
2.2.1	Planung	377
2.2.2	Durchführung	378
2.2.3	Auswertung	380
2.3	Strukturierte Interviews	381
2.3.1	Klinische Psychologie	381
2.3.1.1	Das Diagnostische Interview bei psychischen Störungen (DIPS)	381
2.3.1.2	WHO – Composite International Diagnostic Interview (CIDI; WHO, 1990)	383
2.3.2	Arbeits- und Organisationspsychologie	385
2.3.2.1	Das Behavior Description Interview	385
2.3.2.2	Das Situational Interview	386
2.3.2.3	Das Multimodale Interview	386
2.3.3	Die Entscheidungsorientierte Gesprächsführung	388
2.3.4	Polizeipsychologie: Das Kognitive Interview	388
3	Die Güte des Interviews	389
3.1	Methodologische Aspekte der Evaluation	389
3.2	Psychometrische Bewertungskriterien	391
3.2.1	Objektivität	391
3.2.2	Retest-Reliabilität und Interne Konsistenz	392
3.2.3	Validität	392
3.2.3.1	Zur prädiktiven Validität	393
3.2.3.2	Zur inkrementellen Validität	395
3.2.3.3	Zur Konstruktvalidität	395
3.3	Nichtpsychometrische Bewertungskriterien – Ökonomie, Nutzen, Fairness, Akzeptanz	397
4	Der Interviewer als Schlüsselfigur im Interviewprozess	398
4.1	Interviewereinflüsse und Training des Interviewers	398
4.2	Feedback als Mittel zur Qualitätssicherung im Interviewprozess	399
5	Rechtliche und ethische Rahmenbedingungen	402
5.1	Qualitätsstandards und Grundregeln	402
5.2	Rechtliche Grundlagen	402
6	Schlussfolgerungen für die Gesellschaft	403
	Literatur	404

8. Kapitel: Psychodiagnostische Verfahren im Kulturvergleich Von Beatrice Rammstedt, Janet Harkness und Peter Ph. Mohler

1	Einleitung	415
2	Die Vergleichbarkeit psychischer Gegebenheiten zwischen Kulturen	416
2.1	Konzeptuelle Äquivalenz	417
2.2	Phänomenologische Äquivalenz und Indikatorenäquivalenz	418
2.3	Erhebungsäquivalenz	419
2.4	Itemäquivalenz und Skalenäquivalenz	419
3	Bias und Strategien zu dessen Vermeidung	421
3.1	Konstruktbias	421
3.2	Methodenbias	422
3.3	Itembias	423
4	Methodik der kulturvergleichenden psychologischen Diagnostik	426
4.1	Perspektiven	426
4.2	Methodische Vorgehensweise	427
4.2.1	Der psychometrische Ansatz	427
4.2.2	Der quasi-experimentelle Ansatz	430
5	Entwicklung psychodiagnostischer Verfahren für kulturvergleichende Untersuchungen	432
5.1	Die Entwicklung kulturvergleichend einsetzbarer Inventare	433
5.2	Übersetzung und Kulturelle Adaptation von Inventaren	434
5.2.1	Adaptation	434
5.2.2	Übersetzung	439
5.3	Überprüfung der Angemessenheit einer Adaptation	442
5.3.1	Qualitative Strategien zur Überprüfung der Angemessenheit einer Adaptation	442
5.3.2	Quantitative Strategien zur Überprüfung der Angemessenheit einer Adaptation	444
5.4	Richtlinien für die Adaptation von Testverfahren	446
5.4.1	Richtlinien zum Kontext	446
5.4.2	Richtlinien zur Testentwicklung und Adaptation	447
5.4.3	Richtlinien zur Testanwendung	449
5.4.4	Richtlinien zur Dokumentation/Interpretation	450
6	Schlussfolgerungen	452
	Literatur	453
	Autorenregister	459
	Sachregister	471