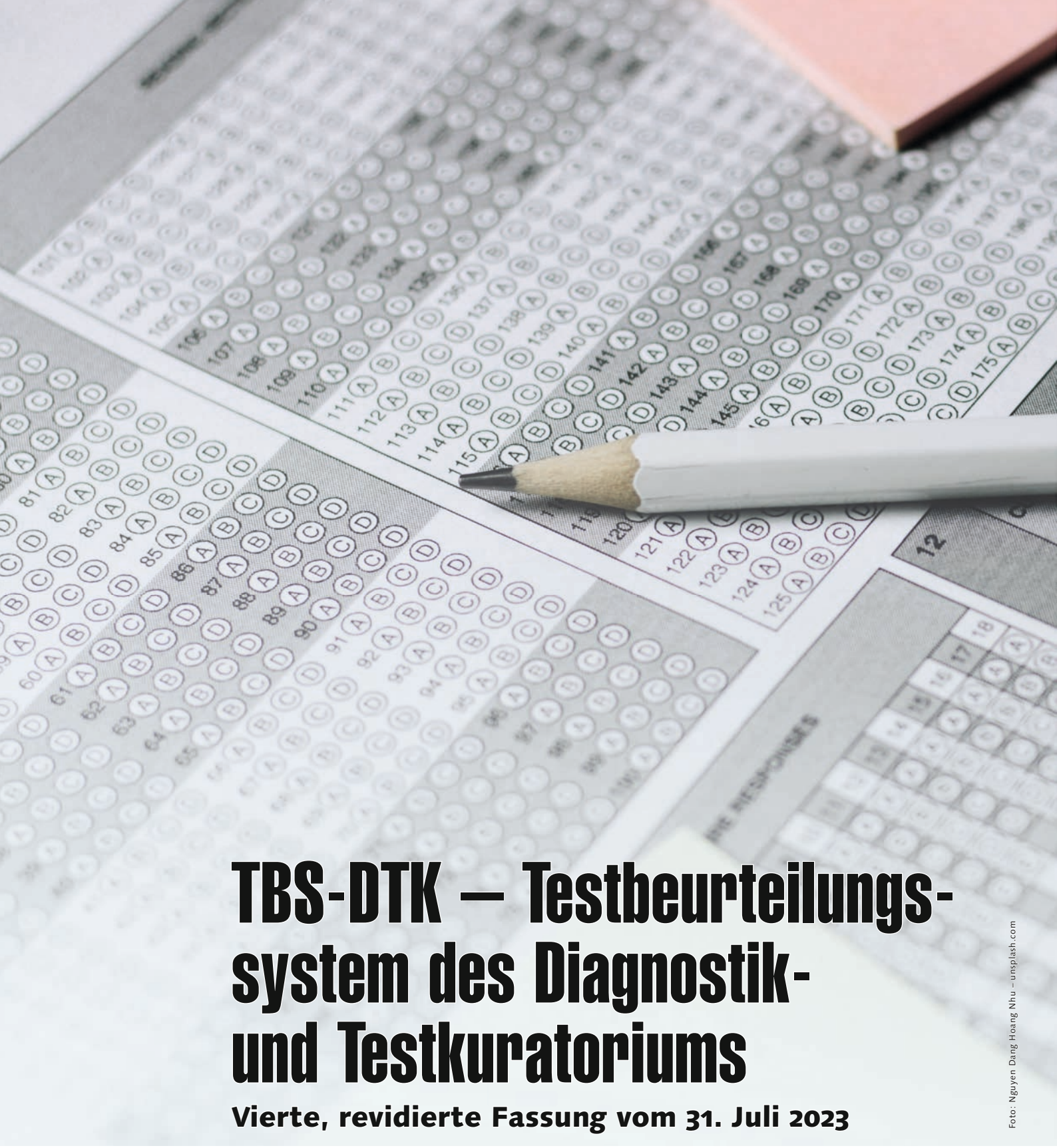

TBS-DTK – Testbeurteilungssystem des Diagnostik- und Testkuratoriums

Vierte, revidierte Fassung vom 31. Juli 2023

Foto: Nguyen Dang Hoang Nhu – unsplash.com

Diagnostik- und Testkuratorium (DTK) der Föderation Deutscher Psychologenvereinigungen[1]

1 Das TBS-DTK wurde vom DTK verfasst. Zum Zeitpunkt der Erstellung der Version von 2023 in das DTK berufen waren: Prof. Dr. Carmen Hagemeister, Prof. Dr. Martin Kersting (Vorsitzender), Dipl.-Psych. Fredi Lang, Prof. Dr. Nikola Stenzel, Dr. Kim-Oliver Tietze und Prof. Dr. Matthias Ziegler. Die KI betreffende Passagen der aktuellen Fassung wurden unter Mitarbeit von Prof. Dr. Clemens Stachl und Dr. Florian Pargent entwickelt. An den vorherigen Fassungen haben zudem folgende ehemalige Mitglieder des DTK mitgewirkt: Prof. Dr. Markus Bühner, Dr. Tom Frenzel, Dipl.-Psych. Lothar Hellfritsch, Prof. Dr. Nina Heinrichs, Prof. Dr. Lutz Hornke, Prof. Dr. Klaus Kubinger, Prof. Dr. Helfried Moosbrugger und Prof. Dr. Karl Westhoff. Das TBS-DTK ist wie folgt zu zitieren: Diagnostik- und Testkuratorium (2023). TBS-DTK. Testbeurteilungssystem des Diagnostik- und Testkuratoriums der Föderation Deutscher Psychologenvereinigungen. Vierte, revidierte Fassung vom 31. Juli 2023. *Report Psychologie*, 48(11+12), 18–27.

Teil 1: Prozedurale Richtlinien zur Erstellung von TBS-DTK-Testrezensionen sowie zur Ausstellung eines TBS-DTK-Transparenzzertifikats

1. Ziel

Das Testbeurteilungssystem des Diagnostik- und Testkuratoriums der Föderation Deutscher Psychologenvereinigungen (TBS-DTK) dient Testautorinnen und -autoren, Verlagen und Personen, die Tests anbieten oder nutzen, zur Qualitätsbeurteilung, -sicherung und -optimierung von Tests.

Das System umfasst zum einen inhaltliche Richtlinien zur Beurteilung von Tests (siehe unten, Teil 2) und zum anderen prozedurale Richtlinien für die Erstellung von Testrezensionen.

2. Geltungsbereich

Der Begriff »Test« hat in der Psychologie und in der Öffentlichkeit eine weit gefasste Bedeutung. Im Folgenden soll die Bezeichnung »Test« als Oberbegriff gelten: Damit sind messtheoretisch fundierte Fragebogen (z. B. Persönlichkeitsfragebogen, Interessenfragebogen) und messtheoretisch fundierte Tests (z. B. Intelligenz- und Wissenstests) gemeint. Dazu gehören auch Tests und Verfahren, die mittels Algorithmen (die z. B. mittels Machine Learning [ML] oder künstlicher Intelligenz [KI] erstellt wurden) Personenkennwerte schätzen.

TBS-DTK-Rezensionen werden zu Verfahren verfasst, die im deutschen Sprachraum verwendet werden.

Etwaige Gutachten bzw. Reporte, die aufgrund der Ergebnisse eines Tests (gegebenenfalls automatisch) erstellt werden, sind nicht Gegenstand der TBS-DTK-Testrezensionen.

3. Durchführung

3.1

Die Auswahl der zu rezensierenden Tests erfolgt durch das DTK. Vorschläge für zu rezensierende Tests nimmt die/der Vorsitzende des DTK entgegen.

3.2

Mit der Beurteilung der ausgewählten Tests beauftragt das DTK zwei »Rezensions-Parteien«. 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 Rezensierenden. Jede einzelne Rezensentin und jeder einzelne Rezensent muss eine Selbsterklärung zu ihrer bzw. seiner Unvoreingenommenheit abgeben.

Sofern ein Verfahren rezensiert wird, zu dem bereits eine TBS-DTK-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 oder eine einzelne rezensierende Person dieser Partei gewonnen werden, die dann um eine neue Rezensions-Partei ergänzt wird. Gelingt dies nicht, werden zwei neue Rezensions-Parteien gewonnen.

3.3

Das DTK sorgt dafür, dass den Rezensions-Parteien sowie dem DTK der für die Beurteilung notwendige Test sowie die dazugehörenden Verfahrenshinweise (auch Testmanual oder Testhandbuch genannt) von den Testanbietenden zur Verfügung gestellt werden. Im Falle von »confidential tests« sichern die Rezensierenden und das DTK den Testanbietenden bei Bedarf die Vertraulichkeit bestimmter (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 Verfahren [Bezeichnung] 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.«*

Sofern Testautorinnen und -autoren oder Testanbietende eines Tests, den das DTK für eine TBS-DTK Rezension ausgewählt hat, angeben, dass der zu rezensierende Test in Teilen oder in Gänze in Kürze modifiziert werden soll, so wird ihnen eine einmalige Frist von sechs Wochen bis zum Einreichen dieser modifizierten Version eingeräumt. Sollten später während des Rezensionsprozesses Modifikationen des zu rezensierenden Tests in Teilen oder in Gänze publiziert werden, wird die Rezension zu der Version verfasst, die zu Beginn der Rezension vorlag. In der Publikation dieser Rezension ist in diesem Fall auf die modifizierte Version hinzuweisen, sofern dies noch möglich ist.

3.4 Beurteilungsprozess

Der Beurteilungsprozess verläuft in zwei Schritten:

3.4.1 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 und durch ein Addendum u. a. um algorithmenspezifische Aspekte ergänzt. 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), die um ein Addendum ergänzt wurde. Hier und im Fol-

genden ist mit »Checkliste« stets die Kombination aus der »DIN-Screen-Checkliste 1« und dem Addendum gemeint. 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 soll Bestandteil der Verfahrenshinweise sein. Ist das nicht der Fall, soll die Checkliste anlässlich der Rezension durch die Testanbietenden oder Testautorinnen und -autoren ausgefüllt werden. In der Checkliste wird angegeben, an welcher Stelle in den Verfahrenshinweisen (Seite oder Abschnitt) sich die jeweiligen Informationen befinden. Die Rezensierenden 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 Verfahren [Bezeichnung] 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.«*

3.4.2 Bewertung des Tests anhand der Besprechungs- und Beurteilungskategorien des DTK

Auf Basis der Angaben in den Verfahrenshinweisen wird eine Bewertung des Tests nach den weiter unten in Teil 2 dargestellten Richtlinien des DTK für die Beurteilung von Tests vorgenommen. Grundlage der Rezension ist die zum Zeitpunkt des Beginns der Rezension aktuelle Version der Verfahrenshinweise. Weitere (gegebenenfalls aktuellere) Informationen müssen von den Rezensierenden nur dann berücksichtigt werden, wenn auch die Anwenderinnen und Anwender explizit auf diese Informationen hingewiesen werden und diese Informationen für alle Testanwenderinnen und -anwender zugänglich sind.

Für die Qualität von Tests ist es von entscheidender Bedeutung, dass alle relevanten Informationen zum Test und seiner Anwendung zentral und zugänglich zur Verfügung stehen und die Anwenderinnen und Anwender auf diese zentrale Informationsquelle hingewiesen werden. Entsprechend sind nur solche Informationen notwendige Grundlage der TBS-DTK-Rezension, die von den Testanbietenden bzw. Testautorinnen und -autoren allen Anwenderinnen und Anwendern zur Verfügung gestellt wurden. Für die Testanbietenden bzw. Testautorinnen und -autoren besteht hinsichtlich der relevanten Informationen zum Test und zu dessen Anwendung eine Bringschuld. Es ist nicht Aufgabe der Anwenderinnen und Anwender, nach solchen Informationen zu suchen.

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 eine formalisierte Bewertung auf einer vierstufigen Skala gemäß Tabelle 2. Eine Erläuterung der Skalenstufen findet sich in Tabelle 3.

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 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 (»DIN-Screen-Checkliste 1«)	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

** 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.*

Tabelle 2
Formalisierte Bewertungsskala

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

Die Gesamtlänge der Rezension darf 12.000 Zeichen (inklusive Leerzeichen) nicht überschreiten. Neben den einschlägigen Bewertungsaspekten sollen in den einzelnen Beurteilungskategorien insbesondere auch spezielle Aspekte beachtet werden, die im Teil 2 dieses 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 publik und transparent werden zu lassen, und darf im Umfang 6.000 Zeichen (inklusive Leerzeichen) nicht überschreiten.

3.5
Die Schritte 3.4.1 und 3.4.2 werden von beiden Rezensions-Parteien unabhängig voneinander vorgenommen. Die Rezensions-Parteien senden ihre Ausarbeitungen zu 3.4.1 und 3.4.2 innerhalb einer vereinbarten Frist an das DTK. Das DTK prüft, ob die Rezensions-Parteien die Richtlinien eingehalten haben, und bittet andernfalls die Rezensions-Parteien darum, die Testrezension zu modifizieren.

3.6
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.

3.7
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 Fall das DTK, wobei explizit zu kennzeichnen ist, dass die Beurteilungen in diesem Fall vom DTK und nicht von den Rezensions-Parteien vergeben wurden.

3.8
Das DTK prüft, ob die von beiden Rezensions-Parteien gemeinsam erstellte Testrezension richtliniengerecht erstellt wurde, und bittet andernfalls die Rezensions-Parteien darum, die gemeinsame Testrezension zu modifizieren.

3.9
Das DTK schickt die Rezension an die erstgenannte deutschsprachige Testautorin bzw. den erstgenannten deutschsprachigen Testautor oder, sofern keine Testautorinnen und -autoren ermittelt werden können, an die Testanbietenden, um den Testautorinnen und -autoren, ersatzweise den Testanbietenden, die Gelegenheit einzuräumen, innerhalb einer Frist von vier Wochen gegenüber dem DTK Stellung zu beziehen. Die Testautorinnen und -autoren bzw. Testanbietenden erhalten dabei lediglich die Gelegenheit, auf mögliche sachliche Fehler in dem Rezensions-Entwurf hinzuweisen. Keinesfalls geht es darum, ob die Testautorinnen und -autoren bzw. Testanbietenden mit der Rezension »einverstanden« sind oder sie Bewertungsänderungen wünschen.

Im Falle einer solchen Stellungnahme entscheidet das DTK, ob und welche Ausschnitte der Stellungnahme

Tabelle 3
Erläuterung der Bewertungsstufen

	Der Test erfüllt die Anforderungen ...			
	voll (höchste Qualität)	weitgehend (gering eingeschränkte Qualität)	teilweise (eingeschränkte Qualität)	nicht (erheblich eingeschränkte Qualität)
Informationsgehalt	Alle notwendigen Informationen sind verständlich im Manual enthalten.	Die Mehrzahl der notwendigen Informationen sind verständlich im Manual enthalten.	Die notwendigen Informationen sind mit Abstrichen im Manual enthalten.	Die notwendigen Informationen sind nicht ausreichend im Manual enthalten.
Objektivität	Es liegen mehrere, überzeugende Objektivitätsbelege für alle Objektivitätsarten vor.	Es liegen einige, zumeist überzeugende Objektivitätsbelege für die meisten Objektivitätsarten vor.	Es liegen einige, zumeist überzeugende Objektivitätsbelege für einige Objektivitätsarten vor.	Es liegen keine überzeugenden Objektivitätsbelege vor.
Zuverlässigkeit	Es liegen mehrere, überzeugende Reliabilitätsbelege für die relevanten Reliabilitätsschätzer vor.	Es liegen einige, zumeist überzeugende Reliabilitätsbelege für die relevanten Reliabilitätsschätzer vor.	Es liegen einige, zumeist überzeugende Reliabilitätsbelege für einige der relevanten Reliabilitätsschätzer vor.	Es liegen keine überzeugenden Reliabilitätsbelege vor.
Gültigkeit	Es liegen mehrere, überzeugende Validitätsbelege für alle relevanten Validitätsarten und Einsatzzwecke vor.	Es liegen einige, zumeist überzeugende Validitätsbelege für die meisten der relevanten Validitätsarten und Einsatzzwecke vor.	Es liegen einige, zumeist überzeugende Validitätsbelege für einige der relevanten Validitätsarten und Einsatzzwecke vor.	Es liegen keine überzeugenden Validitätsbelege vor.

es an die beiden Rezensions-Parteien weitergibt und ob es die beiden Rezensions-Parteien bittet, aufgrund der Stellungnahme die bisherige Testrezension zu modifizieren. Sofern eine vom DTK erbetene Modifikation der Testrezension nicht rechtzeitig erfolgt oder die Modifikation nach Ansicht des DTK die Stellungnahme der Testautorinnen und -autoren bzw. Testanbietenden nicht ausreichend berücksichtigt, behält sich das DTK vor, seinerseits Anpassungen der Rezension vorzunehmen. Dies wird entsprechend ausgewiesen.

4. Publikation

4.1
Die Testrezensionen des DTK werden in den Fachzeitschriften »Report Psychologie« und »Psychologische Rundschau« sowie online veröffentlicht (siehe www.bdp-verband.de/profession sowie www.psyndex.de/tests/testkuratorium).

Sofern die Testrezension in Kooperation mit einer anderen Fachzeitschrift erfolgt, wird die Rezension in dieser Fachzeitschrift sowie online veröffentlicht. Es ist nicht vorgesehen, dass irgendjemand (z. B. Zeitschriften-Herausgebende und das Lektorat des Verlags) den Text der TBS-DTK Rezension inhaltlich und/oder stilistisch verändert.

Andere Medien können die Rezension 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.

4.2
Als Autorinnen und Autoren der Rezension werden die Rezensierenden (hier sind alle Einzelpersonen gemeint) in der von ihnen vereinbarten Reihenfolge genannt, es sei denn, eine oder mehrere Person(en) wollen anonym bleiben; in diesem Fall wird für jede(n) anonym bleibende(n) Rezensierende(n) »N. N.« aufgeführt.

4.3
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.

5. Ausstellung eines TBS-DTK-Transparenzzertifikats
Testautorinnen und -autoren und/oder Testanbietende können für Verfahren ein TBS-DTK-Transparenzzertifikat beantragen. Dies gilt nicht für Tests und Verfahren, die mittels Algorithmen (die z. B. mittels ML oder KI erstellt wurden) Personenkennwerte schätzen.

Hierzu übersenden sie der/dem Vorsitzenden des DTK die Verfahrenshinweise zu ihrem Test sowie die Übersichtstabelle zur »DIN-Screen-Checkliste 1«. Das ist die Tabelle, in der verzeichnet ist, auf welcher Seite oder in welchem Abschnitt der Verfahrenshinweise die Informationen zu finden sind, 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. Zusätzlich müssen sie dem Antrag eine schriftliche und zur Veröffentlichung autorisierte Selbsterklärung beifügen, mit der sie bestätigen, dass (1) es sich um die Verfahrenshinweise handelt, die auch Anwenderinnen und Anwendern zur Verfügung stehen, und (2) alle nach dem genannten Standard geforderten Informationen zur Verfügung stehen.

Voraussetzung für die Beantragung eines TBS-DTK-Transparenzzertifikats ist, dass es Testautorinnen oder -autoren und/oder Testanbietende gibt, welche die Verfahrenshinweise autorisiert haben und für den Inhalt dieser Texte verantwortlich sind.

Auf der Grundlage der Unterlagen entscheidet das DTK über die Vergabe des Zertifikats. Das Zertifikat wird auf Wunsch auch auf Englisch ausgestellt.

Das Zertifikat berechtigt die Testautorinnen und -autoren und Testanbietenden, mit der folgenden Aussage für ihr Verfahren zu werben (siehe Kasten 1).

Kasten 1: Bedeutung des TBS-DTK-Transparenzzertifikats/Werbetext

»Die Verfahrenshinweise zum Test (Bezeichnung) erfüllen den Qualitätsanspruch des Diagnostik- und Testkuratoriums (DTK) 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 entscheidet auf der Basis der Selbsterklärung (ohne Prüfung der Unterlagen), behält sich aber jederzeit stichprobenartige Überprüfungen der Korrektheit der Angaben ebenso vor wie die Veranlassung einer Rezension des Verfahrens auf Basis dieser Informationen. Auf Grundlage des Ergebnisses dieser Überprüfung kann das Zertifikat entzogen werden. Die Testautorinnen und -autoren bzw. Testanbietenden 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 Testautorinnen und -autoren bzw. Testanbietenden, die das TBS-DTK-Transparenzzertifikat beantragt haben.

6. Evaluation und Optimierung
Das DTK evaluiert in regelmäßigen Abständen das hier dargestellte System und nimmt gegebenenfalls Modifikationen vor. Die Rezensierenden werden explizit dazu aufgefordert, an der kontinuierlichen Verbesserung des

Systems mitzuwirken, indem sie z. B. Streichungs- und/oder Ergänzungsvorschläge zu den Beurteilungsrichtlinien einbringen.

Teil 2: Richtlinien des DTK für die Beurteilung von Tests zur Erfassung menschlichen Erlebens und Verhaltens

Neben den einschlägigen Bewertungsaspekten sollen in den einzelnen Besprechungs- und Beurteilungskategorien insbesondere auch die folgenden Aspekte beachtet werden.

Zu 1: Beschreibung des Tests und seiner diagnostischen Zielsetzung

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

- Wird in den Verfahrenshinweisen nachvollziehbar erläutert, …
  - auf welche(s) Konstrukt(e) bzw. Verhalten der Test abzielt?
  - für welche Einsatzzwecke der Test gedacht ist?
  - für welche Zielpopulationen (z. B. Altersgruppen oder Berufsgruppen) der Test gedacht ist?
  - welche Einschränkungen der Anwendbarkeit es gibt?
  - wie der Test aufgebaut ist (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 sowie Addendum.

- Sind alle für die Testdurchführung, -auswertung und -interpretation notwendigen Informationen zugänglich?
- Werden die berichteten empirischen Untersuchungen inklusive der Stichprobenbeschreibung gemäß den Anforderungen der DIN-Screen-Checkliste informativ dargestellt?
- Wird über alle für die Qualität relevanten Aspekte der Testdurchführung und der Durchführungsvoraussetzungen (z. B. Qualifikation der Testleiterinnen bzw. Testleiter, relevante ethische und rechtliche Aspekte des vorgesehenen Testeinsatzes) informiert?
- Wird über alle für die Qualität relevanten Aspekte der Auswertung und Interpretation informiert? (Z. B. 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.)
- Handelt es sich um einen adaptiven Test? Falls ja: Sind die Entscheidungsregeln formuliert und dargestellt, die die Auswahl jedes folgenden Items festlegen?
- Ist aufgeführt, wie viel Zeit für die Testdurchführung sowie für die -auswertung benötigt wird?
- Handelt es sich um einen Test, der Algorithmen nutzt? Falls ja:
  - Liegen nachvollziehbare Informationen über die Entstehung und die Nutzung der eingesetzten Algorithmen vor? (Z. B. genutzte Prädiktoren, Zusammensetzung von Trainings- und Testdatensatz, Begründung des Analysemodells (z. B. Schätzverfahren linear vs. non-linear, mehrparametriges IRT vs. Raschmodell), Aktualisierungsmaßnahmen.)
  - Handelt es sich um ein »Blackbox-Verfahren«? Falls ja: Sind Analysen zu interpretierbarem Machine Learning bzw. Explainable AI dargestellt? (Z. B. worauf basieren Vorhersagen, welche Parameter gehen ein, Änderungen in welchen Wertebereichen der eingehenden Variablen verursachen welchen mittleren Effekt in der Vorhersage des Kriteriums?[2]
  - Sind die zur Schätzung von Personenkennwerten eingesetzten Modelle, Normdaten und statistischen Schätzmethoden so dargestellt, dass die Anwenderinnen und Anwender von Tests den logischen Aufbau der Auswertung der Tests nachvollziehen können?

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) sowie Addendum.

Sehen die Verfahrenshinweise die Tabelle vor, in der verzeichnet ist, auf welcher Seite oder in welchem Abschnitt der Verfahrenshinweise die Informationen zu finden sind, 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 (Übersichtstabelle zur »DIN-Screen-Checkliste 1«)? Falls ja: Sind die Einträge in der Tabelle nachvollziehbar und plausibel?

Zu 4: Theoretische Grundlagen als Ausgangspunkt der Testkonstruktion[3]

DIN-Screen-Aussagen B1, B2.

- Bilden die Testkennwerte eine bestehende Theorie ab oder beziehen sich die Testautorinnen und -autoren auf eine eigene Theorie?

2 Solche Informationen können z. B. in Form von Model Cards dargeboten werden, die darüber hinaus auch über die Qualität der Entscheidungsvorschläge informieren (siehe Validität). Sofern die Testanbietenden den Anwenderinnen und Anwendern darüber hinaus den Modellcode und/oder das fertig trainierte Modell für eigene Analysen zur Beurteilung der Angemessenheit des Modells (z. B. Verfahren des interpretierbaren Machine Learnings; Überprüfung von Modellbias und Modellfairness) für die geplante Anwendung zur Verfügung stellen, sollte dies in der Rezension positiv erwähnt werden.

3 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.

- Wird diese Theorie ausreichend beschrieben? Wird das Konstrukt bzw. werden die Konstrukte hinlänglich beschrieben?
- Wird deutlich, was und was nicht zu dem zu messenden Bereich gerechnet wird (Definition eines nomologischen Netzes)?
- Wird beschrieben, was die Unterschiede und Gemeinsamkeiten gegenüber Tests mit überlappendem Geltungsanspruch sind?
- Wird angegeben, was auf theoretischer Ebene bzw. auf der Ebene des Aufgabenmaterials der Mehrwert des vorliegenden Instruments über bisher existierende Instrumente hinaus ist?
- Wird deutlich, ob ein beliebiges Item zum Test gehören könnte oder nicht?
- Sofern Verfahren der KI für die Erstellung der diagnostischen Verfahren genutzt wurden (z. B. zur Item-Erstellung): Wird nachvollziehbar dokumentiert, in welcher Form KI für die Erstellung der diagnostischen Verfahren (z. B. zur Item-Erstellung) genutzt wurde?
- Ist dargestellt, wie sich die zur Lösung bzw. Beantwortung des Items notwendigen psychologischen Prozesse aus der Theorie bzw. der Konstruktdefinition ableiten lassen?
- Werden das oder die zu messende(n) Konstrukt(e) auf solche Weise (z. B. mit Hilfe von Facetten-Analysen) analysiert, dass deutlich wird, welche Aspekte innerhalb des Konstrukts oder der Konstrukte unterschieden werden können?
- Wird bei kriterienorientierten Tests auf die Bedeutung des Kriteriums für die durch den Test adressierten Fragestellungen eingegangen? Werden die Operationalisierung des Kriteriums und mögliche Kontamination oder Defizienz besprochen?

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:
- Ist der Test so weit wie möglich standardisiert?
- Wird der Test mündlich instruiert? Falls ja: Sind die Instruktionen für die Testleiterinnen und Testleiter so gestaltet, dass sie …
  - möglichst wörtlich vorschreiben, was die Testleiterinnen und Testleiter sagen sollen und was nicht (eine vage Empfehlung in der Art »die Testleiterinnen und Testleiter erklären das Ziel des Tests« – ohne weitere Konkretisierung – ist z. B. als mangelhaft zu werten)?
  - genau angeben, welche Handlungen die Testleiterinnen und Testleiter konkret zu verrichten haben (z. B. das Testmaterial in einer bestimmten Art ordnen)?
  - genau ausführen, wie auf Fragen der Getesteten eingegangen werden muss (es können z. B. Standardtexte für Antworten auf häufig vorkommende Fragen gegeben werden)?
- Enthalten die Instruktionen für die getesteten Personen Beispiel- und Übungs-Items sowie Informationen über die Art, wie die Reaktionen (Antworten) zu geben sind?

Hinsichtlich der *Auswertungsobjektivität* soll insbesondere auf die folgenden Punkte geachtet werden:
- Werden Auswertungsschablonen genutzt? Falls ja: Ist genau angegeben, …
  - wie diese auf die Antwortformulare zu legen sind?
  - zu welcher Version des Tests sie gehören? (Dies ist besonders von Bedeutung, wenn der Test in veränderter Auflage vorliegt.)
- Ist angegeben, welcher Testwert für ein nicht bearbeitetes Item gegeben werden soll bzw. wie mit nicht bearbeiteten Items bzw. Informationseinheiten umzugehen ist?
- Ist angegeben, bis zu welcher Anzahl von nicht bearbeiteten Items oder anderer Informationseinheiten das Testergebnis noch interpretiert werden darf?
- Sieht der Test den Einsatz mehrerer Beurteilenden bzw. Beobachtenden vor? Falls ja: Ist angegeben, wie mit unterschiedlichen Urteilen bzw. Beobachtungen umzugehen ist, um eine zu interpretierende Beurteilung zu erzeugen?

Hinsichtlich der *Interpretationsobjektivität* soll insbesondere auf die folgenden Punkte geachtet werden:
- Sind Fallbeschreibungen in die Verfahrenshinweise aufgenommen?
- Werden unterschiedliche Normgruppen für die Interpretation angeboten? Falls ja: 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, auch getrennt nach Normgruppen)?
- Wird das Ausmaß an Sachkunde angegeben, das nötig ist, um den Testkennwert bzw. die Testkennwerte zu interpretieren?

Zu 6: Anforderungen an die Normierung (Eichung) sowie an Trainingsdatensätze bei algorithmenbasierten Verfahren

DIN-Screen-Aussagen B16 (V6) bis B19 sowie Addendum.

- Wird aufgrund der diagnostischen Zielsetzung (vgl. die Ausführungen zu 1) für die Interpretation der Testwerte die Nutzung von Normen (Eichtabellen) angeboten? Falls ja: Stehen für jedes genannte diagnostische Ziel geeignete Normen (Eichtabellen) zur Verfügung?
- Ist die Eichstichprobe bzw. bei algorithmenbasierten Tests, bei denen die Algorithmen mit Hilfe von Trainingsdatensätzen gewonnen wurden, der Trainingsdatensatz, bezogen auf relevante Personen-

merkmale, repräsentativ für jede angestrebte (Sub-) Population?
- Ist die Repräsentativität für die Zielgruppen nachvollziehbar dargestellt?[4]
- Liegen altersspezifische oder in anderer Hinsicht spezifische Normen (Eichtabellen) vor? Falls ja: Sind die Altersintervallbreite und die betreffende Größe der jeweiligen Eichstichprobe angemessen?
- Ist die Größe von Eichstichproben bzw. Trainingsdatensätzen – auch unter Berücksichtigung des Messfehlers – angemessen?
- Entspricht die für die Umrechnung von Rohwerten in geeichte Testwerte verwendete Skala (z. B. Z-Werte) ...
  - in ihrer Differenziertheit dem in den Verfahrenshinweisen formulierten Anspruch zur Differenzierungsfähigkeit des Tests?
  - in ihrer Differenziertheit dem Range der Rohwerte (z. B. keine T-Werte, wenn die Rohwerte nur von 0 bis 10 gehen)?
  - der Sachkunde des hauptsächlich vorgesehenen Kreises von Anwenderinnen und Anwendern?
- Werden automatisch erstellte Normwerte (z. B. Faktor-Scores oder regressionsbasiertes Continuous Norming) verwendet? Falls ja: Ist dargestellt, auf welchen Analysen diese Normwerte beruhen und wie gut die Schätzgüte ist?

Zu 7: Zuverlässigkeit (Reliabilität/Messgenauigkeit)

DIN-Screen-Aussagen B22 bis B26 sowie Addendum.

- Wurden die jeweiligen Reliabilitätskennwerte für jede (Sub-)Population, für die der Test laut diagnostischer Zielsetzung eingesetzt werden soll, aus einer Stichprobenerhebung geschätzt?[5]
- Sind die jeweiligen Schätzungen der Reliabilitätskennwerte inhaltlich angemessen?[6]
- Werden die Angemessenheit der für die Zuverlässigkeitsbestimmung genutzten Methode(n) sowie das Zutreffen der jeweiligen Voraussetzungen in den Verfahrenshinweisen erläutert? Falls ja: Werden bei der Erläuterung der Angemessenheit die Art der untersuchten Merkmale und der angestrebten Entscheidung sowie die jeweiligen Anwendungs- und Untersuchungsbedingungen berücksichtigt?
- Werden mit den Testkennwerten Merkmale erfasst, für die eine zumindest relative Zeit- und Situationsstabilität angenommen wird, und wird der Test für Prognosezwecke vorgesehen? Falls ja: Wird die Zuverlässigkeit (auch) über die Retest-Methode bestimmt oder die Retest-Reliabilität durch einen geeigneten Untersuchungsplan geschätzt?
- Wurde die Retest-Methode genutzt? Falls ja: Ist das Intervall zwischen Test und Retest angemessen[7]?

Bewertung der Reliabilität: Da es bei Tests eventuell Angaben zu mehreren Reliabilitätsarten gibt und da bei Tests mit mehreren Untertests bzw. Skalen entsprechend mehrere Reliabilitätswerte vorliegen, führen die Rezensierenden die Vielzahl der Informationen zu einem Gesamturteil zur Reliabilität zusammen. Dabei sind vor allem die Reliabilitäten derjenigen Untertests bzw. Skalen zu berücksichtigen, die laut der diagnostischen Zielsetzung (Abschnitt 1) besonders wichtig sind. Zudem ist abzuwägen, ob die über die jeweiligen Schätzer bestimmte Reliabilität den diagnostischen Zielsetzungen (z. B. Statusdiagnostik oder Prognose) genügt.

- Zu berücksichtigen ist auch, dass die Reliabilitätsschätzwerte in Abhängigkeit von den untersuchten Gruppen variieren (eine besondere Bedeutung kommt der Homogenität der Gruppe hinsichtlich des gemessenen Konstrukts zu).
- Die Rezensierenden berücksichtigen darüber hinaus, dass die Schätzung der Reliabilität über Homogenitätsmaße (z. B. interne Konsistenz) ...
  - bei nahezu identisch gestalteten Items eine Überschätzung darstellt und
  - bei Tests mit einer Speed-Komponente, bei denen also nicht alle Testpersonen auch zur Bearbeitung der letzten Items kommen, unangemessen ist und ebenfalls zu einer Überschätzung führt.
- 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.
- Bei Tests, die von den Testautorinnen und -autoren bzw. Testanbietenden für die Einzelfalldiagnostik vorgesehen werden, sind Aspekte der Messpräzision zu berücksichtigen. Die Messpräzision bezieht sich auf das Verhältnis der Breite des Konfidenzintervalls zur Breite des Punkteranges. Da dieses Verhältnis angibt, wie groß der Bereich an Rohwerten ist, die sich lediglich aufgrund der Messungenauigkeit unterscheiden könnten, sollte ein möglichst geringer Wert resultieren. Je größer das resultierende Verhältnis, desto ungenauer fallen einzelfalldiagnostische Personenvergleiche oder Vergleiche mit Cut-offs aus.
- Bei algorithmenbasierten Tests sind je nach Einsatzzweck Test-Retest-Reliabilitäten anzugeben. Hier ist

4 Dabei geht es auch um eine angemessene Beschreibung sowohl der Population als auch der Art der Stichprobenziehung oder Datensammlung. Hierzu zählen zumindest Angaben zum Erhebungszeitpunkt und -ort, zu Umfang und Art (demografische Merkmale) der berücksichtigten Personen sowie zu den Umständen der Erhebung (z. B. Anonymität, Art der Teilnahme [z. B. anonym, freiwillig, vergütet usw.]). Des Weiteren geht es darum, ob bei der Datensammlung bloß von einer »anfallenden Stichprobe« Gebrauch gemacht wurde. Beispielsweise werden oft nur Schülerinnen und 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, da diese leicht verfügbar sind.

5 Bei der Beurteilung der Höhe sind die Einsatzzwecke zu berücksichtigen.

6 Die Bestimmung der internen Konsistenz ist beispielsweise keine angemessene Art der Zuverlässigkeitsbestimmung für Testkennwerte, die heterogene Konstrukte abbilden. Die Bestimmung der Retest-Reliabilität ist keine angemessene Art der Zuverlässigkeitsbestimmung für Verfahren zur Messung rasch veränderlicher Merkmale (z. B. Stimmungen).

7 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.

darauf zu achten, dass Angaben zur Accuracy eher die Modellgüte anzeigen und nicht die Reliabilität des zu interpretierenden Kennwerts. Auch ist darauf zu achten, ob das ML- oder KI-Modell eine Abschätzung der Messpräzision für individuelle Personenkennwerte ermöglicht (z. B. Konfidenzintervalle).

Zu 8: Gültigkeit (Validität)

DIN-Screen-Aussagen B27 bis B54 sowie Addendum.

- Werden die Validitätskoeffizienten für alle (Sub-) Populationen aus einer Stichprobenerhebung geschätzt, für die der Test laut diagnostischer Zielsetzung eingesetzt werden soll?
- Wird Kriteriumsvalidität beansprucht? Falls ja: Hat die Untersuchung zur Kriteriumsvalidität unter Testbedingungen stattgefunden, die den Bedingungen bei der Nutzung des Tests in der guten Praxis weitgehend entsprechen?
- Ist der Test für die Einzelfalldiagnostik vorgesehen? Falls ja: Liegen für Testkennwerte (z. B. Cut-offs) Angaben zur Sensitivität und Spezifität vor? Abhängig von der diagnostischen Zielsetzung sollten weitere Klassifikationsmaße (z. B. positiver und negativer prädiktiver Wert) angegeben sein.
- Wird Eindimensionalität angenommen? Falls ja: Liegen für die Testkennwerte Belege für die Eindimensionalität vor?
- Wird eine Hierarchie und/oder Mehrdimensionalität der Testkennwerte angenommen? Falls ja: Liegen empirische Belege für die postulierte Datenstruktur vor?
- Werden statistische Modelle eingesetzt (z. B. Strukturgleichungsmodelle, IRT-Modelle, ML- oder KI-Modelle)? Falls ja:
  - Ist deren Güte (mit Hilfe von für die entsprechende Modellklasse angemessenen Kennwerten) angegeben?
  - Ist ausgewiesen, ob und welche Daten für eine mögliche Kreuzvalidierung des Modells genutzt wurden?
- Handelt es sich um einen Test, der Algorithmen nutzt? Falls ja:
  - Ist angegeben, auf welche Kriterien die Algorithmen trainiert wurden und welche Kriterien dann für die Kriteriumsvalidität genutzt wurden?
  - Werden entscheidungstheoretische Kennzahlen berichtet? Falls ja: Erfolgt dies multiperspektivisch[8]?
  - Wurden die für den jeweiligen Anwendungsfall relevanten Kennwerte empirisch ermittelt?
  - Werden die für den jeweiligen Anwendungsfall relevanten Kennwerte in den Verfahrenshinweisen nachvollziehbar berichtet?

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

Da es bei Tests eventuell Angaben zu den Ergebnissen mehrerer Validierungsuntersuchungen gibt (z. B. Konstrukt- und Kriteriumsvalidierungen an verschiedenen Stichproben) und da bei Tests mit mehreren Untertests bzw. Skalen entsprechend mehrere Validitätswerte vorliegen, führen die Rezensierenden die Vielzahl der Informationen zu einem Gesamturteil zur Validität zusammen. Dabei sind vor allem die Validitäten der Interpretationen derjenigen Untertests bzw. 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. Die Rezensierenden prüfen, inwieweit die in den Verfahrenshinweisen berichteten Validitätsbelege zur Stützung des Testeinsatzes gemäß der diagnostischen Zielsetzung überzeugen.

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

- Die Rezensierenden berücksichtigen bei ihrem Urteil über die Angaben zur Validität der Testkennwertinterpretation, 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.
- In dem Fall, dass die Validitätsbefunde auf Mittelwertvergleichen beruhen (etwa bei einer Extremgruppenvalidierung), sollen die Rezensierenden die Relevanz des Effekts des Mittelwertunterschieds bewerten.
- Die Rezensierenden prüfen, ob die Validitätsuntersuchungen theoriegeleitet begründet werden und nicht nur ohne vorherige theoretische Einbettung Korrelationen als Validitätsbelege angeführt werden.
- Des Weiteren ist die inhaltliche und psychometrische Qualität der zur Validierung herangezogenen Maße (z. B. andere Tests zur Konstruktvalidität; Kriteriumsmaße) von den Rezensierenden zu beurteilen.
- Wenn Übereinstimmungsvaliditäten mit Kennwerten aus gleichartigen Tests angeführt werden, soll in die Beurteilung mit einfließen, inwieweit die konkurrierenden Tests selbst das Gütekriterium der Validität erfüllen.
- Die Rezensierenden 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 sind die psychometrische Qualität des Kriteriums (z. B. Reliabilität) zu beurteilen sowie die inhaltliche Qualität (z. B. inhaltliche Gültigkeit bzw. Relevanz). Zu bewerten sind schließlich die Art der Beziehung zwischen Prädiktor und Kriterium (z. B. linear/non-linear) sowie die Art der Analyse dieser Be-

Kontakt
Diagnostik- und Testkuratorium
Vorsitz: Prof. Dr. Martin Kersting
Justus-Liebig-Universität Gießen
Fachbereich 06 Psychologie und Sportwissenschaft
Abteilung für Psychologische Diagnostik
Otto-Behaghel-Str. 10F
35394 Gießen
E vorsitz-dtk@psychologie.de

8 Es sollte also z. B. nicht nur die Accuracy (Anteil von korrekten Vorhersagen an allen getroffenen Vorhersagen), sondern auch die Precision bzw. Sensitivität (Anteil der korrekten positiven Vorhersagen von allen als positiv klassierten Fällen) sowie die Spezifität berichtet werden.

ziehungen (z. B. einfache oder multiple Regression; Kreuzvalidierung an unabhängiger Stichprobe; Miteinbeziehung von Moderator- und Suppressor-Variablen, Sensitivität bzw. Spezifität).
- Falls in den Verfahrenshinweisen 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.
- Bei Tests, die Algorithmen nutzen, berücksichtigen die Rezensierenden, ob sich die Daten auf den Trainingsdatensatz oder auf einen Anwendungsdatensatz beziehen.

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

DIN-Screen-Aussage B14 (V4) sowie Addendum.

Bewertung weiterer Gütekriterien:

Störanfälligkeit: Die Rezensierenden berücksichtigen, in welchem Ausmaß der Test empfindlich ist gegenüber aktuellen Zuständen der Testperson und situativen Faktoren der Umgebung. Insbesondere soll geprüft werden, ob eine solche Störanfälligkeit angesichts der diagnostischen Zielsetzung ein Problem darstellt.

Unverfälschbarkeit: Die Rezensierenden beurteilen, inwieweit es beim gegebenen Test möglich ist, dass die Testperson durch ein gezieltes Testverhalten die konkrete Ausprägung ihres Testwerts steuern bzw. kontrollieren kann. 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 Zielsetzung ein Problem darstellen.

Skalierung: Die Verwendung formaler Testtheorien und insbesondere die IRT erlauben es, bei Tests kritisch zu hinterfragen, inwieweit die Zahlenrelationen der Testwerte mit den Relationen der beobachtbaren Verhaltensweisen – sowohl innerhalb ein und derselben Testperson als auch zwischen verschiedenen Testpersonen – übereinstimmen. Sie ermöglichen, basierend auf dem Testmodell, eine Abschätzung der Präzision für individuelle Personenkennwerte. Die Rezensierenden sollten im Fall, dass die Testkonstruktion und Auswertung nicht nach einem formalen Testmodell erfolgte, anführen, inwieweit in den Verfahrenshinweisen die Frage aufgegriffen und diskutiert wird, ob die laut den Verrechnungsvorschriften resultierenden Testwerte die empirischen Verhaltensrelationen adäquat und mit ausreichender Messpräzision abbilden. Dies gilt insbesondere auch für mit Hilfe von ML- oder KI-Modellen konstruierte Testverfahren.

Testfairness: Sie bezieht sich u. a. auf das Ausmaß einer eventuell bestehenden systematischen, auf den Testwerten aufsetzenden, Diskriminierung bestimmter Testpersonen, z. B. aufgrund ihrer ethnischen Herkunft, des Geschlechts, der Religion oder Weltanschauung, einer Behinderung, des Alters oder der sexuellen Identität. Für algorithmenbasierte Tests sollten potenzielle Diskriminierungen in den Verfahrenshinweisen diskutiert und die Ergebnisse von Bias-Checks dokumentiert werden. Dies gilt für Variablen, für die es aufgrund theoretischer Überlegungen oder empirischer Befunde naheliegend ist, dass sie Varianz in den Prädiktoren erzeugen (z. B. Geschlecht, Dialekt bei Sprachanalysen, Piercing bei Gesichtsanalysen).

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. https://doi.org/10.1007/978-3-662-53772-5