Foto: Nguyen Dang Hoang Nhu – unsplash.com

Qualitätssicherung- und -optimierung von Tests

Die vierte Fassung des Testbeurteilungssystems des Diagnostik- und Testkuratoriums (TBS-DTK)

Die gesellschaftliche Bedeutung von Tests[1] ist enorm. Testergebnisse beeinflussen wichtige Lebensbereiche der diagnostizierten Personen und der Gesellschaft, beispielsweise Bildung, Beruf und (psychische) Gesundheit. Sie wirken auf das Selbstbild und Fremdbild von Personen ein. Gleichzeitig prägen Tests als sichtbares »Produkt« und »Werkzeug« das Bild der Psychologie in der Gesellschaft.

Der sinnvolle Einsatz von qualitativ hochwertigen Tests, die professionell durchgeführt und fachgerecht interpretiert werden, kann sich außerordentlich positiv auf Menschen und auf die Gesellschaft auswirken. Allerdings können der fehlerhafte Einsatz von Tests sowie die mangelhafte Qualität von Tests auch immense negative Wirkungen verursachen. Ungeachtet dieser hohen Bedeutung gibt es – außer in sehr spezifischen Anwendungsbereichen – in Deutschland keine Gesetze und Verordnungen, die sich mit der Qualität von Tests und Testungen befassen. Die Nutzung des Begriffs »Test« ist ungeschützt. Eine staatliche Qualitätskontrolle von Tests oder von Personen, die Tests entwickeln oder anwenden, findet in den meisten Bereichen nicht statt.

Das Diagnostik- und Testkuratorium (DTK) ist für alle Aspekte der Qualitätssicherung und Qualitätsoptimierung des auf menschliches Erleben und Verhalten bezogenen diagnostischen Prozesses zuständig. Ein wesentliches Instrument dabei ist das Testbeurteilungssystem des Diagnostik- und Testkuratoriums (TBS-DTK).

Mit dem TBS-DTK werden klare Anforderungen an den Informationsgehalt von Verfahrenshinweisen (Testhandbücher/-manuale) sowie Qualitätsstandards in Form von Beurteilungsrichtlinien formuliert. Auf diese Weise trägt das DTK zur Verbreitung verlässlicher Informationen sowie fachlich fundierter, nachvollziehbarer Urteile über Tests bei. Darüber hinaus dient das System als Leitfaden für die fachgerechte Entwicklung und Optimierung dieser Instrumente und ist somit kulturprägend.

Die erste Fassung des Systems wurde 2006 publiziert. 2009 wurde die zweite, 2018 die dritte und 2023 die vierte Fassung publiziert. Für eine Kurzdarstellung der früheren Fassungen verweisen wir auf Hagemeister, Kersting und Stemmler (2012) sowie auf Moosbrugger, Stemmler und Kersting (2008). Die Notwendigkeit, ein solches System einzuführen, begründete Kersting (2006). Im Folgenden geben wir einen kurzen Überblick über die Ergebnisse der bislang publizierten Rezensionen. Außerdem benennen und kommentieren wir die wesentlichen Weiterentwicklungen, die in der vierten Fassung vorgenommen wurden.

Ergebnisse der publizierten Rezensionen

Die erste Rezension nach dem TBS-DTK System wurde 2008 publiziert. In dem 15 Jahre umfassenden Zeitraum von 2008 bis Mitte 2023 wurden insgesamt 59 Rezensionen zu Tests aus verschiedensten Anwendungsbereichen veröffentlicht oder befanden sich »im Druck«. Seit der dritten, 2018 publizierten Version des TBS-DTK Systems gibt es zehn »Besprechungs- und Beurteilungskategorien«, vorher waren es neun. Neu hinzugekommen ist 2018 eine formalisierte Bewertung (»ja«/»nein«) zur Frage, ob die Verfahrenshinweise eine Übersichtstabelle zur DIN-Screen-Checkliste (Kersting, 2018) enthalten.

In dieser Tabelle sind alle Informationen aufgeführt, die nach Ansicht des DTK zu einem Test vorliegen müssen (DTK-Testinformationsstandard). Beispiele sind In-

1 Die Bezeichnung »Test« wird im vorliegenden Text als Oberbegriff genutzt: Damit sind messtheoretisch fundierte Fragebogen (z. B. Persönlichkeitsfragebogen, Interessenfragebogen) und messtheoretisch fundierte Tests (z. B. Intelligenz- und Wissenstests) gemeint. Auch Tests und Verfahren, die mittels Algorithmen (die z. B. mittels Machine Learning [ML] oder künstlicher Intelligenz [KI] erstellt wurden) Personenkennwerte schätzen, fallen unter den Begriff »Test«. Der Artikel thematisiert ausschließlich Tests, die sich auf menschliches Erleben und Verhalten beziehen.

formationen zur Untersuchungsgruppe, zum Jahr der Datenerhebung, zu den Reliabilitäts- und Validitätsstudien usw. Für die Kategorien sind freie und/oder formalisierte Bewertungen vorgesehen. Formalisierte Bewertungen auf einer vierstufigen Skala gibt es für die Kategorien (1) Bewertung des Informationsgehalts der Verfahrenshinweise, (2) Objektivität, (3) Reliabilität und (4) Validität. Die Skalenstufen lauten: »Der Test erfüllt die Anforderungen …«

- »voll«
- »weitgehend«
- »teilweise«
- »nicht«.

Die ab 2018 neu aufgenommene Frage, ob die Verfahrenshinweise eine Übersichtstabelle zur DIN-Screen-Checkliste umfassen, wird dichotom (»ja«/»nein«) beantwortet.

Tabelle 1 gibt eine Übersicht über die formalisierten Bewertungen aller bis Mitte 2023 publizierten TBS-DTK Rezensionen.

Um einen ersten Gesamteindruck von den bislang vorliegenden formalisierten Bewertungen zu gewinnen, kann man die vierstufige Bewertungsskala in zwei Hälften aufteilen und die beiden Einstufungen »nicht« und »teilweise« vereinfacht als »eher negative« und die beiden Einstufungen »weitgehend« und »voll« vereinfacht als »eher positive« Bewertung interpretieren. Dies entspricht der Erläuterung der vier Bewertungsstufen, die in der vierten Fassung des TBS-DTK ergänzt wurde.

Geht man so vor, erzielen in Bezug auf die Häufigkeit der Urteilskategorien 88 % aller betrachteten Verfahren eine positive Bewertung in Hinsicht auf den Informationsgehalt der Verfahrenshinweise. Bei 84 % der Verfahren wird die Objektivität und bei 70 % der Verfahren die Reliabilität positiv beurteilt. Lediglich bezüglich des Hauptgütekriteriums Validität überwiegt mit 58 % ein negatives Urteil.

Für eine genauere Analyse wurden für die vier Bewertungskriterien mit einer quantitativen Bewertung die Mittelwerte pro Kategorie bestimmt. Dabei wurden die Einstufungen »nicht«, »teilweise«, »weitgehend« und »voll« mit »1«, »2«, »3« und »4« codiert. Die Mittelwerte lauten: Bewertung des Informationsgehalts der Verfahrenshinweise: 3.3; Objektivität: 3.1; Reliabilität: 2,8; und Validität: 2,4. Das Urteil fällt also kritischer aus, wenn es um die Gütekriterien Reliabilität und Validität geht.

Schätzwerte für die Reliabilität und Validität werden in der Regel aus empirischen Studien abgeleitet, deren Durchführung und Interpretation aufwendig und anspruchsvoll ist. Aus dem Bewertungsmittelwert allein lässt sich nicht schlussfolgern, ob die Verfahren selbst den Anforderungen nicht gerecht werden oder ob die entsprechenden empirischen Studien den Anforderungen nicht ausreichend gerecht werden. Insgesamt bedarf es in diesem Bereich verstärkter Forschung und Evaluation.

An dieser Stelle nehmen wir keine Auswertung der Freitexte der Rezensionen vor, die sich beispielsweise auf die Qualität der Normen beziehen. Diese Qualität wird in den TBS-DTK-Rezensionen thematisiert, aber nicht mit einer »Note« beurteilt.

Für zwei Tests musste konstatiert werden, dass die festgelegten Anforderungen bezüglich Information und Dokumentation nicht erfüllt werden. Um einen Test beurteilen zu können, benötigt man Informationen. Tests, zu denen keine ausreichenden Informationen vorliegen, entziehen sich einer Bewertung; ihr Einsatz ist schwer zu verantworten. Das DTK hat sich bezüglich dieser Qualitätsforderung eindeutig positioniert und konkret festgelegt, welche Informationen zu einem Test vorliegen müssen. Dieser Katalog wurde aus der DIN 33430 (DIN, 2016) abgeleitet und in Form einer Checkliste (»DIN-Screen-Checkliste«, Kersting, 2018) aufbereitet. Wenn zu einem Test diese Informationen nicht vorliegen, 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.«

Dieses Urteil trifft für die beiden Tests »InCheck« und »Lumina Spark« zu. Beide Verfahren wurden von den weiter oben stehenden Auswertungen ausgenommen.

Die Verfahrenshinweise zu einem Test müssen die für den Testeinsatz relevanten Informationen liefern. Idealerweise befindet sich in den Verfahrenshinweisen bereits eine Tabelle, in der auf einen Blick abgelesen werden kann, an welcher Stelle der Verfahrenshinweise die jeweiligen Informationen zu finden sind. Für Anwenderinnen und Anwender ist das Vorhandensein einer solchen tabellarischen Übersicht ein erster Indikator für die Qualität des Tests. Seit der dritten Version des TBS-DTK gehört die Prüfung, ob eine solche Übersicht vorhanden ist, zu den formalen Bewertungskriterien des TBS-DTK. Wie die Bewertung in der Spalte »DIN-Screen« (Tabelle 1) zeigt, wird die Möglichkeit, die Qualität eines Tests durch eine solche tabellarische Übersicht herauszustellen, so gut wie nicht genutzt.

Modifikationen des Systems

Das DTK überprüft regelmäßig, ob das TBS-DTK noch zeitgemäß ist. Den Entwurf der vierten Fassung des TBS-DTK hat das DTK in der Fachöffentlichkeit zur Diskussion gestellt. Dieser partizipative Prozess sollte sicherstellen, dass die Modifikation vom Fachwissen vieler Kolleginnen und Kollegen mit einem breiten Erfahrungsspektrum profitiert.

Prozess und Struktur des TBS-DTK haben sich bewährt und wurden – ebenso wie die Bewertungskriterien – für die vierte Fassung übernommen.

Tabelle 1
Übersicht über die formalisierten Bewertungen aller bis Mitte 2023 publizierten TBS-DTK Rezensionen (Anmerkungen siehe S. 17)

Testname	Publikation der Rezension	Allgemeine Information	Objektivität	Zuverlässigkeit	Validität	DIN-Screen
DISK-Gitter	2010	++	++	++	++	*
DERET	2011	++	++	+	++	*
FPD	2018	++	++	+	++	*
BIP-6 F	2014	++	++	+	+	*
FPI-R	2011	++	++	+	+	*
IBES	2018	++	++	+	+	*
IEA	2018	+	++	++	+	*
IST2000-R	2008	++	++	+	+	*
OLMT	2008	++	++	+	+	*
OPQ32	2010	++	++	+	+	*
WISC-V	2022	++	+	++	+	nein
BDI-II	2008	+	+	+	++	*
CBCL/6-18R	2018	++	+	+	+	*
FAKT-II	2011	+	++	+	+	*
FPP	2020	+	++	+	+	nein
INSBAT	2014	+	+	++	+	*
ISK	2023	+	++	+	+	nein
K-ABC	2012	+	++	+	+	*
KABC-II	2017	++	+	+	+	*
NEO-PI-R	2008	++	+	+	+	*
PSSI	2021	+	++	+	+	nein
S-Tool	2019	++	++	+	-	ja
TOP	2020	++	++	+	-	nein
WISC-IV	2016	++	+	++	-	*
CFT 20-R	2015	+	+	+	+	*
EXPLORIX	2017	++	+	+	-	*
FRAKK	in Druck	++	+	+	-	nein
LSA	2019	++	+	+	-	nein
WAIS-IV	2021	++	+	+	-	nein
WIE	2010	++	+	+	-	*
AID 3	2017	+	+	+	-	*
AKGT	2018	+	+	-	+	*
AVEM	2017	+	+	+	-	*
d2-R	2015	+	+	+	-	*
DESC	2019	++	-	+	-	*
ET_6-6-R	2015	++	+	-	-	*
FEEL-KJ	2015	+	+	+	-	*
IVPE-R	2020	+	+	+	-	*
Klasse 4	2014	++	+	-	-	*
MOA 7.0	2019	-	++	+	-	*
MSCEIT	2015	+	+	+	-	*
SURT	2017	+	+	-	+	*
B5T	2022	+	+	-	-	nein
DCS-II	2016	+	+	-	-	*
PFK 9-14	in Druck	+	+	-	-	nein
SCL-90-R	2012	+	+	-	-	*
TEA-CH	2013	+	+	-	-	*
VVKI	2013	+	-	+	-	*
ET_6-6 (3. Aufl.)	2012	+	-	-	-	*
PRECIRE	2019	-	+	-	-	nein
SKEI	2018	+	-	-	-	*
Cito	2020	-	+	-	--	nein
Frakis	2010	+	-	-	--	*
Persolog	2013	-	-	+	--	*
PPI-R	2010	-	-	-	-	*
GPOP	2009	-	-	-	--	*
Familie in Tieren	2015	-	--	--	--	*
InCheck	2018	Die Verfahren (1) InCheck und (2) Lumina Spark erfüllen 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.				
Lumina Spark	in Druck					

Anmerkungen zu Tabelle 1:
- Skala: »Erfüllt die entsprechenden TBS-DTK Anforderungen …«: »--« = nicht; »-« = teilweise; »+« = weitgehend; »++« = voll
- Die Verfahren sind nach ihrer Bewertung geordnet dargestellt. Bei gleicher Bewertung erfolgt die Anordnung alphabetisch.
- Spalte »DIN-Screen-Checkliste«: »In den Verfahrenshinweisen ist verzeichnet, wo die nach dem DTK-Testinformationsstandard notwendigen Informationen zu finden sind.« (»ja«/»nein«)
- * = Zum Zeitpunkt der Bewertung gab es diese Bewertungskategorie noch nicht.

Die vierte Version wurde vor allem notwendig, um aktuellen Entwicklungen im Kontext der Digitalisierung von Tests Rechnung zu tragen. Vorrangiges Ziel der Modifikation war es, sicherzustellen, dass das TBS-DTK auch auf solche Tests angewendet werden kann, die mittels Algorithmen (die z. B. mit Hilfe von Modellen des maschinellen Lernens oder der künstlichen Intelligenz entwickelt wurden) Personenkennwerte schätzen.[2] Dabei zeigte sich, dass die bislang etablierten Bewertungskriterien ausnahmslos auch für diese »neue Art« von Tests angemessen sind. Insbesondere die im TBS-DTK geforderte Transparenz bezüglich der Datenbasis (z. B. Untersuchungs- und Normgruppen) ist auch ein wesentliches Qualitätskriterium für Tests, die Algorithmen aus (Trainings-)Datensätzen ableiten. Bei Tests, die Algorithmen nutzen, sind aber zusätzliche Anforderungen zu stellen; z. B. bedarf es umfassender Informationen über die Entstehung und die Nutzung der eingesetzten Algorithmen. Ergänzt wurden darüber hinaus Anforderungen an die Nachvollziehbarkeit der mittels solcher Tests gewonnenen Bewertungen. Ein besonderes Augenmerk ist bei diesen Tests außerdem auf potenzielle Diskriminierungen zu richten. Das TBS-DTK wurde entsprechend zeitgemäß erweitert, die DIN-Checkliste wurde durch ein Addendum um diese spezifischen Aspekte ergänzt.

Eine weitere wesentliche Änderung besteht darin, dass eine Erläuterung der Skalenstufen ergänzt wurde. Zunächst wurde die Verbindung zwischen der Erfüllung der Anforderungen und der Qualität des Tests herausgestellt. Die Anforderungen »voll«, »weitgehend«, »teilweise« oder »nicht« zu erfüllen, bedeutet »höchste«, »gering eingeschränkte«, »eingeschränkte« oder »erheblich eingeschränkte« Qualität. Darüber hinaus wurde für jede der vier Beurteilungskriterien (z. B. »Validität«) jeder Skalenpunkt verbal erläutert.

Fazit

Angesichts der hohen Bedeutung von Tests für das Individuum und die Gesellschaft besteht die Notwendigkeit, die Qualität von Tests zu sichern, und die Herausforderung, ihre Qualität beständig zu optimieren. Ein wesentliches Instrument dafür ist das TBS-DTK. Es wird seit 2006 produktiv genutzt, um über Tests zu informieren und Tests zu bewerten. Der Nutzen des TBS-DTK für die Psychologie und die Gesellschaft beschränkt sich nicht auf die vom DTK initiierten Rezensionen, die in Publikationen münden. Jede Person, die sich ein Bild über die Qualität von Tests machen will und über die entsprechende Expertise verfügt, kann für den Test ihres Interesses die TBS-DTK Beurteilungskriterien und -richtlinien anwenden. Organisationen können Tests und Testangebote auf dieser Basis beurteilen. Dozierende an Hochschulen können Studierenden anhand des Systems vermitteln, auf welche Aspekte man bei der Testentwicklung achten muss und wie die Qualität von Tests zu bewerten ist. Zugleich können sich Testentwicklerinnen und -entwickler an diesem Qualitätsstandard orientieren.

Betrachtet man alle bislang publizierten TBS-DTK-Rezensionen, so zeigen sich auf aggregierter Ebene die Stärken und Entwicklungsbereiche von Tests insgesamt. Damit wird deutlich, in welchen Qualitätsbereichen eine verstärkte Anstrengung notwendig ist, damit die Testqualität insgesamt noch weiter optimiert wird. Hinsichtlich des Bewertungskriteriums »Reliabilität« werden 70 % und hinsichtlich des Bewertungskriteriums »Validität« 42 % der Verfahren als positiv bewertet. Die Korrelation zwischen den Bewertungen beträgt .5, was zeigt, dass es tendenziell so ist, dass Verfahren eher in beiden Bereichen positiv abschneiden.

Die bisherigen Rezensionen können als Beleg für die heterogene Qualität auf dem Testmarkt angesehen werden. Um sich auf einem solchen Markt orientieren zu können, bedarf es der Informationen über Tests und der Bewertungen von Tests. Durch das TBS-DTK ist geregelt, welche Informationen zu einem Test vorliegen müssen, und das System definiert, welche Kriterien für die Qualität relevant sind.

Damit das System seinem Qualitätsanspruch selbst gerecht wird, muss von Zeit zu Zeit geprüft werden, ob das Vorgehen noch zeitgemäß ist. Bei der letzten Überprüfung zeigte sich vor allem ein Entwicklungsbedarf infolge der zunehmenden Digitalisierung. Das System musste dahin gehend erweitert werden, dass auch die Qualität von Tests, die Algorithmen nutzen, umfassend beurteilt werden kann. Informationen über Tests sowie nachvollziehbare Beurteilungen der Qualität von Tests sind eine wichtige Grundlage, um Vertrauen in das Gelingen psychologischer Diagnostik zu schaffen. Die Modifikationen, die zur vierten Fassung des TBS-DTK geführt haben, tragen dazu bei, dass das System das zentrale Fundament der Qualitätssicherung für Tests bleibt.

Diagnostik- und Testkuratorium (DTK) der Föderation Deutscher Psychologenvereinigungen[3]

2 Die die KI betreffenden Passagen der aktuellen Fassung wurden unter Mitarbeit von Prof. Dr. Clemens Stachl und Dr. Florian Pargent entwickelt.

3 Mitglieder des DTK waren zum Zeitpunkt der Texterstellung: 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.

Literatur:

Hagemeister, C., Kersting, M. & Stemmler, G. (2012). Test reviewing in Germany. *International Journal of Testing, 12*(2), 185–194.

Kersting, M. (2006). Zur Beurteilung der Qualität von Tests: Resümee und Neubeginn. *Psychologische Rundschau, 57*(4), 243–253.

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

Moosbrugger, H., Stemmler, G. & Kersting, M. (2008). Qualitätssicherung und -optimierung im Aufbruch: Die ersten Testrezensionen nach dem neuen TBS-TK-System. *Report Psychologie, 33*(6), 299–300.

G 3777
FACHZEITSCHRIFT DES BDP
ZEITSCHRIFT DES BERUFSVERBANDES DEUTSCHER PSYCHOLOGINNEN UND PSYCHOLOGEN E.V.
48. JAHRGANG
NOVEMBER/DEZEMBER 2023

11+12|2023

report**psychologie**

Psychologische Kompetenzen