

Enzyklopädie der Psychologie

Leistungs-, Intelligenz-
und Verhaltensdiagnostik

Psychologische Diagnostik

3



Hogrefe · Verlag für Psychologie
Göttingen · Bern · Toronto · Seattle

Enzyklopädie der Psychologie

ENZYKLOPÄDIE DER PSYCHOLOGIE

In Verbindung mit der
Deutschen Gesellschaft für Psychologie

herausgegeben von

Prof. Dr. Niels Birbaumer, Tübingen
Prof. Dr. Dieter Frey, München
Prof. Dr. Julius Kuhl, Osnabrück
Prof. Dr. Wolfgang Schneider, Würzburg
Prof. Dr. Ralf Schwarzer, Berlin

Themenbereich B
Methodologie und Methoden

Serie II
Psychologische Diagnostik

Band 3
Leistungs-, Intelligenz-
und Verhaltensdiagnostik



Hogrefe • Verlag für Psychologie
Göttingen • Bern • Toronto • Seattle

Leistungs-, Intelligenz- und Verhaltensdiagnostik

herausgegeben von

Prof. em. Dr. Lutz F. Hornke, Aachen
Prof. em. Dr. Manfred Amelang, Heidelberg
Prof. Dr. Martin Kersting, Münster



Hogrefe • Verlag für Psychologie
Göttingen • Bern • Toronto • Seattle

© 2011 Hogrefe Verlag GmbH & Co. KG
Göttingen • Bern • Wien • Paris • Oxford • Prag • Toronto
Cambridge, MA • Amsterdam • Kopenhagen • Stockholm
Rohnsweg 25, 37085 Göttingen

<http://www.hogrefe.de>

Aktuelle Informationen • Weitere Titel zum Thema • Ergänzende Materialien

Copyright-Hinweis:

Das E-Book einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung außerhalb der engen Grenzen des Urheberrechtsgesetzes ist ohne Zustimmung des Verlags unzulässig und strafbar.

Der Nutzer verpflichtet sich, die Urheberrechte anzuerkennen und einzuhalten.

Satz: ARThür Grafik-Design & Kunst, Weimar
Format: PDF

ISBN 978-3-8409-1525-3

Nutzungsbedingungen:

Der Erwerber erhält ein einfaches und nicht übertragbares Nutzungsrecht, das ihn zum privaten Gebrauch des E-Books und all der dazugehörigen Dateien berechtigt.

Der Inhalt dieses E-Books darf von dem Kunden vorbehaltlich abweichender zwingender gesetzlicher Regeln weder inhaltlich noch redaktionell verändert werden. Insbesondere darf er Urheberrechtsvermerke, Markenzeichen, digitale Wasserzeichen und andere Rechtsvorbehalte im abgerufenen Inhalt nicht entfernen.

Der Nutzer ist nicht berechtigt, das E-Book – auch nicht auszugsweise – anderen Personen zugänglich zu machen, insbesondere es weiterzuleiten, zu verleihen oder zu vermieten.

Das entgeltliche oder unentgeltliche Einstellen des E-Books ins Internet oder in andere Netzwerke, der Weiterverkauf und/oder jede Art der Nutzung zu kommerziellen Zwecken sind nicht zulässig.

Das Anfertigen von Vervielfältigungen, das Ausdrucken oder Speichern auf anderen Wiedergabegeräten ist nur für den persönlichen Gebrauch gestattet. Dritten darf dadurch kein Zugang ermöglicht werden.

Die Übernahme des gesamten E-Books in eine eigene Print- und/oder Online-Publikation ist nicht gestattet. Die Inhalte des E-Books dürfen nur zu privaten Zwecken und nur auszugsweise kopiert werden.

Diese Bestimmungen gelten gegebenenfalls auch für zum E-Book gehörende Audiodateien.

Autorenverzeichnis

Prof. em. Dr. Manfred Amelang

Universität Heidelberg
Psychologisches Institut
Hauptstraße 47
69117 Heidelberg
E-Mail: manfred.amelang@
psychologie.uni-heidelberg.de

Ass.-Prof. Dr. Pia Deimann

Institut für Entwicklungspsychologie
und Psychologische Diagnostik
Liebiggasse 5
1010 Wien
Österreich
E-Mail: pia.deimann@univie.ac.at

Prof. Dr. Bernad Batinic

Johannes Kepler Universität Linz
Altenbergerstraße 69
4040 Linz
Österreich
E-Mail: bernad.batinic@jku.at

Prof. Dr. Joachim Funke

Universität Heidelberg
Psychologisches Institut
Hauptstraße 47
69117 Heidelberg
E-Mail: joachim.funke@psychologie.
uni-heidelberg.de

Prof. Dr. André Beauducel

Universität Hamburg
Fachbereich Psychologie
Differentielle Psychologie und
Psychologische Diagnostik
Von-Melle-Park 5
20146 Hamburg
E-Mail:
andre.beauducel@uni-hamburg.de

Dr. Timo Gnambs

Johannes Kepler Universität Linz
Altenbergerstraße 69
4040 Linz
Österreich
E-Mail: timo.gnambs@jku.at

Prof. Dr. Klaus Bös

Institut für Sport und
Sportwissenschaft
Karlsruher Institut für Technologie
(KIT)
Kaiserstraße 12
76131 Karlsruhe
E-Mail: boes@kit.edu

PD Dr. Carmen Hagemeister

TU Dresden
Fachrichtung Psychologie
Professur Diagnostik und
Intervention
01062 Dresden
E-Mail: Carmen.Hagemeister@
tu-dresden.de

Prof. Dr. Guido Hertel

Westfälische Wilhelms-Universität
Münster
Institut für Psychologie
Fliednerstraße 21
48149 Münster
E-Mail: ghertel@uni-muenster.de

Prof. em. Dr. Lutz F. Hornke

Institut für Psychologie
RWTH Aachen University
Jägerstraße 17/19
52056 Aachen
E-Mail: lutz.hornke@psych.
rwth-aachen.de

Ass.-Prof. Dr.

Ursula Kastner-Koller

Institut für Entwicklungspsychologie
und Psychologische Diagnostik
Liebiggasse 5
1010 Wien
Österreich
E-Mail:
ursula.kastner-koller@univie.ac.at

Prof. Dr. Martin Kersting

Fachhochschule des Bundes
für öffentliche Verwaltung
Fachbereich Finanzen
Gescher Weg 100
48161 Münster
E-Mail: Martin@Kersting-internet.de

Prof. Dr. Klaus D. Kubinger

Institut für Entwicklungspsychologie
und Psychologische Diagnostik
Liebiggasse 5
1010 Wien
Österreich
E-Mail: klaus.kubinger@univie.ac.at

Dr. Jonas W. B. Lang

Maastricht University
Department of Work & Social
Psychology
Postbus 616
6200 MD Maastricht
Niederlande
E-Mail:
jonas.lang@maastrichtuniversity.nl

Prof. Dr. Thomas Rammsayer

Universität Bern
Institut für Psychologie
Abteilung Persönlichkeits-,
Differentielle Psychologie und
Diagnostik
Muesmattstraße 45
3012 Bern
Schweiz
E-Mail:
thomas.rammsayer@psy.unibe.ch

Prof. Dr. Bernd Reuschenbach

Katholische Stiftungsfachschule
München
Preysingstraße 83
81667 München
E-Mail:
bernd.reuschenbach@ksfh.de

Prof. Dr. Birgit Spinath

Psychologisches Institut –
Pädagogische Psychologie
Universität Heidelberg
Hauptstraße 47–51
69117 Heidelberg
E-Mail: birgit.spinath@psychologie.
uni-heidelberg.de

Prof. Dr. Jutta Stahl

Universität zu Köln
Professur für Differentielle Psychologie und Psychologische Diagnostik
Höniger Weg 115
50969 Köln
E-Mail: jutta.stahl@uni-koeln.de

Dr. Otto B. Walter

Institut für Psychologie
RWTH Aachen University
Jägerstraße 17/19
52056 Aachen
E-Mail: walter@psych.rwth-aachen.de

Prof. Dr. Heinz-Martin Süß

Otto-von-Guericke-Universität
Magdeburg
Institut für Psychologie I
Abteilung für
Psychologische Methodenlehre,
Psychodiagnostik und Evaluationsforschung
Universitätsplatz 2, Gebäude 24
39106 Magdeburg
E-Mail: heinz-martin.suess@ovgu.de

Prof. Dr. Karl Westhoff

TU Dresden
Fachrichtung Psychologie
Professur Diagnostik und
Intervention
01062 Dresden
E-Mail: Karl.Westhoff@tu-dresden.de

PD Dr. Susanne Tittlbach

Universität Bayreuth
Institut für Sportwissenschaft
95440 Bayreuth
E-Mail: Susanne.Tittlbach@uni-bayreuth.de

Vorwort

Eine Enzyklopädie ist zunächst sehr vereinfacht viel Arbeit, die auf viele Köpfe und fleißige Hände verteilt ist. Dabei handelt es sich um eine besondere Textsorte, die mit einer ganz besonderen fachlichen Absicht erstellt wurde.

Seit dem 5. Jh. v. Chr. versteht man unter „Enzyklopädie“ die universelle Bildung oder doch zumindest die Zusammenstellung derselben (siehe dazu den Eintrag in der wohl größten Enzyklopädie der Gegenwart, der Wikipedia, die zudem noch internetbasiert und mit vielen Querverweisen organisiert und allseits zugänglich ist). Enzyklopädisten meinten damit vom Wortstamm her zunächst einen geschlossenen Kreis an Wissen, und später stand die Idee des kreisförmig immer wiederkehrenden und zu erweiternden bzw. erneuernden Wissens im Vordergrund.

Sicher wird niemand von den vorliegenden Enzyklopädiebänden eine vollständige Darstellung von Fragen, Methodiken und Antworten der Psychologischen Diagnostik erwarten, wohl aber Hinweise auf neue Trends und Ansätze. Die Herausgeber haben versucht, dies zu ermöglichen und mit weitgefassten Vorgaben an die Autoren sowie entsprechenden Begutachtungen und Hinweisen die Grundfragen und -antworten der Psychologischen Diagnostik zu Beginn des 21. Jahrhunderts vorzustellen. Wenngleich andere, historische Enzyklopädien viel umfangreicher waren, so kann sich diese hier doch sehen lassen: 47 Artikel von 74 Autorinnen und Autoren.

Dass eine solche Arbeit mehrere Jahre in Anspruch nahm, verwundert nicht. Autoren und Herausgeber haben mit großem Elan begonnen, aber manche mussten zwischenzeitlich ihre anfänglichen Zeitplanungen angesichts ihrer auch vielfältigen anderen Aufgaben revidieren. Doch auch die traurigen und erfreulichen privaten und beruflichen Lebensereignisse verzögerten die Arbeit: Wissenschaftler, auch und gerade Psychologen, sind Menschen; Herausgeber haben das zu akzeptieren und hoffen, all dem Beruflichen und Privaten durch Benevolenz begegnet zu sein. Die Leser werden das sicher respektieren.

Jeder Beitrag ist wie bei Zeitschriften auch entsprechend begutachtet worden. Dabei war stets eine Rückmeldung anonym, und die Herausgeber haben sich entschlossen, offen ihre Korrekturen, Kommentare und Diskussionen anzu-

bringen. Hieraus ergab sich mit einigen Autoren ein sehr fruchtbarer Dialog; die Beiträge haben dadurch fraglos gewonnen, wenngleich die Autorinnen und Autoren jeweils den eigenen Beitrag selber verantworten, denn bei ihnen liegt letztlich die Expertise, wegen derer sie zur Mitarbeit eingeladen wurden.

Im Ergebnis ist durch die Aufgeschlossenheit und die Bemühungen aller eine Enzyklopädie der Psychologischen Diagnostik entstanden, die wie ihre Vorgänger fachlich über Fragen, Methoden und Antworten informiert. Auch wenn nicht alle anfangs geplanten Beiträge geschrieben werden konnten, so liegt doch nun ein Überblick vor, der allgemein Interessierte, Studierende und Fachwissenschaftler gleichermaßen, wenngleich aus unterschiedlicher Perspektive, ansprechen kann und soll. Zumindest hat die Psychologische Diagnostik wieder eine Plattform gefunden, von der aus Gegenwarts- und Zukunftsfragen gestellt und beantwortet werden können. Die hier vorliegende Enzyklopädie zeigt die Nähe aller Autorinnen und Autoren zueinander und zum Fach. Sie ist eine Art Zugehörigkeitsausweis; denn alle, die sich hier zu Wort melden, sind die Mitgestalter und -träger zukünftiger Entwicklungen in dieser Teildisziplin der Psychologie. Dass vieles zwar auch international mitteilbar ist, aber manches doch nur im deutschsprachigen Raum rezipiert wird, das liegt an der Natur des Gebietes: Diagnostik menschlichen Erlebens und Verhaltens geht zwar in seinen allgemeinen Grundfragen über den eigenen Sprachraum hinaus, aber die gesellschaftliche Anwendung vollzieht sich nun einmal dominant innerhalb einer (Sprach-) Kultur. Beides aber bedingt sich und ist allein genommen eher problematisch.

Nach allem gebührt aus Sicht der Herausgeber allen Autorinnen und Autoren der Dank, an dieser Enzyklopädie Serie „Psychologische Diagnostik“ mitgewirkt zu haben. Die vielen Leser werden ihren Dank dadurch abstaten, dass sie die Artikel zitieren, die Ideen aufgreifen, diskutieren und weiterentwickeln. Dazu gehört auch die Frage, ob die Textsorte Enzyklopädie zukünftig nicht doch auch einer anderen Präsentationsform oder einer Ergänzung durch diese bedarf.

Die Gesellschaft aber ist eingeladen, das Nachstehende zur Kenntnis zu nehmen und dort, wo es nicht unmittelbar verständlich ist, die Autorinnen und Autoren zu fragen, anzusprechen und sich mit ihnen auseinanderzusetzen. Gerade die vielen jungen Autorinnen und Autoren garantieren, dass fachlich „nachgehalten“ werden kann. Nur so wird die eigentliche Absicht einer *Enkyklios Paideia* eingelöst: Psychologie und Psychologische Diagnostik im Besonderen ist der gesellschaftlichen Wirklichkeit und Aufklärung verpflichtet; hier entfaltet sie ihre vornehmsten Wirkungen.

Lutz F. Hornke, Aachen,
Manfred Amelang, Heidelberg und
Martin Kersting, Münster

Inhaltsverzeichnis

1. Kapitel: Testorientierte Leistungsdiagnostik: Grundlagen und Probleme, Verfahrensinventar und besondere Einsatzgebiete Von Klaus D. Kubinger

1	Einleitung	1
2	Einteilung von Leistungstests	2
3	Voraussetzungen von Tests zur Leistungsdiagnostik	3
4	Gestaltungsweisen von Tests zur Leistungsdiagnostik	5
4.1	Antwortformat	5
4.2	Zeitbegrenzung	7
4.3	Sozialform	8
4.4	Technik	8
5	Technische Möglichkeiten und Grenzen der Leistungsdiagnostik	10
5.1	Messgenauigkeit	10
5.2	Ökologische Validität	15
5.3	Stabilität des gemessenen Merkmals	17
5.4	Globalisierungsanspruch	18
5.5	Anforderungsmodell	22
6	Probleme von Tests zur Leistungsdiagnostik	24
6.1	Freies Antwortformat versus Multiple-Choice-Format	24
6.2	Power- versus Speed-and-Power-Test	27
6.3	Gruppen- versus Individualtests	28
6.4	Papier-Bleistift-Tests versus Computertests	28
7	Typische Tests zur Leistungsdiagnostik	29
7.1	Merkmalsbereich Intelligenz	29
7.2	Merkmalsbereich Sprachverständnis	31
7.3	Merkmalsbereich Räumliches Vorstellungsvermögen	32
7.4	Merkmalsbereich Gedächtnis	34
7.5	Merkmalsbereich Verarbeitungsgeschwindigkeit	35
7.6	Merkmalsbereich Logisches Denken	36
7.7	Merkmalsbereich Intellektuelle Lernfähigkeit	37
7.8	Merkmalsbereich Aufmerksamkeit und Konzentration	38
7.9	Merkmalsbereich Technisches Verständnis	39

8	Besondere Einsatzgebiete	39
8.1	Hochbegabtdiagnostik	40
8.2	Simulationsdiagnostik	42
8.3	Verkehrspsychologische Diagnostik	44
9	Schlussfolgerungen für die Gesellschaft	46
	Literatur	46

2. Kapitel: Konzentrationsdiagnostik

Von Carmen Hagemeister und Karl Westhoff

1	Theorien und Konzepte von Konzentration	51
1.1	Konzentration als Aspekt des Arbeitens	51
1.2	Konzentration als Zustand und als Persönlichkeitsmerkmal	52
1.3	Zentrale Merkmale der Konzentration	53
1.4	Das Akku-Modell der Konzentration	55
1.5	Verständnis von Konzentration im Alltag	56
1.6	Abgrenzung von Aufmerksamkeit und Konzentration	56
1.6.1	Begriffliche Abgrenzung	56
1.6.2	Empirische Abgrenzung	58
2	Konzentrationstests	59
2.1	Bedingungen für das Messen von Konzentration	59
2.1.1	Hirnorganische Gesundheit	59
2.1.2	Wahrnehmung: Tempo der Aufgabendarbietung	60
2.1.3	Wahrnehmung: Erkennbarkeit der Aufgaben	61
2.1.4	Gedächtnis	61
2.1.5	Lernfähigkeit	61
2.1.6	Strategien der Testbearbeitung	62
2.1.7	Geübtheit	62
2.1.8	Motivation: Umgang mit Über- und Unterforderung	63
2.2	Maße für Konzentration	63
2.2.1	Tempo konzentrierten Arbeitens	63
2.2.2	Fehlerneigung beim konzentrierten Arbeiten	64
2.2.3	Kombination von Tempo und Fehleranteil zu einem Maß	65
2.2.4	Verlauf einer länger dauernden Konzentrationsleistung	67
2.2.5	Unterschiede zwischen Verhalten in Konzentrationstests und alltäglichem konzentrierten Arbeiten	67
2.3	Zusammenfassende Definition von Konzentrationstest	69
2.4	Einfache und komplexe Konzentrationstests	70
2.5	Übung in Konzentrationstests	71
2.5.1	Übung und Transfer bei Konzentrationstests	71
2.5.2	Informationen in Konzentrationstests über Übungseffekte	73

2.5.3	Veränderung der Validität mit der Übung	74
2.5.4	Vermeiden von Übungseffekten	75
2.5.4.1	Tests mit geringen Übungseffekten	75
2.5.4.2	Tests mit verschiedenen großen Übungseffekten	75
2.5.4.3	Austrainieren der Getesteten	75
2.5.5	Ansätze zum Erkennen von Übung	76
2.5.5.1	Verwendung von EEG-Parametern	76
2.5.5.2	Unterschiede zwischen verschiedenen Items	77
2.6	Vortäuschen schwächerer Konzentrationsleistungen	77
2.7	Weitere Aspekte der Validität	78
2.7.1	Konzentrationsmessung bei Kindern	78
2.7.2	Schule	79
2.7.3	Berufsausbildung	80
2.7.4	Fahrzeugführung	81
2.7.5	Autofahren im Alter	84
3	Andere Methoden zur Erfassung von Konzentration	85
3.1	Unsystematische Beobachtungen von Konzentration	85
3.2	Systematische Berichte von Konzentration in Fragebögen	85
4	Schlussfolgerungen für die Praxis und die Gesellschaft	88
	Literatur	90

3. Kapitel: Intelligenztests und ihre Bezüge zu Intelligenztheorien

Von Heinz-Martin Süß und André Beauducel

1	Einleitung	97
2	Begriffsexplikation	100
3	Historischer Abriss der Intelligenzmessung	102
3.1	Die Anfänge der Intelligenzdiagnostik	102
3.2	Binet und Simons Stufentests der Intelligenz	103
3.3	Wechslers Beitrag zur Intelligenzdiagnostik	104
4	Strukturmodelle der akademischen Intelligenz	105
4.1	Zwei-Faktorentheorie	106
4.2	Primärfaktorentheorie	107
4.3	Structure-of-Intellect Model	107
4.4	Erweiterte Theorie der fluiden und kristallinen Intelligenz	109
4.5	Three-Stratum Theory	118
4.6	Cattell-Horn-Carroll-Theorie	119
4.7	Berliner Intelligenzstrukturmodell	120

4.8	Vergleich und kritische Diskussion der integrativen Modelle	122
4.8.1	Hierarchieannahme	122
4.8.2	Unimodalität versus Mehrmodalitätsannahme	124
4.8.3	Generelle Fähigkeitskonstrukte (Übersicht)	125
4.8.4	Fluide Intelligenz (logisch-schlussfolgerndes Denken)	126
4.8.5	Kristalline Intelligenz	127
4.8.6	Gedächtnis	127
4.8.7	Divergentes Denken	128
4.9	Geltungsbereiche faktorenanalytischer Strukturmodelle der Intelligenz	129
5	Beschreibung und Klassifikation ausgewählter Intelligenztests	131
5.1	Adaptives Intelligenz Diagnostikum	169
5.2	Berliner Intelligenzstrukturtest	171
5.3	Grundintelligenztest	173
5.4	Intelligenzstrukturtest	174
5.5	Kognitiver Fähigkeits-Test	176
5.6	KIT of Factor-Referenced Cognitive Tests	176
5.7	Mannheimer Intelligenztest	177
5.8	Wechsler Intelligenztest für Erwachsene	178
5.9	Wilde-Intelligenz-Test	180
5.10	Woodcock-Johnson-Tests	181
5.11	Folgerungen für die Diagnostik	182
6	Zur Kriteriumsvalidität von Intelligenztests	185
7	Neuere Entwicklungen in der Intelligenzdiagnostik	187
7.1	Testtheoretische Grundlagen	187
7.2	Regelgenerierte Itemkonstruktion	188
7.3	Computergestütztes Testen	189
7.4	Computergestütztes adaptives Testen	189
7.5	Lerntestkonzept (dynamisches Testen)	190
7.6	Zusammenfassung	191
8	Neue Intelligenzkonstrukte und ihre Messkonzepte	192
8.1	Operative Intelligenz	194
8.2	Praktische Intelligenz (traditioneller Begriff)	196
8.3	Praktische Intelligenz (nach Sternberg und Wagner)	196
8.4	Soziale Intelligenz	198
8.5	Emotionale Intelligenz	202
8.6	Gardners multiple Intelligenzen	205
8.7	Zusammenfassende Kritik	206
9	Schlussfolgerungen	207
	Literatur	208

4. Kapitel: Wissensdiagnostik: Allgemeine und spezielle Wissenstests

Von André Beauducel und Heinz-Martin Süß

1	Einleitung	235
2	Validitäten im Kontext der Wissensdiagnostik	240
	2.1 Inhaltsvalidität	241
	2.2 Konstruktvalidität	244
	2.3 Kriteriumsvalidität	250
3	Ein Orientierungsrahmen und exemplarische Einordnung von Wissenstests	253
4	Konsequenzen für die zukünftige Wissensdiagnostik	263
5	Gesellschaftliche Relevanz	265
	Literatur	267

5. Kapitel: Entwicklungstests

Von Ursula Kastner-Koller und Pia Deimann

1	Theoretische und methodische Grundlagen von Entwicklungstests	275
	1.1 Die Bedeutung neuerer Entwicklungstheorien für Entwicklungstests	275
	1.2 Methodische Anforderungen an Entwicklungstests	276
	1.3 Aufgaben und Ziele von Entwicklungstests	278
2	Entwicklungstests bei Säuglingen, Kindern und Jugendlichen	279
	2.1 Stellenwert der Entwicklungsdiagnostik in dieser Altersgruppe	279
	2.2 Entwicklungstests für Säuglinge und Kleinstkinder	280
	2.3 Entwicklungstests für Klein- und Vorschulkinder	283
	2.3.1 Allgemeine Entwicklungstests	283
	2.3.2 Spezielle Entwicklungstests	285
	2.3.2.1 Entwicklungstests zur Überprüfung der motorischen Entwicklung	285
	2.3.2.2 Entwicklungstests zur Überprüfung der Wahrnehmungsentwicklung	286
	2.3.2.3 Entwicklungstests zur Überprüfung der kognitiven Entwicklung	287
	2.3.2.4 Entwicklungstests zur Überprüfung der sprachlichen Entwicklung	290
	2.4 Entwicklungstests für Schulkinder und Jugendliche	292
	2.5 Screeningverfahren	293

3	Entwicklungstests bei Erwachsenen	298
3.1	Stellenwert der Entwicklungsdiagnostik in dieser Altersgruppe	298
3.2	Entwicklungstests zur Diagnose von Abbauprozessen im Alter	299
4	Schlussfolgerungen für die Gesellschaft	300
	Literatur	300

6. Kapitel: Schultests

Von Birgit Spinath

1	Diagnostik zu Schulbeginn	307
1.1	Einschulung, Schulreife und Schulfähigkeit	307
1.1.1	Traditionelle Schuleingangsdiagnostik und ihre Probleme	308
1.1.2	Moderne Schuleingangsdiagnostik	310
1.2	Sprachstandsdiagnostik	311
2	Feststellung spezifischer Förderbedarfe	313
2.1	Feststellung von Teilleistungsstörungen	314
2.2	Feststellung sonderpädagogischen Förderbedarfs	318
3	Übergangsempfehlungen	320
4	Schulleistungstests	322
5	Schlussfolgerungen für die Gesellschaft	326
	Literatur	329

7. Kapitel: Funktionendiagnostik

Von Susanne Tittlbach und Klaus Bös

1	Einführung	333
2	Sensorische Sinnessysteme	335
2.1	Visuelles Sinnessystem	336
2.2	Vestibuläres Sinnessystem	337
2.3	Kinästhetisches Sinnessystem	338
2.4	Akustisches Sinnessystem	339
3	Motorik	340
3.1	Motorische Grundeigenschaften	341
3.2	Motorische Funktionen	343
3.2.1	Anatomisch-physiologische Funktionen	345
3.2.2	Komplexe motorische Funktionen	345

4	Diagnose motorischer Grundeigenschaften und motorischer Funktionen	347
4.1	Klassifizierung von Testverfahren	347
4.2	Motorische Funktionstests	349
4.2.1	Manuelle und klinische sportmedizinische Diagnostik des Halte- und Bewegungssystems	349
4.2.2	Muskelfunktionstests	350
4.2.3	Erfassung sensomotorischer Leistungen	351
4.2.4	Funktionstests zur Erfassung motorischer Grundeigenschaften	352
4.2.5	Tests zur Erfassung des Gleichgewichts	352
4.2.6	Tests zur Erfassung situationsbezogener motorischer Funktionen	353
4.3	Selbst- und Fremdeinschätzungsskalen	354
4.4	Apparative Messverfahren	356
4.5	Motorische Verhaltenstests	358
4.6	Strategie zum Einsatz von Diagnostik	360
5	Ausgewählte Testverfahren	363
5.1	Beispiele motorischer Funktionstests	363
5.2	Beispiele von Selbsteinschätzungsskalen	365
5.3	Beispiele motorischer Verhaltenstests	369
5.4	Tabellarische Übersicht zu motorischen Testverfahren	371
6	Schlussfolgerungen für die Gesellschaft	379
	Literatur	381

8. Kapitel: Grundlagen des adaptiven Testens

Von Otto B. Walter

1	Adaptives Testen als diagnostische Strategie	389
2	Adaptives Testen und computeradaptives Testen	391
3	Item-Response-Theorie als Grundlage für adaptives Testen	392
4	Konstruktion von adaptiven Tests	394
4.1	Item-Response-Modelle	394
4.2	Itempool und Itemkalibrierung	396
4.3	Bestandteile eines adaptiven Testalgorithmus	396
5	Fazit	399
	Literatur	400

9. Kapitel: Computer-adaptives Testen

Von Jonas W. B. Lang

1	Illustratives Beispiel	405
2	Historische Entwicklung und heutiger Einsatz von CATs	408
2.1	Historische Vorbilder	408
2.2	1970er Jahre	410
2.3	1980er Jahre bis heute	410
3	Testtheoretische Grundlagen für computer-adaptives Testen	411
3.1	Das logistische Ein-Parameter- oder Rasch-Modell (1PLM)	411
3.2	Das logistische Zwei-Parameter-Modell (2PLM)	413
3.3	Das logistische Drei-Parameter-Modell (3PLM)	414
3.4	Weitere Modelle für computer-adaptives Testen	415
4	Computer-adaptive Testkonstruktion	416
4.1	Erstellung von Itembanken	416
4.2	Modelwahl, Modellgültigkeit und Kalibrierung	418
4.3	Testalgorithmen	421
5	Praktischer Einsatz von CATs	427
5.1	CATs in wehrpsychologischen Auswahlprogrammen	427
5.2	CATs in großen Bildungsdiagnostikprogrammen	428
5.3	CATs in der psychologischen Diagnostik	429
6	Fazit	439
	Literatur	440

10. Kapitel: Internetbasierte psychologische Diagnostik

Von Timo Gnams, Bernad Batinic und Guido Hertel

1	Einleitung	447
2	Internetbasierte Erhebungstechniken	451
2.1	Internetbasierte Tests und Fragebögen	451
2.2	Internetbasierte Interviews und Gruppendiskussionen	453
2.3	Internetbasierte Beobachtungen	454
3	Umsetzung diagnostischer Aufgaben mithilfe Neuer Medien	457
3.1	Exploration und Anamnese	457
3.2	Persönlichkeitsdiagnostik	459
3.3	Intelligenz- und Leistungsdiagnostik	460
3.4	Klinisch-psychologische Diagnostik	462
3.5	Berufliche Eignungsdiagnostik	463
4	Gütekriterien internetbasierter Methoden	466
4.1	Objektivität	466

4.2	Reliabilität	467
4.3	Validität	472
4.4	Normierung	475
4.5	Fairness	476
4.6	Verfälschbarkeit	478
4.7	Akzeptanz	483
5	Resümee und Ausblick	485
	Literatur	488

11. Kapitel: Apparative Diagnostik

Von Thomas Rammsayer und Jutta Stahl

1	Von der apparativen zur computergesteuerten Diagnostik	499
2	Computergesteuerte apparative Verfahren: Möglichkeiten und Grenzen . .	501
2.1	Zur Genauigkeit der Zeitmessung beim computergesteuerten apparativen Testen	501
2.2	Der Einsatz von Peripheriegeräten zur Reaktionszeiterfassung	502
2.3	Zum zeitlichen Auflösungsvermögen des Computerbildschirms	503
3	Psychophysiologische und bildgebende Verfahren	505
3.1	Elektroenzephalogramm	505
3.1.1	Grundlagen	505
3.1.2	Roh-EEG	506
3.1.3	Ereigniskorrelierte Potenziale	507
3.1.4	Bewertung	508
3.2	Magnetenzephalogramm	509
3.2.1	Grundlagen	509
3.2.2	Bewertung	510
3.3	Positronen-Emissions-Tomografie	510
3.3.1	Grundlagen	510
3.3.2	Bewertung	511
3.4	Magnetresonanztomografie	512
3.4.1	Grundlagen	512
3.4.2	Bewertung	513
3.5	Nahinfrarot-Spektroskopie	513
3.5.1	Grundlagen	513
3.5.2	Bewertung	514
3.6	Elektrokardiogramm	515
3.6.1	Grundlagen	515
3.6.2	Bewertung	516
3.7	Arterieller Blutdruck	516
3.7.1	Grundlagen	516
3.7.2	Bewertung	517

3.8 Elektrodermale Aktivität	518
3.8.1 Grundlagen	518
3.8.2 Bewertung	519
3.9 Elektromyogramm	519
3.9.1 Grundlagen	519
3.9.2 Bewertung	519
3.10 Polygrafische Messungen	520
3.10.1 Grundlagen	520
3.10.2 Bewertung	520
4 Abschließende Bewertung	522
Literatur	524

12. Kapitel: Ambulantes Assessment

Von Bernd Reuschenbach und Joachim Funke

1 Einleitung	529
2 Ambulantes Assessment: Definitionen und Abgrenzungen	530
3 Anwendungsbegründung	536
4 Technische Voraussetzungen	541
5 Methodische Aspekte	547
5.1 Sampling	547
5.2 Auswertung	550
6 Anwendungsbeispiele	553
6.1 Stressforschung	553
6.2 Schmerzmessung	554
7 Bewertung des Ansatzes	555
7.1 Äquivalenz	556
7.2 Reaktivität	557
7.3 Akzeptanz und Compliance	559
7.4 Ethische und rechtliche Aspekte	562
7.5 Probleme und Nachteile des ambulanten Assessments	563
8 Zukünftige Entwicklung des ambulanten Messens	565
8.1 Erweiterte technische Möglichkeiten in der Diagnostik	565
8.1.1 RFID-Chips	569
8.1.2 Tracking und Lokalisierung	571
8.2 Ambulantes Assessment als Beitrag zur diagnostischen Grundlagenforschung	572
8.3 Automatisierte Verbindung von Diagnostik und Intervention	574
8.4 Risiken und Probleme zukünftiger Entwicklungen	576
9 Schlussfolgerungen	580
Literatur	580

13. Kapitel: Einsatz technischer Mittel in der psychologischen Diagnostik

Von Joachim Funke und Bernd Reuschenbach

1	Einleitung	595
2	Systematisierung des technischen Einsatzes in der Diagnostik	596
3	Beispiele moderner technischer Diagnostika	600
3.1	Computersimulationen	602
3.1.1	Anwendungsbegründung	604
3.1.2	Systematisierung	607
3.1.3	Bewertung des Ansatzes	608
3.2	Digitale Lernspiele („serious games“)	609
3.2.1	Anwendungsbegründung	610
3.2.2	Systematisierung	610
3.2.2.1	Schule	611
3.2.2.2	Arbeitswelt	611
3.2.2.3	Militär	612
3.2.2.4	Umwelt	612
3.2.3	Bewertung des Ansatzes	613
3.3	Virtuelle Welten	613
3.3.1	Anwendungsbegründung	615
3.3.2	Beispiele	615
3.3.3	Bewertung des Ansatzes	616
3.4	Videotests	617
3.4.1	Anwendungsbegründung	618
3.4.2	Beispiele	618
3.4.3	Bewertung des Ansatzes	620
3.5	Diagnostikbedarf durch technische Entwicklungen	620
4	Chancen und Risiken zukünftiger Entwicklungen	622
4.1	Chancen	622
4.2	Risiken	623
5	Schlussfolgerungen	624
	Literatur	625
	Autorenregister	633
	Sachregister	659

1. Kapitel

Testorientierte Leistungsdiagnostik: Grundlagen und Probleme, Verfahrensinventar und besondere Einsatzgebiete

Klaus D. Kubinger

1 Einleitung

Psychologisch-diagnostische Verfahren werden zumeist eingeteilt in solche zur Persönlichkeitsdiagnostik und solche zur „Leistungsdiagnostik“; demgemäß spricht man bei letzteren von „Leistungstests“. Allerdings ist dies unscharf: Einerseits wird vieles in der modernen Persönlichkeitsdiagnostik primär leistungsbezogen erfasst (vgl. dazu die sog. „Objektiven Persönlichkeitstests *sensu R. B. Cattell*“ bei Kubinger, 2003b, bzw. Ortner, Proyer & Kubinger, 2006); andererseits sind die Ergebnisse von Leistungstests durch die Persönlichkeit der Testperson stark (mit-)beeinflusst. Und schließlich beobachten wir zwar aktuelle Leistungen, interessieren uns aber für überdauernde Fähigkeiten – d. h., die beobachtbare Leistung wird als Ausdruck einer eigentlich zu messen beabsichtigten, jedoch nicht direkt beobachtbaren Fähigkeit aufgefasst. Unter diesen Einschränkungen soll hier die Bezeichnung „Leistungstest“ beibehalten und verstanden werden.

Ausgehend von der allgemeinen Definition aller psychologisch-diagnostischer Verfahren (vgl. z. B. Kubinger, 2009a) ist ein Leistungstest wie folgt definiert: Ein psychologischer Leistungstest erhebt unter standardisierten Bedingungen eine Verhaltensstichprobe über die Testperson, indem mit systematisch erstellten Aufgaben die interessierenden Verhaltensweisen oder psychischen Vorgänge ausgelöst, beobachtet und bewertet werden.

2 Einteilung von Leistungstests

Psychologische Leistungstests sind vornehmlich durch die Klasse der „Intelligenztests“ bekannt. Wobei dem einzelnen Test jeweils eine spezifizierte Intelligenztheorie zugrunde liegt (zu einer Übersicht gängiger Intelligenztheorien vgl. z. B. Amelang, Bartussek, Stemmler & Hagemann, 2006; zu einer beispielhaften Verankerung eines bestimmten Intelligenztests innerhalb der von Intelligenztests vgl. Kubinger, Litzenberger & Mrakotsky, 2006); und zwar auch dann, wenn dem Test bloß ein pragmatischer Ansatz zugrunde liegt und (nur) versucht wird, „die Gesamtheit aller kognitiven Voraussetzungen“ zu erfassen, „die notwendig sind, um Wissen zu erwerben und Handlungskompetenzen zu entwickeln“ (Kubinger, 2009b, S. 23). Mit „Kognition“ wird jeder Prozess gemeint, durch den sich der Mensch seiner Umwelt bewusst wird bzw. sich über sie Kenntnisse verschafft (also vor allem Wahrnehmen und Erkennen, Merken, Lernen und Urteilen sowie Vorstellen und Denken).

Zumeist handelt es sich bei Intelligenztests eigentlich um eine Zusammenstellung von mehreren Untertests zu einer Testbatterie. Über diese Intelligenz-Testbatterien hinaus gibt es spezielle Leistungstests, zum Beispiel für die Messung bestimmter neuropsychologischer Funktionstüchtigkeiten. Manche solcher speziellen Leistungstests sind insofern ebenfalls den Intelligenztests zuzurechnen, als sie einen einzelnen etablierten Intelligenzfaktor erfassen. Umgekehrt existieren auch Testbatterien mit mehreren Untertests, die nur auf spezielle Leistungen bzw. Fähigkeiten abzielen und mit Intelligenz im engeren Sinn weniger zu tun haben, wie zum Beispiel sogenannte „Konzentrationstests“. Eine mögliche Einteilung der speziellen Leistungstests könnte sein:

- Sprachverständnis,
- räumliches Vorstellungsvermögen,
- Gedächtnis,
- Verarbeitungsgeschwindigkeit,
- logisches Denken,
- intellektuelle Lernfähigkeit,
- Aufmerksamkeit und Konzentration,
- technisches Verständnis.

Intelligenz-Testbatterien wie spezielle Leistungstests sind also dadurch gekennzeichnet, dass sie die kognitiven Fähigkeiten eines Menschen erfassen wollen – insofern wäre auch der Oberbegriff „Kognitionstest“ angemessen.

3 Voraussetzungen von Tests zur Leistungsdiagnostik

Mit der *Psychologischen Diagnostik* hat sich ausgehend von der Testkonstruktionslehre Gullikssens (1950) die sogenannte „Klassische Testtheorie“ entwickelt. Sie formuliert maßgebliche Haupt- und Nebengütekriterien, die für in der Praxis einsetzbare Tests gegeben sein müssen. Darunter fallen vor allem die bekannten testtheoretischen Ansprüche wie Objektivität, Reliabilität und Validität. Nicht nur wegen der methodischen Weiterentwicklung wurden die ursprünglichen Gütekriterien um andere erweitert, sondern vor allem, weil heutzutage der Konsumentenschutz auch für psychologische Tests relevant ist. Das heißt, der Konsument (der Begutachtete, der Klient oder Patient, die Testperson) sollte die Garantie haben, fachgemäß psychologisch behandelt und begutachtet zu werden. Die gegenwärtig wohl bedeutendste rechtsnahe Vorgabe ist die DIN 33430 (DIN Deutsches Institut für Normung e. V., 2002). Sie regelt konkret die „Anforderungen an Verfahren und deren Einsatz bei der berufsbezogenen Eignungsbeurteilung“, wobei sie sich in wesentlichen Teilen auf die bereits angesprochenen Gütekriterien bezieht (vgl. Westhoff et al., 2005).

Ein weiteres, relativ neues Gütekriterium ist die der „Modernen Testtheorie“ (Kubinger, 1989b) entstammende Skalierung, die u. a. Tests zur Leistungsdiagnostik erfüllen sollten (vgl. Testkuratorium, 2006). Hierunter versteht man (vgl. z. B. Kubinger, 2009a, S. 82): „Ein Test erfüllt das Gütekriterium Skalierung, wenn die laut Verrechnungsvorschriften resultierenden Testwerte die empirischen Verhaltensrelationen adäquat abbilden.“ Gemeint ist damit, dass die vom Testautor festgelegte Regel, wie die einzelnen Testleistungen bzw. -reaktionen einer Testperson zu einem numerischen Testwert zu verrechnen sind, „angemessen“ insofern ist, als sie im Sinn der Theorie des Messens faktisch gerechtfertigt und nicht nur „apodiktisch“ bzw. willkürlich festgesetzt ist.

Die angesprochene Moderne Testtheorie oder „Item-Response-Theorie“ (IRT) bietet nun Erkenntnisse und mathematisch-statistische Modelle, die im Wesentlichen auf eindimensionale Messungen abzielen (vgl. eine darüber hinausgehende mehrdimensionale Messung, insbesondere das sog. *multidimensional Rasch model* bei Adams, Wilson & Wang, 1997). Freilich versuchen die meisten Leistungstests der *Psychologischen Diagnostik*, (je Untertest) eindimensional zu messen. Dies ist nicht nur daraus abzuleiten, dass die Messintentionen von Tests regelmäßig „eindimensional“ begrifflich benannt/beschrieben werden (z. B. „schlussfolgernde Denkfähigkeit“), sondern auch daraus, dass bei der Interpretation der erhaltenen Testwerte es regelmäßig um das Abwägen als „mehr“ oder „weniger“ geht. Bei der Personalauswahl zum Beispiel geht es

darum, einen möglichst hohen Testwert für eine positive Entscheidung zugrunde zu legen und einen zu niedrigen Testwert als Indiz für zu geringe berufliche Eignung anzusehen. Umgekehrt ist es offensichtlich kritisch, wenn die bei einem Test zu erbringende Leistung nicht nur von einer einzigen, bestimmten Fähigkeit abhängt, sondern auch noch von anderen.

Dann ergibt sich das Problem, dass der berechnete Testwert für den Anwender oft wertlos ist, und somit die gegebene psychologisch-diagnostische Fragestellung nicht beantwortet werden kann. Soll ein Test „Lernfähigkeit“ erfassen, aber misst im hohen Grad auch „Lesefähigkeit“ (weil es sich um syntaktisch besonders anspruchsvolle Texte handelt), dann ist ein niedriger Testwert nicht eindeutig zu interpretieren: Entweder die Testperson verfügt über eine geringe Lernfähigkeit oder sie verfügt über eine geringe Lesefähigkeit; oder es mangelt ihr gar an beidem. Misst zum Beispiel ein Test mit der Messintention „Konzentrationsfähigkeit“ im hohen Grad auch „Instruktionsverständnis/sprachliches Auffassungsvermögen“ (weil es sich um semantisch besonders anspruchsvolle Texte handelt), dann ist ein niedriger Testwert ebenfalls nicht eindeutig zu interpretieren: Entweder die Testperson verfügt über eine geringe Konzentrationsfähigkeit oder sie verfügt über ein geringes Instruktionsverständnis; oder es mangelt ihr gar an beidem. Eindimensionalität der gemessenen Fähigkeit ist also fast immer eine wesentliche Voraussetzung von Tests zur Leistungsdiagnostik.

Aufbauend auf dem Konzept eindimensionalen Messens geht es nun innerhalb der IRT im Zusammenhang mit dem Gütekriterium Skalierung darum, je Verrechnungsvorschrift ein Modell anzubieten, welches einen zur Diskussion stehenden psychologischen Test in Bezug auf die oben angesprochene „Angemessenheit“ zu prüfen erlaubt. Das bedeutet: Modelle, die erstens empirisch prüfbar sind und die zweitens mindestens eine hinreichende, besser die notwendige Bedingung darstellen, die erfüllt sein muss, damit die jeweilige Verrechnungsvorschrift fair ist – also die empirisch gegebenen Verhaltensrelationen bei den einzelnen Testaufgaben sowohl innerhalb einer Testperson als auch zwischen verschiedenen Testpersonen mittels der resultierenden Testwerte adäquat abgebildet werden. Einen Katalog solcher Modelle bzw. Verrechnungsvorschriften gibt Kubinger (1989a). Im Folgenden soll es genügen, ein einziges solches Modell näher zu betrachten. Dieses ist das sogenannte *Rasch-Modell* und bezieht sich auf die Verrechnungsvorschrift: Testwert ist gleich Anzahl gelöster Aufgaben. Es handelt sich dabei sowohl um den einfachsten Fall einer Verrechnungsvorschrift als auch um den am weitesten verbreiteten.

Das Rasch-Modell beschreibt die Wahrscheinlichkeit, dass die Testperson v Item i löst („+“), in Abhängigkeit eines (eben eindimensionalen) Personenpara-

mers, ξ_v , als der (wahren) Fähigkeit von v , und eines (gleichfalls eindimensionalen) Aufgabenparameters, σ_i , als der (wahren) Schwierigkeit von i . Das heißt, eine bestimmte Fähigkeit ξ_v bedingt nicht deterministisch, ob es zu einer Lösung kommt oder nicht, sondern nur probabilistisch in der Hinsicht, dass die Lösungswahrscheinlichkeit für größere ξ , bei konstantem σ , ebenfalls größer wird. Konkret lautet das Modell:

$$P(+|\xi_v, \sigma_i) = \frac{e^{\xi_v - \sigma_i}}{1 + e^{\xi_v - \sigma_i}}. \quad (1)$$

Fischer (1995) hat bewiesen, dass dieses Modell notwendigerweise gelten muss, damit die genannte Verrechnungsvorschrift fair ist. Die Tragweite des Beweises liegt im Attribut „notwendigerweise“. Mit ihm ist nämlich der Umkehrschluss zwingend, dass Tests nicht verrechnungsfair sind, für die das Rasch-Modell nicht gilt. Möglich wäre es zwar, dass schlicht ein anderer Verrechnungsmodus verhaltensadäquat ist, zum Beispiel wenn die einzelnen Antworten aufgabenspezifisch gewichtet würden; bezüglich des festgelegten Verrechnungsmodus: Testwert ist gleich Anzahl gelöster Aufgaben, sind sie es jedenfalls dann nicht. Und hier ist es auch tatsächlich möglich, die Geltung des Modells empirisch zu prüfen – im Gegensatz zu vielen anderen (testtheoretischen) Modellen, für die sogenannte *Goodness-of-fit*-Tests lediglich feststellen, wie gut die beobachteten Daten durch das Modell erklärt werden können. Wie dabei im Sinne eines Standards Modelltests zum Rasch-Modell vorzugehen ist, zeigt Kubinger (2005).

4 Gestaltungsweisen von Tests zur Leistungsdiagnostik

Grundsätzlich können die Tests zur Leistungsdiagnostik nach vier Gesichtspunkten eingeteilt werden:

1. freies Antwortformat versus Multiple-Choice-Format,
2. Power- versus Speed-and-Power-Test,
3. Gruppen- versus Individualtest,
4. Papier-Bleistift-Test versus Computertest.

4.1 Antwortformat

Grundsätzlich ist das sogenannte „freie Antwortformat“, manchmal auch als „offenes“ Antwortformat bezeichnet, leicht zu beschreiben: Die Testperson reagiert auf die gestellte Aufgabe beim Lösungsversuch meist mit einer münd-

lichen, zumindest sprachlichen Aussage in mehr oder weniger detaillierter selbstgewählter Ausführlichkeit. Es sind aber auch nonverbale Tätigkeiten, zum Beispiel mathematische Ausarbeitungen oder gewisse (fein-)motorische Aktionen denkbar. Das gezeigte Verhalten muss der Testleiter dann nach vorgegebenen Bewertungsregeln verrechnen, und zwar wieder am einfachsten nach „richtig“ oder „falsch“. Zur Illustration gelte folgendes Item: „*Warum sollte man Obst vor dem Essen waschen?*“ (Lösungen: „um Schmutz abzuwaschen“ oder „um Pflanzenschutzmittel abzuwaschen“).

Allerdings gibt es eine besondere Variante des freien Antwortformats, die sehr einfach zu verrechnen ist: Kubinger, Holocher-Ertl und Frebort (2006) bezeichnen es als „Kästchen“-Format, wenn bei einer Testaufgabe das numerische Ergebnis einer Berechnung in vorgegebene, dem Stellenwert entsprechende Kästchen eingetragen werden muss (prinzipiell möglich ist auch das Eintragen von kurzen Wörtern). Werden die Kästchen als Lesebereich definiert, ist eine Auswertung auch über Scanner bzw. Computer möglich. Als Beispiel sei folgendes Item entworfen: „*Wie lautet die um 2 größere Zahl von 29?*“

Als „gebundenes“ Antwortformat wird das bekannte Multiple-Choice-Format in all seinen Variationen bezeichnet: Aus einer Liste von Antwortmöglichkeiten (Lösungsvorschlägen) muss die Testperson auswählen, welche sie davon für richtig erachtet. Herkömmlich beinhaltet eine Aufgabe im Multiple-Choice-Format eine einzige Lösung, die übrigen Antwortmöglichkeiten nennt man „Distraktoren“: Letztere sollen der Lösung möglichst nahe kommen, und trotzdem in gewisser Hinsicht falsch sein. Außerhalb der *Psychologischen Diagnostik* wird das Multiple-Choice-Format zumeist in Form von vier Antwortmöglichkeiten benutzt, eine Lösung samt drei Distraktoren. Um erfolgreiches Raten seitens der Testperson wenig wahrscheinlich zu machen, verwendet man allerdings innerhalb der *Psychologischen Diagnostik* besser mehr Antwortmöglichkeiten.

Damit die Rateeffekte so klein wie möglich gehalten werden, gibt es die Variante, unter den Antwortmöglichkeiten mehr als nur eine einzige Lösung zu bieten. Kubinger, Holocher-Ertl und Frebort (2006) haben vorgeschlagen, im Gegensatz zum klassischen Multiple-Choice-Format „1 aus 5“ oder „1 aus 6“, „2 aus 5“ oder „x aus 5“ zu verwenden. Gemeint ist in den beiden ersten Fällen, dass eine Antwort der fünf bzw. sechs Antwortmöglichkeiten richtig ist, bzw. bei den letzten Fällen, das entweder genau zwei der fünf Antwortmöglichkeiten richtig sind (und die Testperson darüber explizit informiert ist) oder mehrere der fünf Antwortmöglichkeiten richtig sind: Keine Antwortmöglichkeit, eine, zwei, drei, vier oder sogar alle fünf Antwortmöglichkeiten (worüber die Testperson gleichfalls informiert sein muss). Zur Illustration sei das Item in Kasten 1 entworfen.

Kasten 1:

Beispielitem mit dem Multiple-Choice-Format „x aus 5“
(in Klammer sind die richtigen Antwortmöglichkeiten mit Häkchen gekennzeichnet)

An Deutschland grenzt/grenzen:	
a) Slowenien	()
b) Polen	(✓)
c) Italien	()
d) Österreich	(✓)
e) Niederlande	(✓)

4.2 Zeitbegrenzung

Sogenannte „Power-Tests“ messen eine bestimmte (kognitive) Fähigkeit ohne Zeitbegrenzung, die die Leistung beeinflussen könnte; „Speed-Tests“ dagegen messen als Fähigkeit die Verarbeitungsgeschwindigkeit, ohne anspruchsvolle (sonstige) kognitive Anforderung. „Speed-and-Power-Tests“ erfassen sowohl eine bestimmte (kognitive) Fähigkeit als auch (Verarbeitungs-)Geschwindigkeit, d. h. sie stellen anspruchsvolle Leistungsanforderungen unter Zeitdruck. Genau genommen ist dabei zu differenzieren zwischen solchen Speed-and-Power-Tests, bei denen die insgesamt Bearbeitungszeit beschränkt ist, so dass nicht alle Testpersonen tatsächlich alle Items bearbeiten, und solchen, bei denen jedes einzelne Item mit einer Zeitbegrenzung versehen ist, d. h., jede Testperson bearbeitet das Item zeitlich begrenzt, ohne es notwendigerweise auch (richtig oder falsch) zu beantworten. Als Beispiel sei das in Abbildung 1 dargestellte Item aus dem AID 2 (Kubinger, 2009b) angeführt: „Wenn du (alle) diese Teile richtig aneinandersetzt, wird eine Figur daraus ...“ (verfügbare Lösungszeit für eine Verrechnung von 1 Punkt: 1 Minute 30 Sekunden, für 2 Punkte: 20 Sekunden).

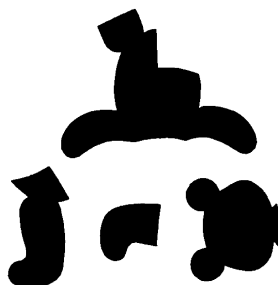


Abbildung 1:

Item „ⓐ-1“ des Untertests *Antizipieren und Kombinieren*-figural aus dem AID 2;
als Lösungsfigur ergibt sich „Teddybär“

4.3 Sozialform

Trivialerweise können Gruppentests gleichzeitig, auch ohne Nutzung eines Computers mehreren Testpersonen vorgegeben werden. Bei Individualtests gilt, dass die besondere Vorgabemodalität bzw. besondere Materialien eine Testvorgabe an mehr als eine einzige Testperson gleichzeitig praktisch nicht ermöglichen. Computertests sind in gewisser Weise als Individualtests aufzufassen, auch wenn mehrere Computer-Testplätze in ein und demselben Raum verfügbar sind, und mehr als nur eine einzige Testperson getestet wird. Ein wesentlicher Vorteil der Individualtestung ist nämlich hier erfüllt, nämlich die Möglichkeit, individuell verschiedene Tests vorzugeben und nicht zwingend mit jedem Test zum selben Zeitpunkt zu beginnen. Was Gruppentests betrifft, gilt erfahrungsgemäß, dass pro Testleiter nicht mehr als sechs bis höchstens acht Testpersonen sachgemäß getestet werden können; dies trifft wohl auch für Computer-Testplatzsysteme zu (vgl. zur Illustration die Raumanordnung eines in der Praxis bewährten Computer-Testplatzsystems in Abb. 2).

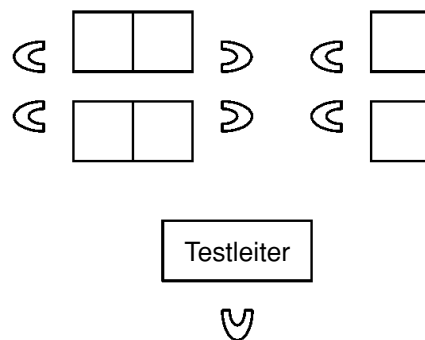


Abbildung 2:

Raumanordnung eines in der Praxis bewährten Computer-Testplatzsystems.

4.4 Technik

Papier-Bleistift-Tests und Computertests unterscheiden sich ganz offensichtlich und selbst bei solchen Tests, die ursprünglich zur Testung mit Papier und Bleistift vorgesehen waren, kann die Übertragung auf den Computer psychologisch relevant abweichende Ergebnisse bedingen; sogenannte Äquivalenzprüfungen (vgl. Wagner-Menghin, 2003a) sind bei solchen Übertragungen unabdingbar. Prinzipiell gibt es zwischen Papier-Bleistift-Tests und Computertests ähnliche Unterschiede wie zwischen Gruppentests und Einzeltests. Dies ist nämlich dadurch bedingt, dass nicht jede Vorgabemodalität mit beiden Techniken realisierbar ist, nicht jedes Testmaterial am Computer umsetzbar ist, und nicht jede Population mit beiden Techniken zumutbar getestet werden kann (vgl. das Gütekriterium Zumutbarkeit z. B. bei Kubinger, 2009a). Zur Illustration sei der

Test *Signal Detection* (Schuhfried, 1986) angeführt, bei dem jeweils dann eine Reaktionstaste gedrückt werden muss, wenn in der laufend wechselnden, komplexen Punktkonfiguration ein Quadrat bestimmter Größe auftaucht (vgl. Abb. 3).

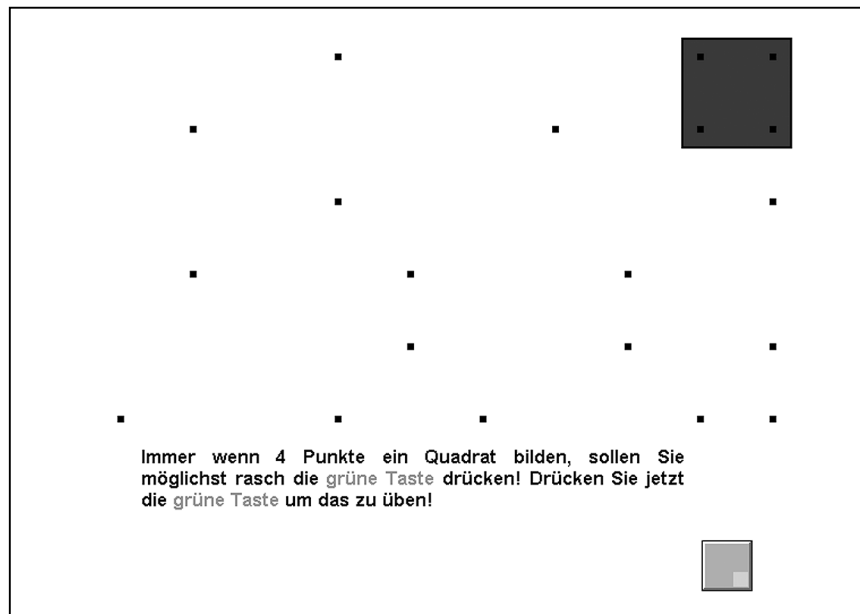


Abbildung 3:

Screenshot aus der Instruktion zum Test *Signal Detection*
(der Abdruck erfolgt mit freundlicher Genehmigung von Schuhfried GmbH)

Grundsätzlich sind die verschiedenen Gestaltungsweisen nicht beliebig kombinierbar. Ein freies Antwortformat (mit Ausnahme des Kästchen-Formats, vgl. Abschnitt 4.1) findet sich heutzutage noch (!) eher nur bei (Papier-Bleistift-)Individualtests, kaum bei Gruppen- und Computertests, bei welchen dieses Antwortformat schlicht unpraktisch weil unökonomisch ist. Umgekehrt verbietet sich das Multiple-Choice-Format eigentlich für (Papier-Bleistift-)Individualtests; wenn sich schon ein Testleiter ausschließlich einer einzigen Testperson widmen muss, dann sollte auch die regelmäßig beim freien Antwortformat zusätzlich anfallende Information für den diagnostischen Prozess genutzt werden. (Papier-Bleistift-)Gruppentests wiederum machen Speed-and-Power-Tests (bei denen die insgesamt Bearbeitungszeit beschränkt ist) deshalb praktisch notwendig, selbst wenn man ursprünglich an der Messung einer bestimmten (kognitiven) Fähigkeit ohne leistungsbeeinflussende Zeitbegrenzung interessiert ist, weil dabei, anders als bei Individual- und Computertests, ein individuelles Zeitbudget bei der Testbearbeitung schwerlich zugestanden werden kann. Allerdings erfordern wiederum solche Speed-and-Power-Tests, bei denen jedes einzelne Item mit einer Speed-Komponente versehen ist, im Fall eines Papier-Bleistift-Tests auch einen Individualtest, weil anders die Zeitkontrolle nicht möglich ist.

5 Technische Möglichkeiten und Grenzen der Leistungsdiagnostik

Die Praxis *psychologischen Diagnostizierens* läuft manchmal an den Konzepten und Erkenntnissen der *Psychologischen Diagnostik*, insbesondere ihrer Grundlagenforschung, vorbei. Was die eben angesprochenen unterschiedlichen Gestaltungsweisen betrifft, wird darauf weiter unten nochmals eingegangen. Davon abgesehen finden sich aber in der geübten Praxis etliche Unzulänglichkeiten, die zu irreführenden Interpretationen führen; das betrifft folgende Themenbereiche:

- Messgenauigkeit,
- Ökologische Validität,
- Stabilität des gemessenen Merkmals,
- Globalisierungsanspruch,
- Anforderungsmodell.

5.1 Messgenauigkeit

Zwar ist die Messgenauigkeit grundlegender Ausbildungsstoff wohl in jedem Studiengang Psychologie, trotzdem ist ihre kritische Reflexion/Bewertung kaum in psychologischen Gutachten der Praxis zu finden.

Innerhalb der Klassischen Testtheorie wird bekanntlich der „Standardmessfehler“ über die geeignet ermittelte Reliabilität eines Tests bestimmt. Damit kann ein Konfidenzintervall für den wahren, gesuchten Messwert (ω) der Testperson errechnet werden. Nach üblicher Festlegung beinhaltet dann in 95 % der Fälle das solcherart berechnete Konfidenzintervall den gesuchten, wahren Messwert irgendwo in diesem Intervall; in 5 % der Fälle liegt der wahre Wert aber außerhalb. Mit ρ als der Reliabilität eines Tests und damit $1-\rho$ als der „Unreliabilität“ sowie mit σ als der Standardabweichung der geeichten Testwerte ergibt sich der angesprochene Standardmessfehler (δ) als $\sigma \times \sqrt{1-\rho}$. Das fragliche Konfidenzintervall für die angedachte Irrtumswahrscheinlichkeit von $\alpha = .05$ bestimmt sich dann für einen beobachteten Testwert (X) wie folgt:

$$\omega_{\min, \max} = X \pm 1.96 \times \delta \approx X \pm 2\delta. \quad (2)$$

In der Praxis wird regelmäßig übersehen, dass dieses Konfidenzintervall für übliche Testreliabilitäten extrem groß und daher interpretationsbezogen ziemlich uninformativ ist (vgl. z. B. Kubinger, 2009a): Für einen Test mit einer Reliabilität $\rho = .80$ folgt etwa in IQ-Werten (mit $\sigma = 15$) bei einem beobachteten Testwert $IQ = 100$, dass wegen $\omega_{\min, \max} \approx X \pm 2\delta = 100 \pm 2 \times 15 \times \sqrt{1-.80} = 100 \pm 13.4$ der wahre Messwert ω im Intervall von 86.6 bis 113.4 angenommen werden muss

– und das in 5 von 100 Fällen auch noch falsch ist. Da bekanntlich die IQ-Werte extra so geeicht sind, dass zwischen $IQ = 90$ und $IQ = 110$ die leistungsbezogenen mittleren 50% einer Population zu finden sind, ist die Aussage „der wahre Messwert liegt zwischen 86.6 bis 113.4“ ziemlich „leer“. Sie beinhaltet nämlich nicht nur die (konventionelle) Interpretation, dass der wahre Messwert als „durchschnittlich“ bezeichnet werden kann, sondern auch die Möglichkeit, dass der zu $IQ = 100$ gehörige wahre Messwert als unterdurchschnittlich (unter dem Prozentrang 25) zu beurteilen ist, wie sogar die Möglichkeit, dass der wahre Messwert als überdurchschnittlich (über dem Prozentrang 75) zu beurteilen ist. Für den analogen Fall einer Testreliabilität von $\rho = .95$ ergibt sich übrigens ein Konfidenzintervall von 93.3 bis 106.7. Wagner-Menghin (2003b) bietet genauer eine Grafik, ab welcher Reliabilität das Konfidenzintervall um einen exakt durchschnittlichen beobachteten Testwert nicht mehr den Prozentrang 25 bzw. 75 unter- bzw. überschreitet: .87. Aus der Grafik ist auch zu entnehmen, dass für einen Test mit Reliabilität $\rho = .87$ der IQ-Wert $IQ \geq 122.5$ betragen müsste, um unter dem einkalkulierten Risiko von 5% (nämlich dass diese Schlussfolgerung falsch ist) den wahren Testwert als überdurchschnittlich interpretieren zu können – oder $IQ \leq 77.5$, um den wahren Testwert als unterdurchschnittlich zu interpretieren.

Für die Praxis bedeutet das aber, dass Testwerte von wenig reliablen Tests fahrlässig unsachlich interpretiert werden. Die Reliabilität eines Tests ist also keine Kenngröße, deren Abweichung vom Ideal $\rho = 1$ beliebig toleriert werden kann.

Ein methodischer Zugang, um bei weniger reliablen Tests eher trotzdem aussagekräftige Interpretationen zu erhalten, ist der hypothesengeleitete Ansatz, nur einseitige Konfidenzintervalle im Blick auf eine konkrete Entscheidungsfindung zu bestimmen. Bei gleichen Risiko (nämlich dass die letztlich getroffene Schlussfolgerung falsch ist), wird dabei die Intervallbreite im relevanten Testwertebereich kleiner. Für das angesprochene Risiko von wieder 5% ergibt sich zum Beispiel $\omega_{\max} = X + 1.645 \times \delta \approx X + 1.5 \times \delta = 85 + 1.5 \times 15 \times \sqrt{1 - .87} = 93$. Damit sind etwa Fragestellungen beantwortbar, ob trotz unterdurchschnittlich beobachteten Testwerts (im Beispiel $IQ = 85$) der wahre Testwert als durchschnittlich (im Sinn eines Prozentrangs größer als 25) bezeichnet werden kann.

Wenig empfehlenswert ist dagegen die jüngst manchmal feststellbare Vorgehensweise, das eingegangene Risiko (nämlich für eine letztlich falsch getroffene Schlussfolgerung) zu erhöhen, zum Beispiel auf 10% – statt $\omega_{\min, \max} \approx X \pm 2\delta = 100 \pm 2 \times 15 \times \sqrt{1 - .80} = 100 \pm 13.4$ zum Beispiel ergäbe sich dabei nämlich $\omega_{\min, \max} = X \pm 1.645 \times \delta \approx 100 \pm 1.5 \times 15 \times \sqrt{1 - .80} = 100 \pm 8.1$. Es scheint kaum verantwortlich, durchschnittlich jede zehnte Testperson bei einer solchen Vorgehensweise in Bezug auf ihren wahren Testwert fälschlich zu interpretieren.

Besonders kritisch wird die Missachtung der Messgenauigkeit bei einer sogenannten „Profilinterpretation“. Dabei geht es darum, die einzelnen Untertestleistungen einer Testperson zueinander in Relation zu setzen, also „Höhen“ und „Tiefen“ zu bestimmen. Offensichtlich gibt es bei wenig reliablen Untertests kaum Testwerte, die sich, kalkuliert man den wahrscheinlichen Messfehler mit ein, tatsächlich unterscheiden. Es gibt andernorts (vgl. z. B. Huber, 1973) publizierte Formeln, um die „Signifikanz“ eines Testprofils zu bestimmen, d. h. auszurechnen, ob sich unter Miteinbeziehung des Messfehlers tatsächlich mindestens zwei Testwerte unterscheiden. Diese Formeln beruhen allerdings auf sehr strikten Annahmen, die kaum je praktisch erfüllt sind, so dass der oben angesprochene hypothesengeleitete Ansatz unter Berechnung einseitiger Konfidenzintervalle in Bezug auf bestimmte vermutete Höhen und Tiefen pragmatisch vorzuziehen ist.

Hinsichtlich des möglichen Messfehlers eines Tests ist jedoch über alles bisher Gesagte hinaus überhaupt der Ansatz der Klassischen Testtheorie grundsätzlich kritisch. Er beruht, wie ausgeführt, auf dem Konzept der sogenannte „Reliabilität“, also im Wesentlichen auf einer Korrelation, wie sie in einer (mehr oder weniger für die fragliche Population repräsentativen) Stichprobe bestimmt wurde. Abgesehen davon, dass die Korrelation als eine statistische Methode extrem stichprobenabhängig ist (vgl. Kubinger, 2009a), ist es anschaulich sehr unplausibel anzunehmen, dass die Messgenauigkeit eines Tests für alle Testwerte im gesamten Testwertbereich gleich ist, wie das eben die Klassische Testtheorie mit ihrer Bestimmung des Standardmessfehlers annimmt. Vielmehr ist es erfahrungsgelitet wahrscheinlich, dass im oberen und unteren Testwertbereich wesentlich ungenauer gemessen wird, weil dort jeweils weit weniger Items zwischen ähnlichen Fähigkeitsausprägungen differenzieren können. Dem gegenüber ist der entsprechende Ansatz der IRT bestechend: Dieser ermöglicht einen sogenannten „(Standard-)Schätzfehler“ spezifisch, pro konkret erbrachter Leistung einer Testperson zu berechnen: Geht es zum Beispiel beim Rasch-Modell darum, den (unbekannten) Fähigkeitsparameter ξ_v zu schätzen (vgl. Abschnitt 3), so ist über *Fishers* Theorie der „*information in the sample*“ für die diesbezügliche Maximum-Likelihood-Schätzung zunächst die person- und item-

spezifische „Information“ $I(i, v) = \frac{[P'("+"|i, v)]^2}{P'("+"|i, v) \cdot P'("-" |i, v)}$ für Person v und Item i

zu bestimmen, und daraus der Standardschätzfehler für ξ_v :

$$\varepsilon(\xi_v) = \sqrt{\frac{1}{\sum_i I(i, v)}} = \left[\sum_i \frac{e^{\xi_v - \sigma_i}}{1 + e^{\xi_v - \sigma_i}} \times \frac{1}{1 + e^{\xi_v - \sigma_i}} \right]^{-\frac{1}{2}}, \quad (3)$$