Nachrichten

Föderation Deutscher Psychologenvereinigungen

DGPs Deutsche Gesellschaft für Psychologie

Berufsverband Deutscher Psychologinnen und Psychologen

Diagnostik- und Testkuratorium. (2018). TBS-DTK. Testbeurteilungssystem des Diagnostik- und Testkuratoriums der Föderation Deutscher Psychologenvereinigungen. Revidierte Fassung vom 03. Jan. 2018. Psychologische Rundschau, 69, 109– 116. (Unter Mitarbeit von Martin Kersting.)

TBS-DTK – Testbeurteilungssystem des Diagnostik- und Testkuratoriums der Föderation Deutscher Psychologen-vereinigungen. Revidierte Fassung vom 03. Januar 2018

Das TBS-DTK wurde vom Diagnostik- und Testkuratorium (DTK) erstellt. Mitglieder des Diagnostik- und Testkuratoriums zum Zeitpunkt der Erstellung der Version von 2018 waren: Dr. Tom Frenzel, Prof. Dr. Carmen Hagemeister, Prof. Dr. Nina Heinrichs, Prof. Dr. Martin Kersting (Vorsitzender), Dipl.-Psych. Fredi Lang, Prof. Dr. Matthias Ziegler. An den vorherigen Fassungen haben zusätzlich zu den genannten Personen die folgenden ehemaligen Mitglieder des Testkuratoriums mitgewirkt: Prof. Dr. Markus Bühner, Dipl.-Psych. Lothar Hellfritsch, Prof. Dr. Lutz Hornke, Prof. Dr. Klaus Kubinger, Prof. Dr. Helfried Moosbrugger, Prof. Dr. Karl Westhoff.

Das TBS-TK ist in folgender Weise zu zitieren:

Diagnostik- und Testkuratorium. (2018). TBS-DTK. Testbeurteilungssystem des Diagnostik- und Testkuratoriums der Föderation Deutscher Psychologenvereinigungen. Revidierte Fassung vom 03. Jan. 2018. *Psychologische Rundschau, 69,* 109–116.

Ziel

Das TBS-DTK dient Testautor_innen, Verlagen und Testanbieter_innen sowie Testnutzer_innen zur Qualitätsbeurteilung, -sicherung und -optimierung von Tests.

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: Damit sind messtheoretisch fundierte Fragebogen (z. B. Persönlichkeitsfragebogen, Interessenfragebogen) und messtheoretisch fundierte Tests (z. B. Intelligenz- und Wissenstests) gemeint.

Durchführung

1. Die Auswahl der zu rezensierenden Tests erfolgt durch das Diagnostik- und Testkuratorium (DTK). Vorschläge für zu rezensierende Tests nimmt die / der Vorsitzende des DTK entgegen.
2. Mit der Beurteilung der ausgewählten Tests werden vom DTK zwei „Rezensions-Parteien“ beauftragt. Eine „Partei“ kann aus mehreren Personen bestehen, von denen mindestens eine Person promoviert sein sollte. Das DTK bürgt für die Qualifikation, Fachexpertise, Unabhängigkeit und Unvoreingenommenheit der Rezensent_innen (hier sind alle Einzelpersonen gemeint). Die Rezensent_innen (hier sind alle Einzelpersonen gemeint) geben zudem eine Selbsterklärung zu ihrer Unvoreingenommenheit ab.
   Sofern ein Verfahren rezensiert wird, zu dem bereits eine TBS-TK-Rezension zu einer früheren Fassung vorliegt, sollen nach Möglichkeit die bisherigen Rezensions-Parteien auch für die Rezension der neuen Version gewonnen werden. Gelingt dies nicht, soll zumindest eine der beiden bisherigen Rezensions-Parteien gewonnen

werden, die dann um eine neue Rezensions-Partei ergänzt wird. Gelingt dies nicht, werden zwei neue Rezensions-Parteien gewonnen.

3. Das DTK sorgt dafür, dass den Rezensions-Parteien sowie dem DTK der für die Beurteilung notwendige Test sowie die dazu gehörenden Verfahrenshinweise (auch Testmanual oder Testhandbuch genannt) von den Testanbieter_innen zur Verfügung gestellt werden. Im Falle von „confidential tests" sichern die Rezensent_innen und das DTK den Testanbieter_innen die Vertraulichkeit, z. B. wettbewerbsrechtlicher Informationen zu. Werden dem DTK die Verfahrenshinweise zu einem Verfahren innerhalb einer Frist von drei Monaten nicht zur Verfügung gestellt, wertet das DTK das Verfahren als „nicht prüffähig" und publiziert eine Rezension mit einer Kurzbeschreibung des Verfahrens sowie der Wertung: „Das xy-Verfahren erfüllt die in den Richtlinien des Diagnostik- und Testkuratoriums für die Beurteilung von Tests zur Erfassung menschlichen Erlebens und Verhaltens festgelegten Anforderungen bezüglich Information und Dokumentation nicht."
4. Der Beurteilungsprozess verläuft in zwei Schritten:
   4.1 Prüfung des Informationsgehalts der Verfahrenshinweise.
   Zunächst prüfen die Rezensions-Parteien, ob und in welchem Ausmaß die in den „Richtlinien des Diagnostik- und Testkuratoriums für die Beurteilung von Tests zur Erfassung menschlichen Erlebens und Verhaltens" festgelegten Anforderungen bezüglich Information und Dokumentation erfüllt sind. Diese Anforderungen an Verfahrenshinweise wurden aus der DIN 33430 (2016) übernommen. Obwohl sich die DIN 33430 auf die berufsbezogene Eignungsdiagnostik bezieht, sind diese Anforderungen auf Tests aus allen Bereichen anwendbar. Die Operationalisierung dieser Anforderungen erfolgt mit der „DIN SCREEN Checkliste 1" (Kersting, 2018). Diese Checkliste dient als „Standard des Diagnostik- und Testkuratoriums der Föderation Deutscher Psychologenvereinigungen hinsichtlich des Qualitätsanspruches an Information und Dokumentation von Instrumenten zur Erfassung menschlichen Erlebens und Verhaltens" (kurz: DTK-Testinformationsstandard). Die Checkliste sollte bereits durch die Testanbieter_innen unter Angabe der Seiten in den Verfahrenshinweisen, auf denen sich die jeweiligen Informationen befinden, ausgefüllt sein. Die Rezensent_innen kontrollieren diese Angaben und korrigieren sie, wenn nötig. Auf Basis der vorliegenden Informationen stellen die Rezensions-Parteien unabhängig voneinander 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 „Das xy-Verfahren erfüllt die in den Richtlinien des Diagnostik- und Testkuratoriums für die Beurteilung von Tests zur Erfassung menschlichen Erlebens und Verhaltens festgelegten Anforderungen bezüglich Information und Dokumentation nicht".
   Testanbieter(innen) und / oder Testautor(inn)en können der / dem Vorsitzenden des DTK auf eigene Initiative hin die Verfahrenshinweise sowie die Tabelle senden, in der verzeichnet ist, auf welcher Seite oder in welchem Abschnitt der Verfahrenshinweise die Informationen, die laut dem „Standard zur Information und Dokumentation von Instrumenten zur Erfassung menschlichen Erlebens und Verhaltens des Diagnostik- und Testkuratoriums der Föderation Deutscher Psychologenvereinigungen" notwendigerweise vorliegen müssen, zu finden sind (Übersichts-Tabelle zur „DIN SCREEN Checkliste 1"). Sofern die Verfahrensanbieter_innen / Testautor_innen zusätzlich eine schriftliche und zur Veröffentlichung autorisierte Selbsterklärung abgeben, mit der sie bestätigen, dass es sich (1) um die Verfahrenshinweise handelt, die auch Anwender_innen zur Verfügung stehen und (2) alle nach dem genannten Standard geforderten Informationen zur Verfügung stehen, können die Testanbieter_innen und / oder Testautor_innen ein Zertifikat beantragen, über dessen Vergabe das DTK entscheidet. Das Zertifikat berechtigt die Testanbieter_innen und / oder Testautor_innen mit der folgende Aussage für ihr Verfahren zu werben: „Die Verfahrenshinweise zum Test (Bezeichnung) erfüllen den Qualitätsanspruch des Diagnostik- und Testkuratoriums an Information und Dokumentation von Instrumenten zur Erfassung menschlichen Erlebens und Verhaltens. Damit erfüllt der Test (Bezeichnung) nach Ansicht des DTK die Voraussetzungen, um einer Qualitätskontrolle unterzogen zu werden. Diese Qualitätskontrolle selbst hat das DTK für den Test (Bezeichnung) nicht vorgenommen".
   Das DTK behält sich jederzeit stichprobenartige Überprüfungen der Korrektheit der Angaben ebenso vor, wie die Veranlassung einer Rezension des Verfahrens auf Basis dieser Informationen. Auf der Basis des Ergebnisses dieser Überprüfung kann das Zertifikat entzogen werden. Die Testanbieter_innen und / oder Testautor_innen müssen in diesem Fall innerhalb von drei Monaten dafür Sorge tragen, dass sie nicht mehr mit dem Zertifikat für das Verfahren werben. Die Kosten für die notwendige Modifikation der Werbung tragen die Testanbieter_innen und / oder die Testautor_innen.
   4.2 Bewertung des Tests anhand der Besprechungs- und Beurteilungskategorien des DTK.
   Auf Basis der Angaben in den Verfahrenshinweisen wird eine Bewertung des Tests vorgenommen. Grundlage der Rezension sind die Verfahrenshinweise sowie solche Updates der Verfahrenshinweise, die den Test-

Tabelle 1. Besprechungs- und Beurteilungskategorien

Kategorien	Bewertung
1. Beschreibung des Tests und seiner diagnostischen Zielsetzung	frei
2. Bewertung des Informationsgehalts der Verfahrenshinweise	frei und formalisiert*
3. Prüfung, ob in den Verfahrenshinweisen verzeichnet ist, wo die nach dem DTK-Testinformationsstandard notwendigen Informationen zu finden sind	formalisiert*
4. Theoretische Grundlagen als Ausgangspunkt der Testkonstruktion	frei
5. Objektivität	frei und formalisiert*
6. Normierung (Eichung)	frei
7. Zuverlässigkeit (Reliabilität, Messgenauigkeit)	frei und formalisiert*
8. Gültigkeit (Validität)	frei und formalisiert*, auch unter Berücksichtigung der Fairness (soweit in Anspruch genommen)
9. Weitere Gütekriterien (Störanfälligkeit, Unverfälschbarkeit und Skalierung)	frei
10. Abschlussbewertung	frei

Anmerkung: * Die formalisierte Bewertung wird auf einer vierstufigen Skala gemäß Tabelle 2 vorgenommen. Ausnahme ist hier die Kategorie 3, bei der eine dichotome Bewertung (Ja / Nein) vorgenommen wird.

anwender_innen von den Testanbieter_innen zur Verfügung gestellt werden. Es ist nicht Aufgabe der Rezensent_innen, weitere Informationen zum Test heranzuziehen. Sofern weitere Quellen für die Rezension herangezogen werden, sind diese Quellen explizit zu benennen.

Die Rezension bezieht sich auf die zum Zeitpunkt des Beginns der Rezension aktuelle Version der Verfahrenshinweise. Änderungen gegenüber diesem Sachstand können nur dann berücksichtigt werden, wenn diese Änderungen sich auf die zu diesem Zeitpunkt aktuelle Version des Verfahrens beziehen und von Testanbieter_innen aktiv an die Zielgruppe der Testanwender_innen herangetragen wurden („Bringschuld" der Testanbieter_innen).

Die Beurteilung gliedert sich in zehn „Besprechungs- und Beurteilungskategorien" gemäß Tabelle 1. Die Bewertung erfolgt kriterienorientiert. Es wird kein Vergleich eines Verfahrens mit einem anderen Verfahren vorgenommen. Für die Kategorien sind gemäß Tabelle 1 freie und / oder formalisierte Bewertungen vorgesehen. Für die Kategorien 2, 5, 7 und 8 erfolgt darüber hinaus eine formalisierte Bewertung auf einer vierstufigen Skala gemäß Tabelle 2. Die Kategorie 3 wird dichotom mit „ja" oder „nein" bewertet.

Die freie Abschlussbewertung ergibt sich nicht „automatisch" aus den formalisierten Einzelbewertungen. Vielmehr ist es Aufgabe der Rezensions-Parteien – in freier Würdigung der Gesamtheit aller Aspekte – eine abschließende Wertung abzugeben. Dabei ist der Test vor allem an den diagnostischen Zielsetzungen zu messen, die in den Verfahrenshinweisen formuliert sind.

Tabelle 2. Formalisierte Bewertungsskala

Der Test erfüllt die Anforderungen ...	voll weitgehend teilweise nicht

Die Gesamtlänge der Bewertung darf 12.000 Zeichen (inkl. Leerzeichen) nicht überschreiten. Neben den einschlägigen Bewertungsaspekten sollen in den einzelnen Beurteilungskategorien insbesondere auch spezielle Aspekte beachtet werden, die im Anhang des vorliegenden Textes aufgeführt sind. Auch in dem Fall, dass ein Test als „nicht prüffähig" eingestuft wird, erscheint eine Rezension zu diesem Test. Sie beschränkt sich allerdings darauf, das Urteil über die mangelhafte Prüffähigkeit transparent werden zu lassen und darf im Umfang 6.000 Zeichen (inkl. Leerzeichen) nicht überschreiten.

5. Die Schritte 4.1 bis 4.2 werden von beiden Rezensions-Parteien unabhängig voneinander vorgenommen. Die Rezensions-Parteien senden ihre Ausarbeitungen zu 4.1 bis 4.2 innerhalb einer vereinbarten Frist an das DTK. Das DTK prüft, ob die Rezensions-Parteien die Richtlinien eingehalten haben und bittet anderenfalls die Rezensions-Parteien darum, die Testrezension zu modifizieren.
6. Sofern ein Verfahren rezensiert wird, zu dem bereits eine TBS-DTK Rezension zu einer früheren Fassung vorliegt und die bisherigen Rezensions-Parteien auch für die Rezension der neuen Version gewonnen wurden, entfällt die Phase, in der eine Rezension in Unkenntnis der anderen Rezensions-Partei erarbeitet wird.

7. Sobald von beiden Rezensions-Parteien Rezensionen vorliegen, die den Richtlinien genügen, hebt das DTK die gegenseitige Anonymität der Rezensions-Parteien auf und bittet beide Rezensions-Parteien um die Erstellung einer gemeinsamen Rezension.
8. Sofern sich die beiden Rezensions-Parteien nicht darauf einigen können, ob der Test prüffähig ist oder sich nicht auf eine in allen Punkten übereinstimmende gemeinsame Fassung einigen können, werden in der Rezension die relevanten Unterschiede der Positionen dargestellt, wobei das DTK die Gesamtlänge der gemeinsamen Fassung bei Bedarf auf bis zu 15.000 Zeichen erweitern kann. Über die formalen Bewertungen entscheidet in diesem Falle das DTK, wobei explizit zu kennzeichnen ist, dass die Beurteilungen in diesem Fall vom DTK und nicht von den Rezensions-Parteien vergeben wurden.
9. Das DTK prüft, ob die von beiden Rezensions-Parteien gemeinsam erstellte Testrezension richtliniengerecht erstellt wurde und bittet anderenfalls die Rezensions-Parteien darum, die gemeinsame Testrezension zu modifizieren.
10. Das DTK schickt die Rezension an den_die erstgenannten deutschsprachigen Testautor_in oder, sofern keine Testautor_innen ermittelt werden können, an die Testanbieter_innen, um den_die Testautor_innen / ersatzweise Testanbieter_innen Gelegenheit einzuräumen, innerhalb einer gesetzten Frist gegenüber dem DTK Stellung zu beziehen. Die Stellungnahme begrenzt sich darauf, dass der_die Testautor_innen / Testanbieter_innen die Gelegenheit erhalten, auf mögliche sachliche Fehler in dem Rezensions-Entwurf hinzuweisen. Im Falle einer solchen Stellungnahme entscheidet das DTK, ob es die beiden Rezensions-Parteien bittet, aufgrund der Stellungnahme die Testrezension zu modifizieren. Sofern eine vom DTK erbetene Modifikation der Testrezension nicht rechtzeitig erfolgt oder die Modifikation nach Ansicht des DTK die Stellungnahme der Testautor_innen nicht ausreichend berücksichtigt, behält sich das DTK vor, seinerseits Anpassungen der Rezension vorzunehmen. Dies wird entsprechend ausgewiesen.
11. Die Testrezensionen des DTK werden in den Fachzeitschriften „Report Psychologie“ und „Psychologische Rundschau“ sowie online veröffentlicht. Sofern die Testrezension in Kooperation mit einer anderen Fachzeitschrift erfolgt, wird die Rezension in dieser Fachzeitschrift sowie online veröffentlicht. Andere Medien können die Rezensionen als Nachdruck veröffentlichen. Dabei müssen die fünf 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.
12. Als Autor_innen der Rezension werden die Rezensent_innen (hier sind alle Einzelpersonen gemeint) in der von ihnen vereinbarten Reihenfolge genannt, es sei denn, ein_e oder mehrere Person_en wollen anonym bleiben; in diesem Fall wird für jede_n anonym bleibende_n Rezensentin und Rezensenten „N.N.“ aufgeführt.
13. Das DTK evaluiert in regelmäßigen Abständen das hier dargestellte System und nimmt ggf. Modifikationen vor. Die Rezensent_innen werden explizit aufgefordert, an der kontinuierlichen Verbesserung des Systems mitzuwirken, indem sie z. B. Streichungs- und / oder Ergänzungsvorschläge zu den Beurteilungsrichtlinien einbringen.
14. Das DTK dokumentiert alle nach dem vorliegenden System erstellten Testbeurteilungen und gewährleistet den Zugriff auf die Testrezensionen. Darüber hinaus bemüht sich das DTK um die Verbreitung der Rezensionen.

Literatur

DIN (2016). *DIN 33430: Anforderungen an berufsbezogene Eignungsdiagnostik*. Berlin: Beuth.

Kersting, M. (2018). Zur Information über und Dokumentation von Instrumenten zur Erfassung menschlichen Erlebens und Verhaltens – Die DIN SCREEN Checkliste 1, Version 3. In Diagnostik- und Testkuratorium (Hrsg.), *Personalauswahl kompetent gestalten: Grundlagen und Praxis der Eignungsdiagnostik nach DIN 33430* (S. 223 – 244). Berlin: Springer.

Kontakt: Diagnostik- und Testkuratorium, Vorsitz: Prof. Dr. Martin Kersting, Justus-Liebig-Universität Gießen, Fachbereich 06 Psychologie und Sportwissenschaft, Abteilung für Psychologische Diagnostik, Otto-Behaghel-Straße 10F, 35394 Gießen, martin.kersting@psychol.uni-giessen.de.

Anhang: Richtlinien des Diagnostik- und 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: Beschreibung des Tests und seiner diagnostischen Zielsetzung

DIN Screen Aussagen A1 bis A3 (V1), B1.

- Diagnostische Zielstellung
  - Einsatzzwecke
  - Altersgruppen
  - Einschränkungen der Anwendbarkeit
- Testaufbau (z. B. Zahl der Items, Subskalen, Beantwortungsmodus, Testformen)

Zu 2: Bewertung des Informationsgehalts der Verfahrenshinweise

DIN Screen Aussagen A4 (V2) bis A11, B3 bis B13.

- Zugänglichkeit von Informationen / Informationspolitik.
- Informationsgehalt der Darstellung aller empirischen Untersuchungen inklusive Stichprobenbeschreibung gemäß DIN Screen Checkliste.
- Durchführung und Durchführungsvoraussetzungen (z. B. Qualifikation der Testleiter(innen), relevante ethische und rechtliche Aspekte des vorgesehenen Testeinsatzes).
- 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 Entscheidungsregeln formuliert sein, die die Auswahl jedes folgenden Items festlegen.
- Zeiten (Durchführung, Auswertung).

Zu 3: Prüfung, ob in den Verfahrenshinweisen verzeichnet ist, wo die nach dem DTK-Testinformationsstandard notwendigen Informationen zu finden sind

Alle DIN Screen Aussagen (A1 bis B54).

Die Verfahrenshinweise sollen eine Tabelle enthalten, in der verzeichnet ist, auf welcher Seite oder in welchem Abschnitt der Verfahrenshinweise die Informationen, die laut dem „Standard zur Information und Dokumentation von Instrumenten zur Erfassung menschlichen Erlebens und Verhaltens des Diagnostik- und Testkuratoriums der Föderation Deutscher Psychologenvereinigungen“ notwendigerweise vorliegen müssen, zu finden sind (Übersichts-Tabelle zur „DIN SCREEN Checkliste 1“). Die Rezensent_innen prüfen, ob diese Tabelle vorhanden ist. Darüber hinaus werden Prüfungen zur Nachvollziehbarkeit und Plausibilität der Tabelleneinträge vorgenommen.

Zu 4: Theoretische Grundlagen als Ausgangspunkt der Testkonstruktion

DIN Screen: Aussagen B1, B2

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 entwickeln die Testautor(inn)en 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 messende(n) Konstrukt(e) auf solche Weise (z. B. mit Hilfe von Facetten-Analyse) analysiert, sodass deutlich wird, welche Aspekte innerhalb des Konstrukts oder der Konstrukte unterschieden werden können?

Zu 5: Objektivität

DIN Screen Aussagen A12 (V3) bis A15, B15 (V5), B20 (V9) bis B21

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

- Der Test muss so weit wie möglich standardisiert sein.
- Die Instruktionen für die Testleiter(innen) müssen
  - möglichst wörtlich vorschreiben, was die Testleiter (innen) sagen sollen und was nicht (so ist z. B. die Empfehlung „die Testleiter(innen) erklären das Ziel des Tests“ als mangelhaft zu werten),
  - genau angeben, welche Handlungen die Testleiter (innen) konkret zu verrichten haben (z. B. das Testmaterial in einer bestimmten Art ordnen),
  - genau ausführen, wie auf Fragen der Teilnehmer(innen) 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 insbesondere 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(innen) / Beobachter(innen) erfordert, muss angegeben sein, wie mit unterschiedlichen Urteilen / Beobachtungen umzugehen ist.
- Bei Tests, die am Computer durchgeführt und ausgewertet werden, müssen die Anwender(innen) die Auswertung vom Prinzip her nachvollziehen 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 insbesondere auf die folgenden Punkte geachtet werden:

- Wurden einzelne Fallbeschreibungen in die Verfahrenshinweise (das Testmanual) aufgenommen?
- Wurden, sofern unterschiedliche Normgruppen für die Interpretation angeboten werden, Hinweise gegeben, wie die Entscheidung, welche Normgruppe in welchem Fall heranzuziehen ist, zu treffen ist?
- Wird bei der beispielhaften Interpretation von Testergebnissen darauf eingegangen, welchen möglichen Einfluss bestimmte Hintergrundvariablen und (Test-) Erfahrung auf die Testwerte haben können bzw. wie mit möglichen Messfehlern umzugehen ist (z. B. Konfidenzintervalle oder kritische Differenzen)?
- Wird das Ausmaß an Sachkunde angegeben, das nötig ist, um den Test zu interpretieren?

Zu 6. Normierung (Eichung)

DIN Screen Aussagen B16 (V6) bis B19

Von einschlägig bekannten Aspekten abgesehen, soll auf Folgendes geachtet werden:

- Sofern die diagnostische Zielstellung (vgl. die Ausführungen zu 1) bei der Interpretation der Testwerte Normen (Eichtabellen) nötig macht, prüfen die Rezensent_innen, ob tatsächlich für jedes genannte diagnostische Ziel Normen (Eichtabellen) zur Verfügung stehen.
- Die Eichstichprobe muss repräsentativ sein für jede angestrebte (Sub-)Population. Die Rezensent_innen 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_innen 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, da diese leicht verfügbar sind.
- Im Fall altersspezifischer oder in anderer Hinsicht spezifischer Normen (Eichtabellen) beurteilen die Rezensent_innen, ob die Altersintervallbreite und die betreffende Größe der jeweiligen Eichstichprobe angemessen sind.
- Bei der Beurteilung der Angemessenheit der Größe von Eichstichproben ist der Messfehler zu berücksichtigen.
- Beim Umrechnen von Rohwerten in geeichte Testwerte beurteilen die Rezensent_innen, ob die verwendete Skala (z. B. *Z*-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 7. Zuverlässigkeit (Reliabilität / Messgenauigkeit)

DIN Screen Aussagen B22 bis B26

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

- Die Rezensent_innen prüfen, ob die jeweiligen Reliabilitätskennwerte für alle (Sub-)Populationen aus einer Stichprobenerhebung geschätzt wurden, für die der Test laut diagnostischer Zielsetzung eingesetzt werden soll. Hierbei sind die Einsatzzwecke zu berücksichtigen.
- Die Rezensent_innen prüfen, ob die jeweiligen Reliabilitätskennwert-Schätzungen inhaltlich angemessen sind. Die Bestimmung der internen Konsistenz ist beispielsweise keine angemessene Art der Zuverlässigkeitsbestimmung für Verfahren mit heterogenen Inhalten. Die Bestimmung der Retest-Reliabilität ist keine angemessene Art der Zuverlässigkeitsbestimmung für Verfahren zur Messung rasch veränderlicher Eignungsmerkmale (z. B. Stimmungen). Die Angemessenheit der für die Zuverlässigkeitsbestimmung genutzten Methode(n) sollte in den Verfahrenshinweisen (im Testmanual) erläutert werden. Bei der Begründung der Angemessenheit soll die Art der untersuchten Eignungsmerkmale und

der angestrebten Entscheidung ebenso berücksichtigt werden wie die jeweiligen Anwendungs- und Untersuchungsbedingungen.
- Sofern mit dem Verfahren Eignungsmerkmale erfasst werden, für die eine zumindest relative Zeit- und Situationsstabilität angenommen wird, sollte die Zuverlässigkeit (auch) über die Retest-Methode bestimmt oder die Retest-Reliabilität durch einen geeigneten Untersuchungsplan geschätzt werden.
- Die Rezensent_innen beurteilen 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.
- 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).
- Die Rezensent_innen prüfen darüber hinaus, ob eine sehr hohe interne Konsistenz auf nahezu identisch gestaltete Items zurückzuführen ist.
- Die Rezensent_innen prüfen 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.
- 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 Rezensent_innen 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. Zudem ist abzuwägen, ob die Schätzer den diagnostischen Zielsetzungen (z. B. Statusdiagnostik oder Prognose) genügen.

Zu 8. Gültigkeit (Validität)

DIN Screen Aussagen B27 bis B54

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 Rezensent_innen 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.
- Die Rezensent_innen prüfen, ob die Validitätskoeffizienten für alle (Sub-)Populationen 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 Mittelwertunterschieds von den Rezensent_innen als inhaltlich relevant oder irrelevant bewertet werden.
- Die Rezensent_innen 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. Die Rezensent_innen prüfen im Manual die hypothesengeleitete Prüfung von Validitätsbelegen zur Stützung des Testeinsatzes gemäß der diagnostischen Zielsetzung.
- Die Rezensent_innen überprüfen, ob 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 Rezensent_innen 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 Rezensent_innen 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.
- Die Rezensent_innen beurteilen 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, Sensitivität / Spezifität).
- 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.

Da es bei Tests eventuell Angaben zu den Ergebnissen mehrerer Validierungsuntersuchungen gibt und da bei Tests mit mehreren Untertests / Skalen entsprechend mehrere Validitätswerte vorliegen, führen die Rezensent_innen die Vielzahl der Informationen zu einem Gesamturteil zur Validität zusammen. Dabei sind vor allem die Validitäten der Interpretationen derjenigen Untertests / Skalen zu berücksichtigen, die laut der diagnostischen Zielsetzung (Abschnitt 1) besonders wichtig sind. Zudem ist abzuwägen, ob die Validierungsprüfungen den diagnostischen Zielsetzungen (z. B. Statusdiagnostik oder Prognose) genügen.

Zu 9: Weitere Gütekriterien (Störanfälligkeit, Unverfälschbarkeit und Skalierung)

DIN Screen Aussage B14 (V4)

Die Rezensent_innen 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 Rezensent_innen 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 Vehaltensweisen – sowohl innerhalb ein und derselben Testperson als auch zwischen verschiedenen Testpersonen – übereinstimmen („Skalierung"). Da eine entsprechende empirische Absicherung durch die Testautor_innen eben nur durch den Einsatz der Modelle der IRT möglich ist, sollten die Rezensent_innen 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.

https://doi.org/10.1026/0033-3042/a000401