Föderation Deutscher Psychologenvereinigungen

DGPs Deutsche Gesellschaft für Psychologie

Berufsverband Deutscher Psychologinnen und Psychologen

TBS-TK – Testbeurteilungssystem des Testkuratoriums der Föderation Deutscher Psychologenvereinigungen. Revidierte Fassung vom 09. September 2009[1]

Ziel

Das TBS-TK dient Testautoren, Verlagen und Anbietern sowie Rezensenten zur Qualitätssicherung von Tests.[2]

Geltungsbereich

Der Begriff „Test" hat in der Psychologie und erst recht in der nicht psychologischen Öffentlichkeit eine sehr weit gefasste Bedeutung: Er wird praktisch für alle psychologisch-diagnostischen Verfahren, die beim psychologischen Diagnostizieren eingesetzt werden, benutzt. Obwohl ein psychologischer Test im engeren Sinne nur eine besondere Untergruppe solcher psychologisch-diagnostischer Verfahren darstellt, soll die Bezeichnung „Test" im vorliegenden Zusammenhang als Oberbegriff gelten: Es sind also neben Intelligenz- und allgemeinen Leistungstests insbesondere Persönlichkeitsfragebogen, Objektive Persönlichkeitstests sowie Projektive Verfahren, aber auch standardisierte Interviews sowie Erhebungsverfahren zur Arbeitsplatzanalyse gemeint.

Durchführung

1. Die Auswahl der zu rezensierenden Tests erfolgt durch das Testkuratorium (TK). Das TK nimmt Vorschläge für zu rezensierende Tests entgegen.
2. Mit der Beurteilung der ausgewählten Tests werden vom TK zwei Rezensenten beauftragt. Das TK bürgt für die Unabhängigkeit und Unvoreingenommenheit der Rezensenten.
3. Das TK sorgt dafür, dass den Rezensenten der für die Beurteilung notwendige Test sowie weiteres notwendiges Informationsmaterial zur Verfügung stehen. Im Falle von „confidential tests" sichert das TK den Testanbietern die Vertraulichkeit z. B. wettbewerbsrechtlicher Informationen zu.
4. Der Beurteilungsprozess verläuft in drei Schritten:

4.1 Prüfung, ob die Verfahrenshinweise (auch Testmanual, Testhandbuch genannt) alle für eine Testbeurteilung notwendigen Informationen über den Test enthalten. In einem ersten Schritt prüfen die Rezensenten, ob und in welchem Ausmaß die Anforderungen der DIN 33430 an Verfahrenshinweise (Testmanuale) erfüllt sind. Diese Anforderungen sind auf Tests aus allen Bereichen anwendbar. Die Operationalisierung dieser Anforderungen erfolgt mit der „Checkliste 1" der Publikation „DIN Screen" von Kersting (2008). Diese Checkliste dient als Standard des Testkuratoriums der Föderation Deutscher Psychologenvereinigungen hinsichtlich des Qualitätsanspruches an Information und Dokumentation von Instrumenten zur Erfassung menschlichen Erlebens und Verhaltens. Die Checkliste sollte nach Möglichkeit bereits durch den Testanbieter unter Angabe der Seiten in den Verfahrenshinweisen (im Testmanual), auf denen sich die jeweiligen Informationen befinden, ausgefüllt sein. Die Rezensenten kontrollieren diese Angaben und korrigieren sie, wenn nötig. Auf Basis der vorliegenden Informationen stellen die Rezensenten fest, ob der Test „prüffähig" ist. Ein Test, der in diesem Sinne nicht prüffähig ist, weil wesentliche Angaben gemäß DIN 33430 fehlen, erhält ohne weitere Prüfung die Beurteilung „Der Test erfüllt die durch DIN 33430 festgelegten Anforderungen bezüglich Information und Dokumentation nicht".

4.2 Testkategorisierung nach ZPID- und EFPA-System. Im zweiten Schritt erfolgt eine Testkategorisierung, und es werden formale Merkmale des Tests für Datenbanken angegeben. Hierzu werden das ZPID-System (http://www.zpid.de) und Teile des EFPA-Systems (1.10.1 – 3 und 1.11) benutzt (www.bdponline.de/web/report/test.html). Die Angaben sollten nach Möglichkeit bereits durch den Testanbieter unter Angabe der Seiten in den Verfahrenshinweisen (im Testmanual), auf denen sich die jeweiligen Informationen befinden, vorgenommen werden. Die Rezensenten kontrollieren diese Angaben und korrigieren sie, wenn erforderlich.

4.3 Bewertung des Tests anhand der Besprechungs- und Beurteilungskategorien des TK. Im dritten und letzten Schritt erfolgt die eigentliche Beurteilung des Tests durch die Rezensenten. Mit diesem Schritt wird eine Bewertung des Tests auf Basis der Angaben in den Verfahrenshinweisen (im Testmanual) vorgenommen. Die „Richtlinien des Testkuratoriums für die Beurteilung von Tests zur Erfassung menschlichen Erlebens und Verhaltens" (siehe Anhang) erläutern, wie die in der DIN 33430 spezifizierten Informationsanforderungen (DIN-Screen Checkliste 1) den Beurteilungsaspekten des TBS-TK Systems zuzuordnen sind. Die Rezensenten können weitere Erkenntnisquellen zum Verfahren heranziehen, sofern diese allgemein zugänglich sind.

Die Beurteilung gliedert sich in sieben „Besprechungs- und Beurteilungskategorien des Testkuratoriums" gemäß Tabelle 1 und eine zusammenfassende Abschlussbewertung. Für alle Kategorien ist eine frei formulierte Bewertung im Umfang von maximal 1000 Zeichen

---

[1] Dieser Artikel ist in folgender Weise zu zitieren: Testkuratorium. (2010). TBS-TK. Testbeurteilungssystem des Testkuratoriums der Föderation Deutscher Psychologenvereinigungen. Revidierte Fassung vom 09. September 2009. *Psychologische Rundschau, 61*, 52–56.

[2] Das TBS-TK ersetzt den von der Föderation Deutscher Psychologenverbände (1986) publizierten Kriterienkatalog für die Testbeurteilung.

Tabelle 1. Besprechungs- und Beurteilungskategorien.

Kategorien	Bewertung	max. Zeichenzahl (inkl. Leerzeichen) für die freie Bewertung
1. Allgemeine Informationen über den Test, Beschreibung des Tests und seiner diagnostischen Zielsetzung	frei und formalisiert*	1000
2. Theoretische Grundlagen als Ausgangspunkt der Testkonstruktion	frei	1000
3. Objektivität	frei und formalisiert*	1000
4. Normierung (Eichung)	frei	1000
5. Zuverlässigkeit (Reliabilität, Messgenauigkeit)	frei und formalisiert*	1000
6. Gültigkeit (Validität)	frei und formalisiert*, auch unter Berücksichtigung der Fairness (soweit in Anspruch genommen)	1000
7. Weitere Gütekriterien (Störanfälligkeit, Unverfälschbarkeit und Skalierung)	frei	1000
8. Abschlussbewertung/Empfehlung	frei	2000

Anmerkung: * Die formalisierte Bewertung soll auf einer vierstufigen Skala gemäß Tabelle 2 vorgenommen werden.

(inkl. Leerzeichen) vorgesehen. Für die Kategorien 1, 3, 5 und 6 erfolgt darüber hinaus eine formalisierte Bewertung auf einer vierstufigen Skala gemäß Tabelle 2.

Tabelle 2. Formalisierte Bewertungsskala

Der Test erfüllt die Anforderungen ...	voll	a
	weitgehend	b
	teilweise	c
	nicht	d

Die freie Abschlussbewertung/Empfehlung ergibt sich nicht „automatisch“ aus den formalisierten Einzelbewertungen. Vielmehr ist es Aufgabe der Rezensenten, in freier Würdigung der Gesamtheit aller Aspekte eine abschließende Wertung und Empfehlung abzugeben. Dabei ist der Test vor allem an den diagnostischen Zielsetzungen zu messen, die in den Verfahrenshinweisen (im Testmanual) formuliert sind.

Die Gesamtlänge der Bewertung darf 9000 Zeichen (inkl. Leerzeichen) nicht überschreiten. Neben den einschlägigen Bewertungsaspekten sollen in den einzelnen Beurteilungskategorien insbesondere auch spezielle Aspekte beachtet werden, die im Anhang aufgeführt sind. Auch in dem Fall, dass ein Test als „nicht prüffähig“ eingestuft wird, soll eine Rezension zu diesem Test erscheinen. Sie beschränkt sich allerdings darauf, das Urteil über die mangelhafte Prüffähigkeit transparent werden zu lassen.

5. Die Schritte 4.1 bis 4.3 werden von beiden Rezensenten unabhängig voneinander vorgenommen. Die Rezensenten senden ihre Ausarbeitungen zu 4.1 bis 4.3 innerhalb einer definierten Frist an das TK.
6. Nachdem die beiden Rezensionen eingegangen sind, hebt das TK die gegenseitige Anonymität der beiden Rezensenten auf und bittet beide um die Erstellung einer gemeinsamen Fassung der Rezensionen.
7. Sofern sich die beiden Rezensenten nicht auf eine in allen Punkten übereinstimmende gemeinsame Fassung einigen können, werden in der Rezension die relevanten Unterschiede der Positionen dargestellt, wobei das TK die Gesamtlänge der gemeinsamen Fassung bei Bedarf auf bis zu 12000 Zeichen erweitern kann. Sofern sich die Rezensenten hinsichtlich der Frage der Prüffähigkeit (Punkt 4.1) und/oder der formalisierten Bewertungen nicht einigen können, entscheidet das TK.
8. Das TK schickt die Rezension an den/die Testautor(en), um dem/den Testautor(en) Gelegenheit einzuräumen, innerhalb einer gesetzten Frist gegenüber dem TK Stellung zu beziehen. Im Falle einer solchen Stellungnahme entscheidet das TK, ob es die beiden Rezensenten bittet, aufgrund der Stellungnahme eine Modifikation der Testrezension vorzunehmen. Sofern eine vom TK erbetene Modifikation der Testrezension nicht rechtzeitig erfolgt oder die Modifikation nach Ansicht des TK die Stellungnahme des/der Testautor(en) nicht ausreichend berücksichtigt, behält sich das TK vor, seinerseits Anpassungen der Rezension vorzunehmen.
9. Die Testrezensionen des TK werden in den Fachzeitschriften „Report Psychologie“ und „Psychologische Rundschau“ veröffentlicht. Sofern die Testrezension in Kooperation mit einer anderen Fachzeitschrift erfolgt, ist es möglich, dass die Rezension zuerst von dieser Fachzeitschrift und danach im „Report Psychologie“ und in der „Psychologischen Rundschau“ veröffentlicht wird. Andere Medien können die Rezensionen als Nachdruck veröffentlichen. Dabei müssen die vier formalisierten Bewertungen in jedem Fall vollständig übernommen werden. Sofern in den Texten der Besprechungskategorien eine Informationsauswahl getroffen wird, ist sicherzustellen, dass kein irreführender Eindruck vom Gesamtbild entsteht.
10. Als Autoren der Rezension werden die Rezensenten in alphabetischer Folge oder in der von ihnen vereinbarten Reihenfolge genannt, es sei denn, ein Rezensent oder beide wollen anonym bleiben; in diesem Fall wird für jeden anonym bleibenden Rezensenten „N.N.“ aufgeführt. Sofern eine Rezension nicht nur von dem durch das TK benannten Rezensenten (Regelfall), sondern unter Mitwirkung eines Koautors verfasst wird, werden diese Autoren stets zusammenstehend aufgeführt.

11. Die unter 4.2 erarbeiteten klassifikatorischen Angaben werden im online verfügbaren Datenbanksegment von PSYNDEX, die Testrezension unter http://www.zpid.de/index.php?wahl=Testkuratorium veröffentlicht.
12. Die Rezensenten werden am Ende des Jahres in „Report Psychologie“ genannt. Die Nennung entfällt, wenn der Rezensent dem TK mitteilt, dass er anonym bleiben will.
13. Das TK evaluiert in regelmäßigen Abständen das hier dargestellte System und nimmt ggf. Modifikationen vor. Die Rezensenten werden explizit aufgefordert, an der kontinuierlichen Verbesserung des Systems mitzuwirken, indem sie z. B. Streichungs- und/oder Ergänzungsvorschläge zu den Beurteilungsrichtlinien einbringen.
14. Das TK evaluiert in regelmäßigen Abständen die Ergebnisse der Beurteilungen unter testübergreifenden Gesichtspunkten und publiziert die so gewonnenen Erkenntnisse (vgl. Evers, 2001).
15. Das TK dokumentiert alle nach dem vorliegenden System erstellten Testbeurteilungen und gewährleistet den Zugriff auf die Testrezensionen. Darüber hinaus bemüht sich das TK um die Verbreitung der Rezensionen.

Literatur

Evers, A. (2001). Improving Test Quality in the Netherlands: Results of 18 years of Test Ratings. *International Journal of Testing*, *1*, 137–153.

Föderation Deutscher Psychologenverbände (1986). Beschreibung der einzelnen Kriterien für die Testbewertung. *Diagnostica*, *32*, 358–360.

Kersting, M. (2008). DIN Screen, Version 2. Leitfaden zur Kontrolle und Optimierung der Qualität von Verfahren und deren Einsatz bei beruflichen Eignungsbeurteilungen. In M. Kersting, *Qualität in der Diagnostik und Personalauswahl – der DIN Ansatz* (S. 141–210). Göttingen: Hogrefe.

DOI: 10.1026/0033-3042/a000013

Anhang: Richtlinien des Testkuratoriums für die Beurteilung von Tests zur Erfassung menschlichen Erlebens und Verhaltens

Neben den einschlägigen Bewertungsaspekten sollen in den einzelnen Besprechungs- und Beurteilungskategorien insbesondere auch spezielle Aspekte beachtet werden, die im Folgenden aufgeführt sind.

Zu 1: Allgemeine Information über den Test durch die Verfahrenshinweise und Beschreibung des Tests und seiner diagnostischen Zielsetzung

DIN Screen Abschnitt 1.1 (Aussagen 1–14), 1.2 (Aussagen 15 bis 21 sowie 26–33), 1.4 (Aussagen 45–52).

- Zugänglichkeit von Informationen/Informationspolitik
- Informationsgehalt der Darstellung von empirischen Untersuchungen
- Diagnostische Zielstellung
  - o Altersgruppen
  - o Einschränkungen der Anwendbarkeit
- Testaufbau (Zahl der Items, Subskalen, Itembeispiele, Beantwortungsmodus, Testformen)
- Auswertung und Interpretation (Vorgehen bei der Auswertung [ggf. Schablonen, Auswertungsprogramme], Vergabe von Punktwerten für eine Antwort, Berechnung von Skalen und/oder Gesamtwerten, gegebenenfalls. Umrechnung in Normwerte), (Interpretationshilfen wie Cut-off-Werte, Normen, Vertrauensgrenzen, kritische Differenzen)
- Bei adaptiven Tests müssen die Entscheidungsregeln explizit festgelegt sein, wie die Auswahl jedes folgenden Items getroffen wird.
- Zeiten (Durchführung, Auswertung)
- Durchführungsvoraussetzungen (Qualifikation der Testleiter(innen))

Zu 2: Theoretische Grundlagen als Ausgangspunkt der Testkonstruktion

DIN Screen Abschnitt 1.2 (Aussage 22), 1.5 (Aussagen 53–55)

In dieser Kategorie geht es um die Frage, ob der theoretische Hintergrund beschrieben ist; es geht nicht um die Qualität der Untersuchungsdesigns und der Untersuchungsausführung. Mögliche Besprechungspunkte sind:

- Schließt der Test an eine bestehende Theorie an oder entwickelt der Autor eine eigene Theorie?
- Wird diese Theorie ausreichend beschrieben? Wird das Konstrukt hinlänglich beschrieben?
- Wird deutlich, was und was nicht zu dem zu messenden Bereich gerechnet wird?
- Wird beschrieben, was die Unterschiede und Gemeinsamkeiten gegenüber Tests mit überlappendem Geltungsanspruch sind?
- Wird angegeben, was auf theoretischer Ebene/auf der Ebene des Aufgabenmaterials der Mehrwert des neuen Instruments über bestehende Instrumente hinaus ist?
- Wird deutlich, ob ein beliebiges Item zum Test gehören könnte oder nicht?
- Werden das oder die zu messenden Konstrukt(e) auf solche Weise (z. B. mithilfe von Facetten-Analyse) analysiert, dass deutlich wird, welche Aspekte innerhalb des Konstrukts oder der Konstrukte unterschieden werden können?

Zu 3: Objektivität

DIN Screen Abschnitt 1.2 (Aussage 23–25), 1.3 (Aussagen 34–44), 1.7 (Aussagen 61–67).

Hinsichtlich der *Durchführungsobjektivität* soll auch auf folgende Punkte geachtet werden:

- Der Test muss so weit wie möglich standardisiert sein
- Die Instruktionen für die Testleiter(innen) müssen
  - o möglichst wörtlich vorschreiben, was der Testleiter sagen soll und was nicht (so ist z. B. die Empfehlung „der Testleiter erklärt das Ziel des Tests“ als mangelhaft zu werten)
  - o genau angeben, welche Handlungen der Testleiter konkret zu verrichten hat (z. B. das Testmaterial in einer bestimmten Art ordnen)
  - o genau ausführen, wie auf Fragen eingegangen werden muss (es können z. B. Standardtexte gegeben werden für Antworten auf häufig vorkommende Fragen)

- Die Instruktionen für die getesteten Personen sollten Beispiel- und Übungsitems enthalten sowie Informationen über die Art, wie die Reaktionen (Antworten) zu geben sind.

Hinsichtlich der *Auswertungsobjektivität* soll auch auf die folgenden Punkte geachtet werden:

- Falls Auswertungsschablonen gebraucht werden, muss genau angegeben sein, wie diese auf die Antwortformulare zu legen sind.
- Falls Auswertungsschablonen benutzt werden, muss auf den Schablonen angegeben sein, zu welcher Version des Tests sie gehören (dies ist besonders von Bedeutung, wenn der Test in veränderter Auflage vorliegt).
- Es muss angegeben sein, welcher Testwert für ein nicht bearbeitetes Item gegeben werden soll bzw. wie mit nicht bearbeiteten Items umzugehen ist.
- Es muss angegeben sein, bis zu welcher Anzahl von nicht bearbeiteten Items das Testergebnis noch interpretiert werden darf.
- Falls der Test den Einsatz mehrerer Beurteiler/Beobachter erfordert, muss angegeben sein, wie mit unterschiedlichen Urteilen/Beobachtungen umzugehen ist.
- Bei Tests, die am Computer durchgeführt und ausgewertet werden, müssen die Anwender die Auswertung kontrollieren können.
- Auch für Tests, die definitionsgemäß weniger objektiv sind, z. B. Projektive Verfahren, müssen Prozeduren beschrieben sein, durch die die Objektivität so gut wie eben möglich gewährleistet wird.

Hinsichtlich der *Interpretationsobjektivität* soll auch auf die folgenden Punkte geachtet werden:

- Wurden einzelne Fallbeschreibungen in die Verfahrenshinweise (das Testmanual) aufgenommen?
- Wird bei der beispielhaften Interpretation von Testergebnissen darauf eingegangen, welchen möglichen Einfluss bestimmte Hintergrundvariablen und (Test-)Erfahrung auf die Testwerte haben können?
- Wird das Ausmaß an Sachkunde angegeben, das nötig ist, um den Test zu interpretieren?

Zu 4. Normierung (Eichung)

DIN Screen Abschnitt 1.6 (Aussagen 56–60).

Von einschlägig bekannten Aspekten abgesehen soll auf Folgendes geachtet werden:

- Zu prüfen ist, wenn die diagnostische Zielstellung (vgl. unter 1) bei der Interpretation der Testwerte Normen (Eichtabellen) nötig macht, ob tatsächlich für jedes genannte diagnostische Ziel Normen (Eichtabellen) zur Verfügung stehen.
- Die Eichstichprobe muss repräsentativ sein für jede angestrebte (Sub-)Population. Die Rezensenten sollen prüfen, ob die Repräsentativität für die Zielgruppen nachvollziehbar dargestellt ist. Dabei geht es um eine angemessene Beschreibung sowohl der Population als auch der Art der Stichprobenziehung oder Datensammlung.
- Des Weiteren geht es darum, ob bei der Datensammlung bloß von einer „anfallenden Stichprobe“ Gebrauch gemacht wurde. Beispielsweise werden oft nur Schüler mit Schwierigkeiten bei der Berufswahl in die Stichprobe aufgenommen, die sich ohnehin freiwillig für eine Beratung und Testung interessieren, oder es werden Daten von Studierenden verwendet, weil diese leicht verfügbar sind.
- Im Fall altersspezifischer oder in anderer Hinsicht spezifischer Normen (Eichtabellen) sollen die Rezensenten beurteilen, ob die Altersintervallbreite und die betreffende Größe der jeweiligen Eichstichprobe in entsprechender Relation stehen.
- Bei der Beurteilung der Angemessenheit der Größe von Eichstichproben ist der Messfehler zu berücksichtigen.
- Beim Umrechnen von Rohwerten in geeichte Testwerte sollen die Rezensenten beurteilen, ob die verwendete Skala (z. B. *T*-Werte) in ihrer Differenziertheit dem in den Verfahrenshinweisen (im Testmanual) formulierten Anspruch zur Differenzierungsfähigkeit des Tests entspricht. Die Wahl der Skala muss auch der Sachkunde des hauptsächlich vorgesehenen Anwenderkreises entsprechen.

Zu 5. Zuverlässigkeit (Reliabilität/Messgenauigkeit)

DIN Screen Abschnitt 1.8 (Aussagen 68–76) und 1.13 (Aussagen 133–136).

Bei der Bewertung der Reliabilität (Messgenauigkeit) sind auch die folgenden Umstände mit zu berücksichtigen:

- Zu prüfen ist, ob die jeweiligen Reliabilitätskennwerte für diejenige (Sub-)Population aus einer Stichprobenerhebung geschätzt wurden, für die der Test laut diagnostischer Zielsetzung eingesetzt werden soll.
- Je nach der Art der Varianzquelle, die in der Reliabilitäts-Untersuchung analysiert wird, können verschiedene Reliabilitätsarten unterschieden werden, so dass die Rezensenten darauf eingehen sollten.
- Zu berücksichtigen ist auch, dass die Reliabilitätswerte in Abhängigkeit von den untersuchten Gruppen variieren (eine besondere Bedeutung kommt der Homogenität der Gruppe hinsichtlich des gemessenen Konstrukts zu).
- Darüber hinaus ist zu prüfen, ob eine sehr hohe interne Konsistenz auf nahezu identisch gestaltete Items zurückzuführen ist.
- Zu prüfen ist auch, ob die Messgenauigkeit bei Tests mit einer Speed-Komponente, bei denen also nicht alle Testpersonen auch zur Bearbeitung der letzten Items kommen, zweckmäßiger Weise nicht nach der internen Konsistenz oder mit anderen Homogenitätsmaßen bestimmt worden ist, weil diese die Höhe der Reliabilitätskoeffizienten überschätzen.
- Zu beurteilen ist des Weiteren, ob im Fall von Retest-Reliabilitäten das Intervall zwischen Test und Retest angemessen ist. Werden zu große Intervalle gewählt, weisen geringe Retest-Reliabilitäten nicht zwingend auf eine geringe Messgenauigkeit hin; sie können auch auf eine geringe Merkmalsstabilität zurückführbar sein.
- Bei Tests, die nach der Item-Response-Theorie (IRT) erstellt worden sind, d. h. vor allem nach dem Rasch-Modell, ist zu beachten, ob die Standardschätzfehler im Manual angeführt werden.

Da es bei Tests eventuell Angaben zu mehreren Reliabilitätsarten gibt und da bei Tests mit mehreren Untertests/Skalen entsprechend mehrere Reliabilitätswerte vorliegen, führen die Rezensenten die Vielzahl der Informationen zu einem Gesamturteil zur Reliabilität zusammen. Dabei sind vor allem die Reliabilitäten derjenigen Untertests/Skalen zu berücksichtigen, die laut der diagnostischen Zielsetzung (Abschnitt 1) besonders wichtig sind.

Zu 6. Gültigkeit (Validität)

DIN Screen Abschnitt 1.9 (Aussagen 77–132) und 1.14 (Aussagen 137–140).

Grundsätzlich geht es nicht um die Validität eines Tests, sondern um die Validität der Interpretation der Ergebnisse, die mit dem Test gewonnen werden.

Bei der Bewertung der Validität sind auch die folgenden Umstände mit zu berücksichtigen:

- Die Rezensenten berücksichtigen bei ihrem Urteil über die Angaben zur Validität des Tests, dass die Validitätswerte in Abhängigkeit von den untersuchten Gruppen (eine besondere Bedeutung kommt der Homogenität der Gruppe hinsichtlich des gemessenen Konstrukts zu) und in Abhängigkeit vom Untersuchungsdesign variieren.
- Es ist zu prüfen, ob die Validitätskoeffizienten für diejenige (Sub-)Population aus einer Stichprobenerhebung geschätzt wurden, für die der Test laut diagnostischer Zielsetzung eingesetzt werden soll.
- Im Fall, dass die Validitätsbefunde auf Mittelwertsvergleichen beruhen (etwa bei einer Extremgruppenvalidierung), soll der Effekt des Mittelwertsunterschieds von den Rezensenten als inhaltlich relevant oder irrelevant bewertet werden.
- Die Rezensenten führen die häufig gegebene Vielzahl von Informationen zur Validität (z. B. Kriteriums- und Konstruktvalidität) eines Tests zu einem Gesamturteil über die Validität zusammen.
- Welche Art der Validitätsbestimmung sinnvoll ist und welche Ausprägung der Validität notwendig ist, hängt von der diagnostischen Zielsetzung ab.
- Es ist zu überprüfen, dass die Validitätsuntersuchungen hypothesen- bzw. theoriegeleitet entwickelt wurden und nicht nur im Nachhinein signifikante Korrelationen als Validitätsbeleg angeführt werden.
- Des Weiteren ist die inhaltliche und psychometrische Qualität der zur Validierung herangezogenen Maße (z. B. andere Tests zur Konstruktvalidität; Kriteriumsmaße) von den Rezensenten zu beurteilen.
- Wenn Übereinstimmungsvaliditäten mit gleichartigen Tests angeführt werden, soll in die Beurteilung mit einfließen, inwieweit die konkurrierenden Tests selbst das Gütekriterium der Validität erfüllen.
- Die Rezensenten sollen prüfen, ob die Untersuchung zur Kriteriumsvalidität unter solchen Testbedingungen stattgefunden hat, wie sie den Bedingungen bei der Nutzung des Tests in der Praxis weitgehend entsprechen.

Zu beurteilen ist insbesondere die Art und die Qualität des Kriteriums. Es geht z. B. darum, ob Ausbildungs- oder Berufsleistungen herangezogen wurden, unter welchen Rahmenbedingungen die Kriteriumsleistungen gemessen wurden und ob spezifische Verhaltensweisen oder allgemeine, durchschnittliche oder Maximalleistungen das Kriterium ausmachen. Des Weiteren ist die psychometrische Qualität des Kriteriums (z. B. Reliabilität) zu beurteilen sowie die inhaltliche Qualität (z. B. inhaltliche Gültigkeit/ Relevanz). Zu bewerten ist schließlich die Art der Beziehung zwischen Prädiktor und Kriterium (z. B. linear/non-linear) sowie die Art der Analyse dieser Beziehungen (z. B. einfache oder multiple Regression; Kreuzvalidierung; Miteinbeziehung von Moderator- und Suppressor-Variablen).

Falls in den Verfahrenshinweisen (im Testmanual) eine Validitätsgeneralisierung in Anspruch genommen wird, soll geprüft werden, ob die Situationen und/oder Tests, für die die Generalisierbarkeit in Anspruch genommen wird, mit den Bedingungen der intendierten Nutzung des Tests übereinstimmen.

Zu 7: Weitere Gütekriterien (Störanfälligkeit, Unverfälschbarkeit und Skalierung)

Die Rezensenten berücksichtigen, in welchem Ausmaß der Test empfindlich ist gegenüber aktuellen Zuständen der Testperson und situativen Faktoren der Umgebung („Störanfälligkeit"); insbesondere soll geprüft werden, ob eine solche Störanfälligkeit angesichts der diagnostischen Zielstellung ein Problem darstellt.

Die Rezensenten beurteilen, inwieweit es beim gegebenen Test möglich ist, dass die Testperson durch ein gezieltes Testverhalten die konkrete Ausprägung ihres Testwerts steuern bzw. kontrollieren kann („Verfälschbarkeit"). Je nach diagnostischer Zielsetzung ist dabei darauf zu achten, inwieweit ein Faking-good, ein Faking-bad oder auch beides möglich ist und – falls ja – ob diese Verfälschungen angesichts der diagnostischen Zielstellung ein Problem darstellen.

Insbesondere die IRT, d. h. vor allem das Rasch-Modell, bringt es mit sich, dass bei Tests auch kritisch hinterfragt wird, inwieweit die Zahlenrelationen der Testwerte mit den Relationen der beobachtbaren Verhaltensweisen – sowohl innerhalb ein und derselben Testperson als auch zwischen verschiedenen Testpersonen – übereinstimmen („Skalierung"). Da eine entsprechende empirische Absicherung durch den Testautor eben nur durch den Einsatz der Modelle der IRT möglich ist, sollten die Rezensenten nicht nur eine gegebenenfalls versuchte Absicherung dieser Art beurteilen, sondern auch im Fall, dass die Testkonstruktion nicht nach diesem Modell erfolgte, wenigstens anführen, inwieweit in den Verfahrenshinweisen (im Testmanual) die Frage aufgegriffen und diskutiert wird, ob die laut Verrechnungsvorschriften resultierenden Testwerte die empirischen Verhaltensrelationen adäquat abbilden.

Testkuratorium der Föderation Deutscher Psychologenvereinigungen (2009). TBS-TK. Testbeurteilungssystem des Testkuratoriums der Föderation Deutscher Psychologenvereinigungen. Revidierte Fassung vom 09. September 2009. Report Psychologie, 34, 470-478 sowie Psychologische Rundschau, 2010, 61, 52-56. (Erste Fassung: Report Psychologie, 2006, 31, 492-499. sowie Psychologische Rundschau, 2007, 58, 25-30. (Das TBS-TK wurde 2006 verfasst vom Testkuratorium der Föderation Deutscher Psychologenvereinigungen C. Hagemeister, Gast; L. Hellfritsch, BDP; L.F. Hornke, DGPs; M. Kersting, Gast; K. Kubinger, DGPs; F. Lang, BDP; H. Moosbrugger, DGPs; G. Reimann, BDP & K. Westhoff, DGPs).