Dieter Irblich
Gerolf Renner (Hrsg.)

Diagnostik in der Klinischen Kinderpsychologie

Die ersten sieben Lebensjahre

HOGREFE

Kriterien zur Auswahl von Testverfahren in der klinischen Kinderpsychologie

Gerolf Renner

Psychologische Testverfahren sind wichtige Instrumente der klinischen Kinderpsychologie, deren Ergebnisse mit über Diagnosen und Behandlungsempfehlungen entscheiden (→ Kapitel 7). Die Auswahl eines Testverfahrens sollte daher stets sorgfältig und unter Berücksichtigung der anerkannten Qualitätsmaßstäbe erfolgen, wie sie z. B. in den Standards for Educational and Psychological Testing formuliert sind (American Educational Research Association et al., 1999; deutsche Fassung der ersten Auflage herausgegeben von Häcker, Leutner & Amelang, 1998). Für die Bewertung psychologischer Tests haben sich allerdings noch keine allgemeinverbindlichen Maßstäbe etabliert. Mittlerweile liegen jedoch Kriterienkataloge und Bewertungssysteme vor (z. B. European Federation of Psychologists' Associations, 2005; Rudner, 1994; Testkuratorium der Föderation Deutscher Psychologenvereinigungen, 2006; s. a. Kersting, 2006), die eine Systematisierung und Objektivierung der Testbewertung anstreben.

Die im Folgenden aufgeführten Kriterien sollen in erster Linie als Hilfestellung verstanden werden, die es dem Testanwender erlaubt, Stärken und Schwächen eines Testverfahrens systematisch zu beurteilen und eine verantwortungsvolle Entscheidung für die Testauswahl bei der jeweils gegebenen Fragestellung zu treffen. Die Qualität eines Tests kann jedoch prinzipiell nicht unabhängig von der intendierten Anwendung bewertet werden:

- *Beispiel 1:* Ein Intelligenztest, der deutliche Bodeneffekte aufweist, erlaubt keine aussagekräftigen Ergebnisse bei stark entwicklungsverzögerten Kindern, kann aber, sofern er frei von Deckeneffekten ist, dennoch bei der Untersuchung von gut begabten Kindern eingesetzt werden.
- *Beispiel 2:* Bei der Untersuchung eines hypermotorischen, impulsiven und leicht ablenkbaren Kindes kann ein Verfahren, bei dem der Untersucher während der Testung Bewertungsrichtlinien konsultieren, Zeitmessungen vornehmen und umfangreiche schriftliche Notizen machen muss, die Steuerung des kindlichen Verhaltens erschweren und dadurch zu invaliden Testergebnissen führen, auch wenn bei einer Normalpopulation hervorragende psychometrische Kennwerte ermittelt wurden.

Je nach Art des Verfahrens sind die nachfolgend genannten Kriterien von unterschiedlicher Relevanz. So erwartet man bei einem Intelligenztest hinsichtlich der Durchführung differenziertere Angaben der Testautoren als z. B. bei einem Elternfragebogen zur Verhaltenseinschätzung, bei dem wiederum das Vorliegen einer Parallelform entbehrlich ist, die wichtig sein kann, wenn mit Übungseffekten bei Messwiederholungen zu rechnen ist. Die Kriterien berücksichtigen die ganze Bandbreite von zentralen Gütekriterien bis zu Fragen der Handhabung und Haltbarkeit der Testmaterialien.

Merke:

Ziel einer Testevaluation anhand der hier dargestellten Kriterien ist es nicht, einen Test als allgemein „gut" oder „schlecht" einzustufen, sondern eine reflektierte und gut begründete Entscheidung darüber zu ermöglichen, ob und mit welcher Aussagekraft ein Testverfahren für die bestimmte klinische Anwendung eingesetzt werden kann.

In der Hand des erfahrenen Diagnostikers können bisweilen auch psychometrisch weniger abgesicherte Methoden (z. B. „testing the limits", → Kapitel 2) wertvolle Hinweise liefern, doch sollte die Testauswahl im Allgemeinen von dem Bemühen um exakte und aussagekräftige Ergebnisse geleitet werden. In der klinischen Einzelfalldiagnostik sind bei Entscheidungen, die langfristige und bedeutsame Konsequenzen für den Lebensweg eines Kindes haben, höhere Maßstäbe anzulegen, als beim Einsatz eines Verfahrens zur Hypothesengenerierung oder in wissenschaftlichen Gruppenstudien.

Bei einzelnen Kriterien (z. B. Reliabilität) werden im Folgenden für die Beurteilung numerische Kriterien vorgeschlagen. Natürlich muss dabei berücksichtigt werden, dass es sich nicht um rigide Grenzen handeln kann und diese Empfehlungen z. T. auf subjektiven Entscheidungen beruhen.

Beispiel: Reliabilitätskennwerte sind Schätzungen, die von Stichprobe zu Stichprobe und in Abhängigkeit von der Bestimmungsmethode variieren. Ein Test, der eine Reliabilität von .78 aufweist, muss daher keineswegs unbrauchbarer sein als ein Verfahren mit einer Reliabilität von .82.

Einzelne Forderungen (z. B. hinsichtlich der Qualität der Normierung) werden derzeit nur von wenigen Testverfahren erfüllt. Dies ist kein Grund, den Einsatz von Tests grundsätzlich abzulehnen. Vielmehr geht es darum, anhand der Kriterien deutlich zu machen, wo der Untersucher Testergebnisse u. U. mit besonderer Vorsicht interpretieren, durch weitere Datenquellen absichern und im Verlauf überprüfen muss. Im Vergleich zu einer allein auf subjektiven Maßstäben beruhenden Einschätzung kann ein Testverfahren auch dann zu einer besseren Beurteilung der kindlichen Fähigkeiten beitragen, wenn es nicht alle wünschenswerten Eigenschaften aufweist. Die Spannung zwischen den realen Möglichkeiten des praktischen Testens und den Forderungen der Testtheorie sollte als Aufforderung zu einer kontinuierlichen Verbesserung diagnostischer Instrumente gesehen werden. Für die langfristige Entwicklung der Testqualität ist es eher kontraproduktiv, die Kriterien den gegebenen Möglichkeiten anpassen zu wollen.

Einsatzbereich des Testverfahrens:

- Werden die mit dem Test erfassten Konstrukte (Leistungsbereiche, Fähigkeiten) klar definiert?
- Werden Zielgruppe und Einsatzbereich des Tests klar beschrieben?
- Werden die fachlichen Anforderungen an den Testanwender beschrieben?
- Stehen Parallelformen zur Verfügung?
- Stehen mehrsprachige Instruktionen zur Verfügung?
- Ist der Einsatz des Tests auf bestimmte Zeitintervalle begrenzt (z. B. bei einem Test zur Früherkennung von Lernstörungen auf eine bestimmte Anzahl von Monaten vor der Einschulung)?

- Ist die Einsatzmöglichkeit des Tests bei bestimmten Personengruppen eingeschränkt, z. B. durch Aufgaben, die spezifisches regionales oder kulturelles Wissen voraussetzen?
- Ist der Einsatz des Tests bei Personen mit bestimmten Leistungsbeeinträchtigungen (z. B. motorisch Defizite, Sprachstörungen) eingeschränkt?

Testergonomie und Testmaterialien:

- Wird die allgemeine Testdurchführung übersichtlich und genau beschrieben?
- Sind die spezifischen Instruktionen übersichtlich dargestellt?
- Enthält das Testformular Erinnerungshilfen für die Durchführung?
- Werden bei umfangreichen Skalen altersabhängige Testein- und -ausstiege oder eine adaptive Testvorgabe angeboten?
- Sind alle Vorlagen und Texte, die während der Untersuchung benutzt werden müssen, mühelos aufzuschlagen (z. B. Spiralbindung)?
- Sind Testvorlagen nötigenfalls mit Aufschlaghilfen versehen?
- Zeigen Manuale und Testformulare ggf. die Anordnung des Materials aus Sicht des Untersuchungsleiters?
- Sind die Testmaterialien für eine dauerhafte Anwendung geeignet?
- Können Testmaterialien einzeln nachgekauft werden?
- Sind alle visuellen Materialien hinreichend groß und kontrastreich?
- Sind die Materialien ansprechend (z. B. farbige Gestaltung)?
- Sind Auswertungsschablonen fehlerfrei und einfach zu handhaben?
- Ist das Testformular übersichtlich gestaltet?
- Ist das Testformular völlig fehlerfrei?
- Bietet das Testformular genügend Platz, um die Reaktionen des Kindes aufzeichnen zu können?
- Steht benutzerfreundliche Auswertungssoftware zur Verfügung?

Objektivität:

- Enthält das Manual genaue Angaben zur allgemeinen Gestaltung der Testsituation?
- Wird die Testdurchführung umfassend, detailliert und eindeutig beschrieben?
- Sind die Testinstruktionen wörtlich angegeben?
- Wird beschrieben, wie auf Nachfragen und Bitten um Hilfestellungen oder Rückmeldungen zu reagieren ist? Ist eindeutig geregelt, wann Instruktionen wiederholt oder umformuliert werden dürfen?
- Enthält das Manual Hinweise zum Umgang mit schwierigen Testsituationen?
- Enthält das Manual Hinweise zur Testdurchführung bei Kindern mit Sinnesstörungen oder motorischen Beeinträchtigungen?
- Enthält das Manual Hinweise zur Testdurchführung bei Kindern, die die deutsche Sprache nicht beherrschen?
- Sind die Auswertungsrichtlinien umfassend und eindeutig beschrieben?
- Sind ggf. ausführliche Beispiele für die Bewertung der Antworten vorhanden?
- Enthält das Manual Hinweise auf typische Durchführungsfehler?
- Steht eine Videodemonstration der Testdurchführung zur Verfügung?
- Ist für die Testanwendung der Erwerb spezieller Kompetenzen, z. B. durch Schulungsseminare erforderlich?
- Werden empirische Daten zur Objektivität berichtet?

Reliabilität:
- Liegen Angaben zur split-half-Reliabilität und zur internen Konsistenz vor?
- Wurde die Retestreliabilität bestimmt?
- Liegt die Reliabilität für Skalen, die für diagnostische Entscheidungen relevant sind, über .90?
- Liegt die mittlere Reliabilität von Subskalen über .80?
- Liegt die Reliabilität für einzelne Skalen im inakzeptablem Bereich (< .70)?
- Wurden Reliabilitätskennwerte getrennt für Altersgruppen berechnet?
- Wurden Reliabilitätsschätzungen auch an klinischen Gruppen vorgenommen?
- Gibt es Angaben zu Übungseffekten bei kurz- bis mittelfristiger Testwiederholung?
- Gibt es Angaben zur Profilreliabilität, falls im Manual Profilanalysen vorgeschlagen werden?
- Werden Itemschwierigkeiten und Trennschärfen zumindest zusammenfassend dargestellt?

Validität:
- Werden der theoretische Hintergrund und die erfassten Konstrukte klar beschrieben?
- Werden Unterschiede und Gemeinsamkeiten zu verwandten theoretischen Konstrukten diskutiert?
- Werden bei der Darstellung des theoretischen Hintergrunds aktuelle Forschungsergebnisse berücksichtigt?
- Deckt die Itemauswahl das theoretische Konstrukt umfassend ab?
- Vermeidet die Itemauswahl konstruktirrelevante Einflüsse auf die Testergebnisse?
- Wird die Auswahl der berichteten Daten zur Validität überzeugend begründet und aus den theoretischen Annahmen abgeleitet?
- Finden sich im Manual Daten zu klinischen Gruppen, bei denen der Test angewendet werden kann?
- Gibt es im Hinblick auf den intendierten Anwendungsbereich Daten zur prognostischen Validität?
- Werden Daten (Korrelationen, Mittelwertsunterschiede) zur konvergenten Validität (Zusammenhang mit vergleichbaren Tests) berichtet?
- Werden Daten zur divergenten Validität (Zusammenhang zu konstruktfernen Variablen) berichtet?
- Werden exploratorische Faktorenanalysen nachvollziehbar dargestellt (z. B. Eigenwertverlauf, Kriterium für Zahl der extrahierten Faktoren, Rotationsmethode)?
- Wurde die faktorielle Struktur mit konfirmatorischen Faktorenanalysen überprüft?
- Wurde die Faktorenstruktur für unterschiedliche Altersgruppen geprüft?
- Wurde die Faktorenstruktur für klinische Gruppen geprüft?
- Werden im Manual Daten als Belege für die Validität fehlinterpretiert, die weder theoretisch abgeleitet noch für klinische Entscheidungen relevant sind (z. B. rein deskriptive Daten über Geschlechts- und Alterseffekte usw.)?
- Werden für diagnostische Klassifikationen, die auf den Testergebnissen basieren, Sensitivität, Spezifität, Gesamttrefferquote, positiver und negativer prädiktiver Wert angegeben?

Normierung:
- Sind die Normen aktuell (Kontrolle der Normen nach zehn Jahren)?
- Ist der Zeitpunkt der Normierung im Manual exakt angegeben?
- Ist die Gewinnung der Normstichprobe nachvollziehbar beschrieben?
- Sind Ausschlusskriterien und Umgang mit drop-outs beschrieben?
- Ist die Zusammensetzung der Normstichprobe ausreichend beschrieben (Angaben über Alter, Geschlecht, regionale Herkunft, Bildungsniveau der Eltern, Kindergarten-/Schulbesuch, Mehrsprachigkeit, Migrationshintergrund)?
- Ist die Normstichprobe repräsentativ für die Gesamtpopulation?
- Ist erkennbar, wie die Qualitätssicherung bei der Normierung erfolgte (Qualifikation, Schulung und Kontrolle der Testleiter)?
- Sind die Altersjahrgänge hinreichend stark besetzt (möglichst N > 100, besser N > 200)?
- Ist das Vorgehen bei der Berechnung der Normdaten nachvollziehbar beschrieben?
- Werden Standardnormen verwendet?
- Werden bei nicht normalverteilten Rohwerten Prozentränge oder T-Werte verwendet?
- Ist die Alterseinteilung hinreichend differenziert (Altersgruppen umfassen im Vorschulalter möglichst nicht mehr als 3 Monate (maximal 6 Monate), vor dem 2. Geburtstag maximal 1 bis 2 Monate? Falls dieses Kriterium nicht erfüllt wird, wird dies durch empirische Daten überzeugend begründet?
- Bestehen bei gleichen Rohwerten unverhältnismäßige Unterschiede zwischen den Standardwerten benachbarter Altersgruppen? Wenn ja, wird darauf im Manual ausdrücklich hingewiesen?
- Wird bei der Normierung im Übergangsalter zwischen Kindergarten und Schule berücksichtigt, ob Kinder bereits eingeschult sind?
- Werden Konfidenzintervalle (z. B. auf dem 90 % oder 95 %-Niveau) und ihre Berechnungsmethode angegeben?
- Kann der Testanwender alle Normtabellen problemlos einsehen (keine *ausschließliche* Computerauswertung), so dass er Boden-/Deckeneffekte, Itemgradienten usw. selbst beurteilen kann?
- Weisen die Normen Boden- oder Deckeneffekte auf? Wenn ja, wird darauf im Manual ausdrücklich hingewiesen?
- Erlauben die Itemgradienten eine hinreichende Differenzierung von Leistungsunterschieden? Wenn nein, wird darauf im Manual explizit hingewiesen?
- Ist der Test für das angestrebte Untersuchungssetting (Einzeltest/Gruppentest) normiert?
- Ist die Übertragbarkeit der Normen auf Einzeltestungen nachgewiesen, falls die Normdaten in Gruppentestungen erhoben wurden?

Testinterpretation:
- Sind die Interpretationsmöglichkeiten des Tests im Manual ausführlich beschrieben?
- Korrespondieren die Vorschläge zur Testinterpretation mit den theoretischen Grundlagen?
- Werden die Vorschläge zur Testinterpretation von den Validitätsdaten gestützt?
- Werden Interpretationsvorschläge, die nicht empirisch gesichert sind, eindeutig als solche gekennzeichnet?
- Wird auf potenzielle Interpretationsfehler hingewiesen?

- Wird die Testinterpretation anhand unterschiedlicher Testprotokolle erläutert? Werden dabei auch schwer interpretierbare Testprotokolle berücksichtigt?
- Wird bei der Testinterpretation ggf. auf die Notwendigkeit einer umfassenden und fachgerechten Diagnostik unter Einbeziehung weiterer Datenquellen hingewiesen?
- Stehen für Profilanalysen Signifikanzprüfungen und Daten über die Häufigkeit von Profilunterschieden in der Population zur Verfügung?
- Sind Vorschläge für Profilanalysen empirisch überprüft?
- Werden die Effekte von Boden- und Deckeneffekten bei der Testinterpretation berücksichtigt?
- Wird im Manual eindeutig definiert, ob und ggf. welche Skalen Testrohwerte von Null aufweisen können, ohne die Testinterpretation zu gefährden?

Sonstiges:
- Sind alle Items und Anweisungen in einfacher, klarer und kindgerechter Sprache formuliert?
- Wird das Kind ausreichend in die Aufgabenstellung eingeführt, z. B. durch Übungsitems?
- Werden Angaben zur Testdauer und ihrer Variabilität gemacht?
- Enthält der Test inhaltlich veraltete Items?
- Enthält der Test Items, die bestimmte Personengruppen z. B. aufgrund ihrer Geschlechtszugehörigkeit oder ethnischen Herkunft diskriminieren?
- Werden auch Analysen auf Basis der probabilistischen Testtheorie berichtet, z. B. um die Verrechnungsregeln des Tests zu überprüfen?
- Wird die Testentwicklung (Vorformen, Selektion von Items) zumindest zusammenfassend beschrieben?
- Ist bei Neuauflagen von Tests eindeutig erkennbar, ob und welche Veränderungen vorgenommen wurden?

Literatur

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (1999). *Standards for educational and psychological testing* (2nd ed.). Washington, DC: American Educational Research Association.

European Federation of Psychologists' Associations EFPA (2005) *EFPA review model für the description and evaluation of psychological tests. Test review form and notes for the reviewers.* Verfügbar unter http://www.efpa.be/docpagina.php (Zugriff am 29. 10. 2008).

Häcker, H., Leutner, D. & Amelang, M. (1998). *Standards für pädagogisches und psychologisches Testen.* Göttingen: Hogrefe.

Kersting, M. (2006). Zur Beurteilung der Qualität von Tests: Resümee und Neubeginn. *Psychologische Rundschau, 57,* 243–253.

Rudner, L. M. (1994) *Questions to ask when evaluating Tests.* Washington DC: ERIC Clearinghouse on Assessment and Evaluation. Verfügbar unter http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/14/1b/61.pdf (Zugriff am 31. 10. 2008).

Testkuratorium der Föderation Deutscher Psychologenvereinigungen (2006). TBS-TK Testbeurteilungssystem des Testkuratoriums der Föderation Deutscher Psychologenvereinigungen. *Report Psychologie, 31,* 492–499.

Verbale Beschreibung von Testwerten

Gerolf Renner & Dieter Irblich

Die Kommunikation von Testergebnissen kann zwischen psychodiagnostisch ausgebildeten Fachkräften unproblematisch durch die exakte Weitergabe von numerischen Testwerten und den zugehörigen Konfidenzintervallen erfolgen. In Gutachten- und Berichtstexten ist zusätzlich eine verbale Beschreibung des quantitativen Befundes gebräuchlich. Wenn Untersuchungsbefunde an Personengruppen gerichtet sind, die eine angemessene Interpretation von Testkennwerten nicht leisten können (z. B. Eltern, oft auch Ärzte, Pädagogen und Therapeuten), sind verbale Beschreibungen unabdingbar. Dabei wird kontrovers diskutiert, ob die Weitergabe numerischer Testwerte an nicht qualifizierte Personen sogar einen Verstoß gegen ethische Richtlinien darstellen kann (s. Strauß, Sherman & Spreen, 2006).

Beispiel: Die Mutter eines 6-jährigen Jungen sieht in der Arztpraxis den Bericht einer Frühförderstelle ein, in dem u. a. das Ergebnis eines visuellen Wahrnehmungstests dargestellt wird. Mit einem Prozentrang von 65 hat der Junge ein voll durchschnittliches Ergebnis erzielt. Die Mutter drängt daraufhin auf Aufnahme einer Ergotherapie, da sie aus dem Ergebnis abliest, dass ihr Sohn nur 65 % von dem leistet, was er in seinem Alter können sollte.

Es gibt allerdings bisher keinen Konsens über die Frage, wie Testwerte zu beschreiben sind. Viele Testverfahren machen zwar Vorschläge für die verbale Umschreibung der Ergebnisse, verwenden jedoch unterschiedliche Abstufungen und verbale Bezeichnungen der Kategorien. Dass innerhalb eines psychologischen Berichts, in dem die Ergebnisse mehrerer Testverfahren dargestellt werden, unterschiedliche Bezugssysteme zu einer ausgesprochenen Verwirrung führen würden, versteht sich allerdings von selbst.

Eine konsistente Verwendung verbaler Beschreibungen ist also unbedingt zu empfehlen und kann z. B. durch eine entsprechende Tabelle im Bericht transparent gemacht werden. Dabei sollte unterschieden werden zwischen der Beschreibung eines quantitativen Testbefundes unter Bezug auf die Normgruppe, für die sich am besten Begriffe wie „durchschnittlich" oder „unterdurchschnittlich" eignen (vgl. Westhoff & Kluck, 2008), und der klinischen Interpretation von Testergebnissen, die nicht allein auf dem quantitativen Testwert beruht und für die Begriffe wie „pathologisch" oder „auffällig" vorbehalten bleiben sollten.

Westhoff und Kluck (2008) schlagen vor, bei der verbalen Testbeschreibung auch Konfidenzintervalle (KI) zu berücksichtigen.

Beispiel: Das 90 %-KI eines Testwertes von 84 liegt im Bereich 75 bis 93. Unter Bezug auf die Kategorien in Tabelle 1 wäre der Testwert als „unterdurchschnittlich" zu bezeichnen. Der untere Wert des KI liegt ebenfalls im unterdurchschnittlichen Bereich, der obere Wert ist jedoch als „durchschnittlich" zu bezeichnen, was zu der Formulierung „unterdurchschnittlich bis durchschnittlich" führen würde.