Kriterienkatalog zur Beurteilung psychodiagnostischer Selbstbeurteilungsinstrumente – Empfehlung des Deutschen Kollegiums für Psychosomatische Medizin (DKPM)

Review Model for the Assessment of Psychometric Instruments – Recommendations of the German College of Psychosomatic Medicine (DKPM)

Autoren Heide Glaesmer[1], Thomas Forkmann[2], Andreas Dinkel[3], Inka Wahl[3,4], Heribert Sattel[3], Dorothea Huber[5,6], Lena Spangenberg[1], Sven Rabung[7,8], Sylke Andreas[8], Karin Tritt[9], Gabriele Helga Franke[10], Matthias Rose[11], Bernd Löwe[4]

Institute Die Institutsangaben sind am Ende des Beitrags gelistet

Schlüsselwörter
- Psychodiagnostik
- Instrumente
- Gütekriterien
- Beurteilung
- Katalog

Key words
- psychdiagnostics
- instruments
- psychometric properties
- evaluation
- catalogue

Zusammenfassung

Hintergrund: In der psychosomatischen und psychotherapeutischen Versorgung und Forschung finden psychodiagnostische Instrumente Verwendung im Rahmen von Indikationsstellung, Behandlungsplanung, Verlaufsbeurteilung und -monitoring sowie Qualitätssicherung und Ergebnismessung. Bei der Auswahl dieser Instrumente sollten sowohl deren psychometrische Qualität und Eignung als auch anwendungsökonomische Gesichtspunkte berücksichtigt werden. Es ist davon ausgehen, dass sich der Anwender angesichts der Vielzahl von Instrumenten und Informationsquellen über den neuesten Stand der Forschung kaum einen Überblick verschaffen kann. Es besteht deshalb die Gefahr, dass vorwiegend bekannte, lange etablierte und gut zugängliche bzw. verfügbare Verfahren eingesetzt werden, während inhaltliche und psychometrische Neuentwicklungen kaum rezipiert werden. Diese eher pragmatische Auswahl dürfte nicht immer den verfolgten Zielen dienen.

Material und Methoden: Die Arbeitsgruppe „Psychometrie und Psychodiagnostik“ des Deutschen Kollegiums für Psychosomatische Medizin (DKPM) hat basierend auf verschiedenen internationalen Testbeurteilungssystemen einen für die klinische Forschung und Praxis zugeschnittenen Kriterienkatalog entwickelt und getestet.

Ergebnisse: Die Entwicklungsschritte und der von der Arbeitsgruppe konsentierte Katalog wird vorgestellt. Unter den Oberbegriffen Reliabilität, Validität, Objektivität, Referenzgruppen und Anwendung sind insgesamt 21 Kriterien subsummiert, die mit 0–3 Sternen (*) anhand klar operationalisierter Kriterien eingeschätzt werden sollen. Die praktische Anwendung des Kataloges wird erläutert und diskutiert.

Schlussfolgerung: Mit dem Kriterienkatalog wird ein gut formalisiertes Beurteilungssystem vorgelegt, welches von einem Expertengremium

Abstract

Objectives: Psychometric instruments are commonly applied in psychotherapeutic research and care for the baseline assessment of symptoms, the planning of therapeutic interventions, the assessment of the longitudinal course of symptoms and outcomes of therapeutic interventions as well as quality management of care. Psychometric properties as well as economic aspects should be considered in the selection of specific instruments. It is assumed that users of psychometric instruments face a great variety of instruments and related information. For that reason, it seems challenging to absorb the current knowledge and to integrate it into clinical practice and research. Thus, it is likely that well-known, established and easily accessible instruments are commonly used, while new developed instruments might not be disseminated in research and healthcare.

Methods: Based on available international review models, the working group “Psychometrics and Psychodiagnostics” of the German College of Psychosomatic Medicine (DKPM) has developed and tested a review model specifically tailored for psychotherapeutic research and care.

Results: The different steps of development, as well as the final review model based on the consensus of the working group are presented. The review model contains 6 generic terms (reliability, validity, objectivity, reference groups and aspects of application) with 21 different criteria to be assessed with 0–3 asterisks (*). The criteria are clearly operationalized and the practical use of the review model is explained and discussed.

Conclusions: With the review model for the assessment of psychometric instruments a well-defined evaluation system is made available for research and clinical practice which has been developed by an expert group. The review model facilitates systematic, transparent and compara-

eingereicht 19. Mai 2014
akzeptiert 19. Januar 2015

Bibliografie
DOI http://dx.doi.org/10.1055/s-0034-1398685
Online-Publikation: 28.4.2015
Psychother Psych Med 2015; 65: 246–254

ISSN 0937-2032

Korrespondenzadresse
PD. Dr. Heide Glaesmer
Abteilung für Medizinische Psychologie und Medizinische Soziologie
Universität Leipzig
Phillipp-Rosenthal-Straße 55
Leipzig 04103
heide.glaesmer@medizin.uni-leipzig.de

entwickelt und konsentiert wurde. Eine nach definierten Kriterien verfasste und damit transparente und systematische Testbeurteilung wird damit unterstützt. Es besteht die Möglichkeit verschiedene Tests zu vergleichen. Der Katalog dient aber auch der Unterstützung der Testauswahl durch Praktiker in Forschung und Praxis. Eine wesentliche Aufgabe besteht jetzt darin, den Katalog zu disseminieren und zu implementieren sowie Testbeurteilungen anhand des Kataloges zu erstellen und der Fachöffentlichkeit zur Verfügung zu stellen.

tive evaluation of psychometric instruments along clearly defined criteria. It also supports the selection of psychometric instruments in research and care. Next, the working group aims at disseminating and implementing the review model as well as the application and publication of reviews for different psychometric instruments based on the review model.

Einleitung

In der psychosomatischen und psychotherapeutischen Versorgung und Forschung werden psychometrische und psychodiagnostische Instrumente im Rahmen von Indikationsstellung, Behandlungsplanung, Verlaufsbeurteilung, Qualitätssicherung und Ergebnismessung eingesetzt. So gewinnen in der Behandlung psychischer Störungen wie auch chronischer körperlicher Erkrankungen von den Patienten selbst berichtete Ergebnismaße („Patient Reported Outcomes“, PRO) zunehmend an Bedeutung. Bei der Auswahl psychometrischer Verfahren zur Erfassung relevanter Variablen bzw. Konstrukte für die beschriebenen Anwendungsbereiche sollten sowohl deren psychometrische Qualität und Eignung als auch anwendungsökonomische Gesichtspunkte berücksichtigt werden. In der Psychodiagnostik stehen verschiedenste Haupt- und Nebengütekriterien zur Verfügung, anhand derer die Bewertung und Auswahl von Testverfahren vorgenommen werden kann. Während für einige Zielbereiche (Messkonstrukte) Instrumente erst noch entwickelt werden müssen, gibt es für viele andere Bereiche national und international eine Vielzahl von Verfahren. Auch wenn für viele dieser Instrumente gute psychometrische Eigenschaften berichtet werden (z. B. überzeugende Werte hinsichtlich Validität, Reliabilität und Objektivität), stellen die schwierige Überschaubarkeit und Heterogenität der Instrumente zu einem Konstrukt bzw. Merkmalsbereich und die daraus resultierende mangelnde Vergleichbarkeit der unterschiedlichen Instrumente ein Problem dar. Bei der Erstellung von Metaanalysen oder systematischen Reviews wird das Problem der mangelnden Vergleichbarkeit immer wieder deutlich [1]. Für verschiedene Bereiche, wie z. B. Psychotherapie, liegen Überblickswerke vor [2], in denen die einzelnen Messinstrumente beschrieben werden. Trotzdem muss man davon ausgehen, dass sich der Anwender angesichts der Vielzahl von psychometrischen Instrumenten und Informationsquellen über den neuesten Stand der Forschung kaum einen Überblick verschaffen kann. Im klinischen Alltag besteht deshalb die Gefahr, dass vorwiegend bekannte, lange etablierte und gut zugängliche bzw. verfügbare Verfahren eingesetzt werden, während inhaltliche und psychometrische Neuentwicklungen kaum rezipiert werden. Diese eher pragmatische Auswahl dürfte nicht immer den verfolgten Zielen dienen. Diese Problematik besteht allerdings nicht nur in der psychosomatischen und psychotherapeutischen Versorgung und Forschung, sondern stellt sich zwangsläufig in allen Bereichen ein, in denen psychodiagnostische Verfahren eingesetzt werden, also z. B. auch in der Leistungs-, Entwicklungs- oder Persönlichkeitsdiagnostik. Testbeurteilungssysteme können helfen, die mögliche Gefahr, dass Testverfahren vornehmlich aufgrund von Bekanntheit und Konvention und nicht aufgrund von psychometrischer Qualität ausgewählt werden, abzumildern, indem sie in strukturierter und praktikabler Form Gütekriterien definieren und operationalisieren. Somit können Testbeurteilungssystem eine vergleichende Bewertung verschiedener, bekannter und weniger bekannter und verbreiteter Tests erleichtern. Mit Testbeurteilungen, die auf expliziten Qualitätskriterien beruhen, bekommt der Anwender umfassende Informationen an die Hand, welche ihm eine weitgehend objektive und anwendungsorientierte Auswahl von Testverfahren für seine spezifische Fragestellung ermöglichen. Es existieren international verschiedenste Testbeurteilungssysteme, die aber nicht spezifisch für klinisch-diagnostische Anwendungen konzipiert sind. Einige wesentliche Entwicklungen sollen hier kurz beschrieben werden.

Die European Federation of Psychologists Association (EFPA) hat das *„Review Model for the Description and Evaluation of Psychological Tests“* entwickelt. Dieses Modell soll die einheitliche Testbeschreibung und -bewertung innerhalb Europas unterstützen und liegt in einer von Lindley et al. [3] publizierten Fassung vor. Inzwischen wurde zudem eine Revision des Modells aus dem Jahr 2013 online verfügbar gemacht (http://www.efpa.eu/professional-development/assessment). Es handelt sich bei diesem Beurteilungssystem der EFPA um ein Formular mit definierten Kriterien zur Beurteilung und Beschreibung von Tests und einer Erläuterung für die Beurteiler. Teil 1 erfasst zunächst eine allgemeine Beschreibung des Tests, Teil 2 beinhaltet eine Beschreibung der Anwendungsbereiche (z. B. Thematik, Art der Anwendung, Zielgruppen, Zahl der Items und Skalierung), Teil 3 beschreibt Datenerfassung und -auswertung, Teil 4 beschreibt verfügbare computergestützte Anwendungen und Auswertungen, Teil 5 erfasst Verfügbarkeit und Kosten, Teil 6 nimmt eine Bewertung des Tests anhand verschiedener Kriterien auf einer Skala von 0 („nicht einschätzbar/fehlende Informationen“) über 1 („ungenügend“) bis 5 („ausgezeichnet“) vor. Teil 7 beschreibt Normen und klassische Gütekriterien des Tests, Teil 8 bewertet die Qualität der computergestützten Auswertungen. In Teil 9 wird eine abschließende Gesamtbewertung durchgeführt. Es ist vorgesehen, dass 2 Beurteiler zunächst eine Einschätzung vornehmen, die dann durch einen dritten Gutachter zusammengefügt und geprüft wird. Das aktuelle EFPA-Modell basiert im Wesentlichen auf Systemen aus verschiedenen europäischen Ländern, u. a. aus den Niederlanden (Committee on Test Affairs Netherlands, COTAN; Dutch Association of Psycholgists, NIP) und Großbritannien (British Psychological Society, BPS) [4]. Neben diesem europäischen System existiert auch ein Testbeurteilungssystem der American Psychological Association („Standards of Educational and Psychological Testing“) (http://www.apa.org/science/programs/testing/standards.aspx). Es besteht aus 3 Sektionen: (1) Testkonstruktion, Auswertung und Dokumentation, (2) Fairnessaspekte von Tests und (3) Aspekte der Anwendung von Tests. 2014 wurde eine überarbeitete Version verabschiedet, die derzeit aber noch nicht publiziert ist (http://teststandards.org/).

Im deutschen Sprachraum waren Testbeurteilungen lange nicht einheitlich reguliert. 1986 wurde der „Kriterienkatalog für die Beurteilung psychologischer Tests“ als Orientierungsrahmen definiert [5]. Er enthielt eine Auflistung zu beurteilender Aspekte, jedoch keine konkreten Handlungsanweisungen zur Einschätzung. Ein solches nicht formalisiertes Vorgehen lässt zu viele Gestaltungsmöglichkeiten. In der praktischen Umsetzung waren die Testbeurteilungen oft aber überaus kritisch und zumeist wenig konstruktiv formuliert [6]. Im Jahr 2002 wurde die DIN 33430 „Anforderungen an Verfahren und deren Einsatz für Eignungsbeurteilungen“ etabliert, die jedoch gezielt für berufsbezogene Eignungsbeurteilungen entwickelt wurde. Es handelt sich dabei eher um ein kurzes Screening von Verfahren zur Eignungsbeurteilung, jedoch nicht um ein elaboriertes Beurteilungssystem [7]. Erst im Jahr 2006 wurde das Testbeurteilungssystem des Testkuratoriums der Föderation Deutscher Psychologievereinigungen (TBS-TK) publiziert [8]. Das TBS-TK dient Testautoren, Verlagen und Anbietern sowie Rezensenten zur Qualitätssicherung. Testrezensionen nach dem TBS-TK werden von 2 Rezensenten durchgeführt. Dazu werden die Tests anhand von 7 Kriterien beschrieben. 3 der 7 Kriterien (Objektivität, Normierung, Zuverlässigkeit) werden formalisiert auf einer 4-stufigen Skala beurteilt. Die Bewertungen beider Rezensenten werden dann zusammengeführt und publiziert [8–10]. Die sehr detaillierten Handlungsanweisungen des TBS-TK an die Rezensenten führen allerdings auch zu einem erheblichen Arbeitsaufwand für den jeweiligen Rezensenten, was nicht zuletzt dazu beigetragen haben kann, dass in den 8 Jahren seit Veröffentlichung des TBS-TK erst 19 Testrezensionen publiziert werden konnten (http://www.zpid.de/index.php?wahl=Testkuratorium).

Wie aus diesem kurzen Überblick ersichtlich wird, existieren international verschiedene Testbeurteilungssysteme, die sich in ihren Herangehensweisen grundsätzlich ähneln. Der Formalisierungsgrad der Beurteilungen unterscheidet sich jedoch deutlich. Die Testbeurteilungssysteme sind so konzipiert, dass sie die Beurteilung durch Experten stärker formalisieren und vergleichbar machen, aber sie sind nicht derart gestaltet, dass sie den Testanwender selbst in die Lage versetzen würden, eine Entscheidung über die Auswahl eines von mehreren konkurrierenden Instrumenten zum gleichen Messgegenstand zu treffen. Üblicherweise werden numerische Vorgaben zur Beurteilung der verschiedenen Kriterien (etwa bzgl. der Reliabilität) gemacht, die nicht unbedingt gut begründbar sind, sich aber leicht anwenden lassen [6].

Die verfügbaren Testbeurteilungssysteme sind für die gesamte Bandbreite an psychologischen Tests entwickelt worden und bilden nicht die spezifischen Erfordernisse an Instrumente ab (z. B. Instrumente mit kurzer Bearbeitungsdauer, Verfügbarkeit von Cut-off-Werten zur Einschätzung der Schweregrade, Referenzwerte aus der Allgemeinbevölkerung und aus spezifischen Patientengruppen), die speziell zur Diagnostik und Forschung in Psychosomatik und Psychotherapie eingesetzt werden. Um eine qualitativ hochwertige und evidenzbasierte Beurteilung und Auswahl von psychometrischen Verfahren in der psychosomatischen und psychotherapeutischen Versorgung und Forschung zu unterstützen, bedarf es deshalb aus unserer Sicht eines speziell dafür konzipierten Kriterienkatalogs. Die Arbeitsgruppe „Psychometrie und Psychodiagnostik“ des Deutschen Kollegiums für Psychosomatische Medizin (DKPM) [9] hatte sich deshalb die Aufgabe gestellt, einen Kriterienkatalog für genau diesen Zweck zu erarbeiten. Ziel war es, einen Kriterienkatalog zu entwickeln, der geeignet ist, (1) die Beurteilung klinischer Testverfahren durch Experten zu formalisieren und transparent darstellen zu können und (2) Anwender selbst in die Lage zu versetzen, Verfahren zu beurteilen.

Methodik

Zwischen Mai 2011 und November 2013 wurde im Rahmen der Arbeitsgruppentreffen der DKPM-Arbeitsgruppe „Psychometrie und Psychodiagnostik“ in einem mehrstufigen Verfahren ein Kriterienkatalog zum Vergleich und zur Auswahl psychometrischer Verfahren für die psychosomatische und psychotherapeutische Versorgung und Forschung entwickelt und konsentiert. Die Entwicklungsschritte werden im Folgenden genauer erläutert.

Schritt 1 – Diskussion bestehender Testbeurteilungssysteme

Es fand zunächst eine Sichtung und Diskussion der folgenden Testbeurteilungssysteme statt: Committee on Test Affairs Netherlands (COTAN) [10], Standards of Educational and Psychological Testing (APA) (http://teststandards.org/), EFPA-Modell [3] und das Testbeurteilungssystem des Testkuratoriums der Föderation deutscher Psychologievereinigungen (TBS-TK) [8]. Ziel war es, sich einen Überblick über die Entwicklung, die Zusammenstellung der Kriterien, die Art der Beurteilung der Kriterien (z. B. Formalisierungsgrad), die Durchführung der Beurteilung und die Veröffentlichung der Ergebnisse zu verschaffen, um dann zu entscheiden, wie der Kriterienkatalog des DKPM entwickelt werden soll.

Schritt 2 – Entwicklung, Diskussion und erste Konsentierung von Kriterien

Aufbauend auf den Erkenntnissen des ersten Schrittes wurden Kriterien definiert und möglichst konkret operationalisiert, um nicht nur Experten, sondern auch Anwender in die Lage zu versetzen, den Kriterienkatalog zu nutzen und gleichzeitig eine hohe Objektivität, Reliabilität und Validität der Testbeurteilung zu erreichen.

Schritt 3 – Pilottestung des Kriterienkataloges

Der Kriterienkatalog wurde dann anhand sehr bekannter und etablierter Verfahren (Beck Depressions-Inventar II (BDI-II) [11]; Allgemeine Depressionsskala (ADS) [12]; Depressionsmodul des Patient Health Questionnaire (PHQ-9) [13] geprüft. Im Mittelpunkt standen dabei Verständlichkeit, Anwendbarkeit und Praktikabilität der definierten Gütemaßstäbe der einzelnen Kriterien. Dazu beurteilten jeweils 2 Mitglieder der Arbeitsgruppe und 2 externe Testbeurteiler ein Verfahren. Die Beurteiler führten dazu eine umfassende Literaturrecherche durch, um (möglichst) alle verfügbaren Informationen zu den definierten Gütekriterien des Tests zu identifizieren. Die entsprechenden Quellen, die ihren Urteilen zugrunde lagen, wurden jeweils aufgeführt. Auf Basis der verfügbaren Evidenz wurden dann alle Kriterien beurteilt. Die Beurteilung der jeweiligen Tests erfolgte unabhängig voneinander durch die 4 Beurteiler. Die Ergebnisse und die Erfahrungen bei der Anwendung des Kriterienkataloges wurden in der nächsten Arbeitsgruppensitzung von den Beurteilern dargestellt und in der Gruppe diskutiert.

Schritt 4 – Überarbeitung, erneute Testung und Verabschiedung des Kriterienkataloges

Anhand der Erfahrungen aus der Pilottestung wurde der Kriterienkatalog noch einmal überarbeitet, erneut getestet und von der Arbeitsgruppe verabschiedet. Die wichtigsten Diskussionspunkte und Entscheidungen sind im Ergebnisteil dargestellt. So wurde etwa in der Pilottestung offenbar, dass bislang nicht klar geregelt war, ob ausschließlich Studien zur deutschen Übersetzung des Instrumentes berücksichtigt werden sollen oder ob auch die psychometrischen Befunde zu den Originalskalen berücksichtigt werden sollen. Nach den Erfahrungen aus der Pilottestung wurde dann die weiter unten beschriebene Regelung getroffen.

Der Vorstand des Deutschen Kollegiums für Psychosomatische Medizin (DKPM) gab daraufhin seine Zustimmung, den Kriterienkatalog als Empfehlung des DKPM zu publizieren.

Ergebnisse

Der Kriterienkatalog

Der Kriterienkatalog ist in **Tab. 1** dargestellt. Die Kriterien sind zunächst in 5 Oberkategorien unterteilt: Reliabilität, Objektivität, Referenzgruppen, Validität und Anwendung. Zu den Oberkategorien wurden dann jeweils konkrete Unterkriterien definiert und diese möglichst genau operationalisiert. Die Operationalisierung definiert konkrete Zielkriterien zur Erreichung von 0, 1, 2 oder 3 Sternen (*), wobei die Bedingungen für die Vergabe explizit beschrieben sind. 3* sind bis auf wenige Ausnahmen als vollständige Erfüllung bzw. optimale Erreichung des Kriteriums zu verstehen und 0* als unzureichende Erfüllung des Merkmals bzw. das Fehlen an fundierten Informationen zur Einschätzung des Kriteriums. Es ist zudem möglich, halbe Punkte zur feineren Differenzierung zu vergeben. Es ist nicht vorgesehen, die Bewertungen der verschiedenen Kriterien zu einem Summenwert zu addieren, da keine Äquivalenz der Bedeutung der verschiedenen Kriterien für unterschiedlichste Anwendungsfelder anzunehmen ist (so ist etwa bei einem Instrument, das ein eigentlich stabiles Persönlichkeitsmerkmal misst, von einer geringeren Änderungssensitivität auszugehen als bei einem Instrument, das ein kurzfristig stärker fluktuierendes Merkmal wie z. B. Stimmung misst). Die Skalierung ist als Ordinalskala zu verstehen. Der Anwender ist aufgefordert, vor der Beurteilung zu definieren, welche Kriterien aus seiner Sicht für den intendierten Einsatz von Bedeutung sind, um dann Instrumente auswählen zu können, die diesem Einsatzgebiet möglichst gut entsprechen. Zur Beurteilung der *internen Konsistenz* sollen möglichst Informationen aus deutschsprachigen Studien herangezogen werden, um sich tatsächlich auf die deutsche Version des betreffenden Instruments zu beziehen, da es möglicherweise Unterschiede zwischen Übersetzung und Originalversion geben kann bzw. psychometrische Befunde nicht ohne Weiteres übertragbar sind [14]. Die Problematik der nicht transparenten Übersetzungsqualität und der kulturellen Äquivalenz wurde kritisch diskutiert. Da eine Überprüfung der Übersetzungen nicht möglich oder sehr aufwendig ist, soll auf diese verzichtet werden. Nicht nur bei diesem Beurteilungsaspekt ist es von besonderer Bedeutung zu prüfen, ob die Informationen zu einem Verfahren sich auf die gleiche Version (Itemzahl, -formulierung und Scoring usw.) beziehen, da zum Teil Kurz- und Langformen von Instrumenten existieren und insbesondere bei frei zugänglichen Verfahren verschiedene Versionen (Übersetzungen, Itemzusammensetzung und -zahl) im Umlauf sind. Gibt es keine deutschsprachigen Befunde, sollen internationale Befunde herangezogen werden. Auf die wichtigsten Diskussionspunkte und die konsentierten Empfehlungen/Vorgehensweisen wird hier genauer eingegangen.

Reliabilität

Unter der Oberkategorie Reliabilität wird in diesem Kriterienkatalog nur die *Interne Konsistenz* eingeordnet. Besteht ein Instrument aus mehreren Skalen, soll ein Gesamturteil über alle Skalen gebildet werden. Gibt es einzelne Skalen, die deutlich von diesem Urteil abweichen, soll diese Information in Form eines „Warnhinweises" gegeben werden (Bsp.: Childhood Trauma Questionnaire, Skala „Körperliche Vernachlässigung", [15, 16]). Auch wenn die Test-Retest-Reliabilität ein klassisches Testgütekriterium ist, wurde sie in diesen Katalog nicht aufgenommen, weil für dieses Kriterium bei unterschiedlichen und klinisch häufigen Fragestellungen durchaus unterschiedliche Anforderungen gelten. In entsprechend angepasster Form geht dieser Aspekt im Abschnitt Validität in die Einschätzung der Veränderungssensitivität ein.

Objektivität

In der Oberkategorie Objektivität sind die Itemverständlichkeit, 3 Aspekte der Durchführungsobjektivität und Angaben zur Fairness enthalten.

Die *Itemverständlichkeit* kann mit 1–3 Sternen (*) bewertet werden, wobei 3 Sterne nur dann vergeben werden können, wenn die Itemverständlichkeit empirisch geprüft wurde (z. B. durch Cognitive-Survey-Techniken).

Zunächst wurde diskutiert, die *Durchführungsobjektivität* als ein Kriterium abzubilden. Da die verschiedenen Aspekte der Durchführungsobjektivität jedoch deutlich variieren können, sind diese Aspekte (Instruktion, Auswertung, Interpretation) getrennt dargestellt und können jeweils mit einem * bewertet werden.

Der Aspekt *Fairness* soll abbilden, ob ein Test für verschiedene Zielgruppen (Geschlecht, Alter, ethnische Herkunft, Bildung usw.) gleichermaßen gut einsetzbar ist. Dies wird heutzutage in Studien üblicherweise auf Basis der probabilistischen Testtheorie durch Untersuchungen zum Differential Item Functioning (DIF) [17, 18] untersucht. Da sich dies nicht ohne weiteres auf die klassische Testtheorie anwenden lässt und Studien zu DIF noch relativ selten sind, wurde das Kriterium breiter definiert, indem auch gruppenspezifische Normen als Beitrag zur Fairness berücksichtigt werden.

Referenzgruppen

Die Verfügbarkeit von Referenzgruppen wurde als eigene Oberkategorie in den Katalog aufgenommen. Es wird zwischen Allgemeinbevölkerungs- und klinischen Referenzgruppen unterschieden. Bei den Allgemeinbevölkerungsgruppen wird dabei nach Größe und Repräsentativität der verfügbaren Stichproben unterschieden. Bei den klinischen Referenzgruppen wird vor allem auf konstruktrelevante und homogene Stichproben Wert gelegt.

Validität

Da es für viele etablierte Instrumente widersprüchliche Befunde zur faktoriellen Struktur gibt (z. B. [19]) wurde die *faktorielle Validität* als eigenes Kriterium in den Katalog aufgenommen. Es wurde intensiv diskutiert, wie mit sich widersprechenden Befunden umzugehen ist. Entscheidungsrelevant sollen hier

Tab. 1 DKPM-Kriterienkatalog zur Beurteilung psychodiagnostischer Selbstbeurteilungsinstrumente.

Bewertungskategorien		Bewertungskriterien				Anmerkungen
RELIABILITÄT	**Interne Konsistenz**	*** ≥ 0,80 (>0,90)	** 0.70–0.79	* 0,60–0,69	○ <0,60	bei mehreren Studien gilt Median: SP von n ≥ 100
	Item-Verständlichkeit	*** Item-Verständlichkeit wurde empirisch überprüft (z. B. mittels Cognitive Survey-Techniken)	** alle Items klar verständlich, eindeutig	* wenige Items sind missverständlich	○ Großteil der Items kompliziert, missverständlich	z. B. doppelte Verneinung, mehrere Aspekte in einem Item
OBJEKTIVITÄT	**Instruktion**			* klar und prägnant	○ nicht vorhanden oder missverständlich	klar, prägnant
	Auswertung			* klarer Auswertungsalgorithmus	○ nicht verfügbar	inkl. Angaben zum Umgang mit Missings
	Interpretation			* Richtlinien zur Interpretation der Skalenwerte	○ nicht verfügbar	
	Fairness	*** DIF Test durchgeführt & ggf. Empfehlung zum Umgang mit DIF	** empirische Untersuchung möglicher gruppenspezifischer Effekte auf das Item-Antwortverhalten wurden untersucht	* keine offensichtlichen Verstöße (per Augenschein)	○ keine Information oder offensichtliche Verstöße	beobachtete Verstöße je nach Konstrukt bzgl. folgender Gruppen: klinische Gruppen, Alter, Sprachen, Geschlecht, ethnische Gruppen
REFERENZ-GRUPPEN	**Allgemeinbevölkerung**	*** Repräsentative Stichprobe (n > 1 000)	** große Stichprobe (n > 1 000)	* kleine Stichprobe (>300–1 000)	○ keine Information	
	Klinische Gruppen	*** hohe Qualität (klinisch homogene) & ≥ 5 Gruppen à n >50	** hohe Qualität (klinisch homogene) & 2–4 Gruppen à n >50	* eine konstruktrelevante (klinisch homogene) Gruppe	○ keine Information oder keine Gruppe	Größe, Anzahl der SP, Qualität der SP (transparente Stichprobenrekrutierung und konstruktrelevant)
VALIDITÄT	**Faktorielle Validität**	*** CFA repliziert die angenommene Skalenstruktur im Wesentlichen & faktorielle Invarianz in mindestens 2 Subgruppen	** CFA repliziert die angenommene Skalenstruktur im Wesentlichen	* Faktorstruktur in EFA wird in unabhängiger SP im Wesentlichen repliziert	○ keine Information oder unklare Faktorstruktur in konstruktrelevanten Stichproben	Ladung auf Faktor >.40; Analysen in adäquater SP-Größe (state of the art)
	Inhaltsvalidität	*** Items erfassen Konstrukt hinreichend	** Items erfassen Konstrukt weitgehend		○ Items erfassen Konstrukt unzureichend	
	Konvergente/Diskriminante Validität	*** gute konvergente & diskriminante Validität in MTMM	** gute konvergente & diskriminante V	* gute konvergente V	○ keine Information oder schlechte konvergente V	
	Known Groups Validität	*** Kriteriumsvalidität belegt anhand State of the Art – anerkannter Außenkriterien in mindestens 2 unabhängigen Studien	** Kriteriumsvalidität belegt anhand State of the Art – anerkannter Außenkriterien in einer Studie	* Kriteriumsvalidität belegt anhand klinischer Diagnose oder Schweregrad	○ keine Information	
	Prädiktive Validität	*** Mindestens ein Außenkriterium wird in mehreren Studien vorhergesagt	** Mindestens ein Außenkriterium wird in einer Studie vorhergesagt		○ keine Information	
	Veränderungssensitivität	*** kontrollierte Studien zeigen differentielle Veränderung in verschiedenen Interventionsgruppen	** Studien zeigen Veränderung in verschiedenen Interventionsgruppen	* Studien zeigen Veränderungssensitivität	○ nicht veränderungssensitiv	

Tab. 1 Fortsetzung.

	Minimal Clinical Important Difference			* Studien berichten klinisch relevante, kritische Differenzen (MCID)	○ keine Information	
ANWENDUNG	**Ökonomie (Bearbeitungszeit)**	* * * < 1 min	* * 1–3 min	* 3–8 min	○ ≥ 8 min	Zeit in klinischer SP pro Konstrukt
	Simplizität der Auswertung	* * * leicht (einfache Addition, keine Umpolung, wenig Items bzw. verfügbare, leicht anwendbare PC-Auswertung)	* * mittel (einfache Umrechnungen, leicht handhabbar, geringe Fehlerwahrscheinlichkeit)	* umständlich	○ schwer und fehleranfällig	gilt für klinische Praxis
	Simplizität der Interpretation	* * * instrumenten-unabhängige Darstellung zur leichten Interpretation (z. B. t-Werte, Prozentränge)	* * instrumenten-gebundene Interpretationshilfen (z. B. Cut-off oder Vergleichsgruppenwerte)		○ keine Auswertungshilfen	
	Internationalität	* * * state-of-the-Art Übersetzung in mehrere Sprachen	* * Übersetzung in mehreres Sprachen	* Übersetzung in eine Sprache	○ keine Übersetzung	
	Verfügbarkeit	* * * leichte Zugänglichkeit zum vollständigen Material (Test, Auswertungsanleitung, Manual)	* * leichte Zugänglichkeit zum Test	* mittelschwere Zugänglichkeit zum Test	○ schwere Zugänglichkeit oder Fehlen wesentlicher Materialien	
	Akzeptanz			* kein Hinweis auf Akzeptanzprobleme	○ bekannte Akzeptanzprobleme	wenn Probleme vorhanden, dann explizieren

SP = Stichprobe; MTMM = Multi-Trait-Multi-Method; CFA = Confirmatory Factor Analysis; DIF = Differential Item Functioning; EFA = Exploratory Factor Analysis; MCID = Minimal Clinical Important Difference

Befunde aus relevanten (klinischen) Gruppen sein. Weichen die Befunde für verschiedene Gruppen systematisch voneinander ab (z. B. findet sich eine stabile faktorielle Struktur in klinischen Stichproben, die sich in Bevölkerungsstichproben so nicht zeigen lässt), soll dies in den Kommentaren dargelegt werden.

Zur *Inhaltsvalidität* wurde die Begrifflichkeit „… erfasst ein Konstrukt hinreichend …“ gewählt, weil nicht alle Instrumente alle Aspekte eines Konstruktes erfassen und das möglicherweise auch gar nicht intendiert war. Darüber hinaus sind *konvergente/diskriminante Validität, Kriteriumsvalidität* und *prädiktive Validität* als eigene Kriterien dargestellt. Insbesondere die prädiktive Validität ist für die klinische Forschung und Anwendung von großer Bedeutung, weil es dabei um die Frage geht, ob man mit den Messwerten z. B. Therapieerfolg vorhersagen kann.

Die *Veränderungssensitivität* spielt in der psychosomatischen und psychotherapeutischen Diagnostik und Forschung eine wichtige Rolle, weil damit Therapieprozesse und -ergebnisse abgebildet werden können [20]. Grundsätzlich soll mit diesem Kriterium abgebildet werden, ob ein Instrument zur Veränderungsmessung geeignet ist. In Testhandbüchern ist das leider nicht immer dargestellt, weshalb eine eigene Recherche dazu oft erforderlich ist. Die Befunde zur Veränderungssensitivität sollen zielgruppenspezifisch dargestellt werden. Um die Aussagen zur Veränderungsmessung noch zu spezifizieren, wurde die Verfügbarkeit von Angaben zur *Minimal Clinical Important Difference (MCID)* als weiteres Kriterium aufgenommen [21, 22]. Maximal ein * wird vergeben, wenn Informationen zur MCID verfügbar sind.

Anwendung

In diesem letzten Oberkapitel sind verschiedene Aspekte der praktischen Anwendung eines Verfahrens subsumiert. Die *Bearbeitungszeit (Ökonomie)* sollte möglichst kurz sein. Die Einschätzung dieser Bearbeitungszeit soll dabei pro Konstrukt erfolgen (z. B. Patient-Health-Questionnaire, Einschätzung hier für das Depressionsmodul PHQ-9 und nicht für den gesamten PHQ). Darüber hinaus sollen die *Simplizität der Auswertung* (in der klinischen Praxis am Einzelfall) und die *Simplizität der Interpretation* beurteilt werden.

Der Aspekt *Internationalität* subsumiert Fragen nach der Verfügbarkeit verschiedener Sprachversionen zum internationalen Vergleich von Befunden und der Qualität der Übersetzung, wenn es sich um ein Instrument handelt, das nicht im Deutschen entwickelt wurde. Das Vorgehen bei einer qualitativ hochwertigen Übersetzung ist bei Wild et al. [23] ausführlich beschrieben.

Zum Aspekt der *Verfügbarkeit* wurde intensiv diskutiert, ob hier frei zugängliche Verfahren priorisiert werden sollen. Da sich das Feld der Open-Access-Testverfahren erst langsam entwickelt und nicht bei einem Verlag publizierte Verfahren zum Teil schwer zugänglich bzw. Informationen zum Test nicht gut dokumentiert sind, wurde diese Strategie nicht weiter verfolgt. Sollten sich neuere Initiativen zur Förderung Open Access-basierter Publikationen von Testverfahren (d. h. Testmaterial inkl. wissenschaftlich fundiertem Handbuch) in der Zukunft etablieren (z. B. das Open Access Testportal www.psychometrikon.de [24]) und sich damit die Verfügbarkeit und Qualität von Informationen zu frei publizierten Verfahren verbessern, kann eine solche Priorisierung neu diskutiert werden.

Das letzte zu beurteilende Kriterium ist die *Akzeptanz*. Hier gilt es einzuschätzen, ob bekannte Akzeptanzprobleme vorliegen. Diese sollen ggf. in den Kommentaren dokumentiert werden.

Anwendung des Kriterienkataloges

Wie bereits oben beschrieben ist der Katalog entwickelt worden, um die Bewertung und Auswahl von Testverfahren zu unterstützen und zu formalisieren. Der Testkatalog ist deskriptiv zu verstehen und soll eine überblicksartige Bewertung der Stärken und Schwächen sowie der Einsatzmöglichkeiten eines Tests ermöglichen. Er soll ausschließlich auf von Patienten berichtete Ergebnismaße (Patient-reported Outcomes) angewendet werden. Module eines Instrumentes, die verschiedene Konstrukte erfassen, müssen einzeln bewertet werden. Es ist zunächst vorgesehen, dass durch Mitglieder der DKPM-Arbeitsgruppe und durch andere Experten Testbeurteilungen mit dem Kriterienkatalog vorgenommen und der Öffentlichkeit zugänglich gemacht werden. Folgendes Vorgehen wurde dafür von der Arbeitsgruppe konsentiert:

1. Auswahl eines oder mehrerer Verfahren
2. Auswahl von mindestens 2 unabhängigen Beurteilern
3. Literaturrecherche durch die Beurteiler (Informationen zu Datenbanken, Zeitraum und Suchworten sollen in der Testbeurteilung angegeben werden; die Beurteiler sollen möglichst eine identische Daten- und Publikationsgrundlage benutzen)
4. Einschätzung des Verfahrens anhand des Kriterienkataloges durch die Beurteiler
5. Zusammenführen und Konsentieren der beiden Einschätzungen
6. Rückfrage zu offenen Fragen an den Autor des deutschen Verfahrens
7. Erstellung eines begleitenden „Steckbriefes“ mit Informationen zu Konstrukt, Zielgruppe, Einsatzgebiete, Itemzahl, Bearbeitungszeit, Kosten, Erhebungsmodus (elektronisch, Paper-Pencil-Erhebung), Bezugsquelle, Verfügbarkeit eines Manuals und Internationalität
8. Veröffentlichung der Ergebnisse

Unabhängig von diesem aufwendigen Vorgehen kann der Kriterienkatalog von Praktikern auf einfache Weise als Orientierungshilfe für die Auswahl von Verfahren genutzt werden. Nach einer Festlegung, welche der Gütekriterien für einen bestimmten Einsatzzweck von Bedeutung sind, kann der Praktiker dann einen Test nach diesen Kriterien anhand des Kataloges auswählen bzw. verschiedene Tests miteinander vergleichen. Schließlich kann der Kriterienkatalog auch bei der Planung eigener Erhebungen von Nutzen sein.

Diskussion

Die Arbeitsgruppe „Psychometrie und Psychodiagnostik“ des Deutschen Kollegiums für Psychosomatische Medizin (DKPM) hat zwischen 2011 und 2013 einen Kriterienkatalog zur Beurteilung und Auswahl psychometrischer Instrumente für die psychosomatische und psychotherapeutische Forschung und Praxis entwickelt. Ausgangspunkt für diese Entwicklung war die Feststellung, dass die bereits vorliegenden Testbeurteilungssysteme nicht speziell für die psychosomatische und psychotherapeutische Forschung und Praxis konzipiert wurden und dadurch wesentliche diesbezügliche Anforderungen nicht erfüllen. Entwicklungsschritte des Kriterienkataloges und der konsentierte Katalog mit Erläuterungen wurden dargestellt. Ziel war es, einen leicht anwendbaren Katalog zu entwickeln, mit dem es in Forschung und Praxis tätigen Kollegen ermöglicht wird, mit vertretbarem Aufwand ein Verfahren zu beurteilen und geeignete Verfahren für verschiedenste Anwendungsgebiete evidenzbasiert auszuwählen. Das Kollegium für Psychosomatische Medizin (DKPM) empfiehlt die Beurteilung und Auswahl von klinischen Selbstbeurteilungsinstrumenten nach dem hier vorgestellten Kriterienkatalog.

Mit dem Kriterienkatalog liegt nun ein übersichtliches und gut formalisiertes Beurteilungssystem vor, welches durch ein Expertengremium diskutiert und konsentiert wurde. Eine nach definierten Kriterien verfasste und damit transparente sowie systematische und umfassende Testbeurteilung wird damit erleichtert, und es besteht zudem die Möglichkeit, verschiedene Tests zu vergleichen. Die Arbeitsschritte für eine belastbare Testbeurteilung sind ebenfalls dargestellt. Der Kriterienkatalog dient aber auch einer praktikablen Testauswahl für Forschung und Klinik. Dabei ist es notwendig, dass wichtige Eigenschaften eines auszuwählenden Verfahrens zunächst definiert werden. So dürfte es für die Beurteilung von psychotherapeutischen Behandlungen wichtig sein, dass die grundsätzlichen Testgütekriterien als gut oder sehr gut eingeschätzt werden. Darüber hinaus sind sicherlich die Veränderungssensitivität und die prädiktive Validität von großer Bedeutung, während etwa Internationalität oder die Verfügbarkeit von Allgemeinbevölkerungsnormen nicht ganz so wichtig sein dürften. Für den täglichen Einsatz in der Praxis sind zudem Aspekte der Anwendung eines Tests, wie z. B. Simplizität der Anwendung und Auswertung, sowie Anwendungsökonomie Gesichtspunkte, denen viel Bedeutung beigemessen wird. In einem Forschungskontext hingegen sind Internationalität oder die Verfügbarkeit von Angaben zu „State-of-the-Art“-Kriterien wie bspw. MCID von Bedeutung. Mit diesen Beispielen soll verdeutlicht werden, dass die Anwender des Kataloges aufgefordert sind, sich über die Gewichtung der verschiedenen Kriterien für ihre Absichten vorab Gedanken zu machen und Tests entsprechend dieser priorisierten Kriterien auszuwählen.

Ein weiterer wichtiger Aspekt ist die Verfügbarkeit eines Testverfahrens. Die aufwendige Entwicklung und testpsychologische Evaluation eines Testverfahrens erfolgt seit Jahrzehnten überwiegend an universitären Einrichtungen und ist mitunter ein sehr langwieriger und kostenintensiver Prozess (u. a. durch die Rekrutierung von Probanden für die Normierung). Publiziert werden die Verfahren anschließend in einigen wenigen Testverlagen, die diese dann gegen Nutzungsgebühren verkaufen. Die Rechte an dem Verfahren gehen vollständig an den Verlag über. Die Verlage sorgen für die Produktion und Verbreitung der Instrumente. Die teils erheblichen Kosten für den Erwerb eines Testverfahrens können allerdings auch eine Hürde darstellen, und die Verbreitung innovativer Neuentwicklungen oder psychometrischer Überarbeitungen bestehender Verfahren bremsen. Eine „Open Access“-basierte Publikation, wie sie seit einigen Jahren im Bereich wissenschaftlicher Fachpublikationen stattfindet, würde diese Hürde abbauen helfen und die Verfügbarkeit von Testverfahren für Forschung und Praxis verbessern. Umfragen unter wissenschaftlich tätigen Psychologen zur Einstellung zu Open Access-basierter Publikation zeigen eine positive Grundhaltung [25]. Eine solche positive Einstellung wird sicherlich auch dadurch gefördert, dass Open-Access-Beiträge häufiger zitiert werden und schneller eine größere Aufmerksamkeit erlangen [26–28]. Entsprechend hat sich die Zahl der beim Directory of Open Access Journals (www.doaj.org) gelisteten Open Access Zeitschriften innerhalb von 6 Jahren von 2 140 in 2007 auf 8 341 in 2013 mehr als vervierfacht [29].

Per Open Access veröffentlichte Testverfahren finden sich gegenwärtig noch zumeist auf der Homepage der Autoren oder entsprechender Forschungsprojekte. Eine Entwicklung hin zu einer

systematischen, verlagsähnlichen Open-Access-Publikation als Alternative zu einer verlagsbasierten Publikation wurde aber in jüngerer Zeit angestoßen, etwa durch das Testarchiv des Zentrums für psychologische Information und Dokumentation (ZPID) oder das Open Access Testportal Psychometrikon (www.psychometrikon.de), das eine Testpublikation nach wissenschaftlichen Standards inklusive Peer-Review-Verfahren ermöglicht [24]. Eine solche Entwicklung ist sicher wünschenswert, weil sie die Verfügbarkeit von hochwertigen Testverfahren für Forschung und Praxis erhöht. Der hier vorgestellte Kriterienkatalog kann dazu beitragen, den systematischen Vergleich ggf. veralteter, aber gut etablierter und neuerer, psychometrisch besserer Instrumente zu erleichtern und damit die Begründung für eine Anpassung des diagnostischen Vorgehens an den aktuellen Forschungsstand überzeugender und transparenter zu gestalten.

Ziel der Arbeitsgruppe war es, neben der Entwicklung und Publikation des Kriterienkataloges, auch Testbesprechungen anhand des Katalogs zur Verfügung zu stellen. Die Abfassung publizierbarer Testbeurteilungen wird ähnlich wie bei einem systematischen Review erfolgen und bleibt trotz des gut formalisierten Kataloges aufwendig. In jedem Fall aber kann der Kriterienkatalog auch ohne das Ergebnis einer formalisierten Testbeurteilung als Handreichung und Orientierungshilfe verwendet werden.

Der Kriterienkatalog wurde für die Bewertung und Auswahl von Selbstbeurteilungsverfahren entwickelt und konsentiert. Da in der klinischen Praxis und Forschung häufig auch Fremdbeurteilungsinstrumente eingesetzt werden, wäre es von Interesse, auch für diese Verfahren einen entsprechenden Katalog zur Verfügung zu stellen. Der Kriterienkatalog ist sicherlich weitgehend auf die Bewertung von Fremdbeurteilungsverfahren übertragbar. Da der Katalog dafür nicht entwickelt und konsentiert wurde, muss eine solche Übertragung jedoch geprüft und diskutiert werden. Darüber hinaus werden im Katalog ausschließlich psychometrische Gütekriterien und Anwendungsaspekte berücksichtigt. Grundsätzlich gilt es natürlich auch, konzeptuell-theoretische Aspekte bei der Auswahl von Konstrukten und dazugehörigen Erhebungsinstrumenten zu beachten. Insbesondere in der Psychotherapieforschung stellt sich die Frage nach der Güte der theoretischen Einbettung, das heißt z. B. ob die zu erfassenden Konstrukte aus etablierten und elaborierten Theorien abgeleitet wurden und ob deren Operationalisierung passend ist. Diese Fragestellungen können in einem solchen Kriterienkatalog nicht abgebildet werden, stellen aber ohne Zweifel einen wichtigen Qualitätsaspekt innovativer Forschung dar.

Eine wesentliche Aufgabe besteht nun in der Dissemination und Implementierung des Kataloges in Forschung und Praxis. Aus diesem Grund soll dieser auf verschiedenen Fachtagungen vorgestellt werden. Des Weiteren ist vorgesehen, den Katalog zu testen und Testbesprechungen zu veröffentlichen. Es wurde diskutiert, ob eine Programmierung eines frei zugänglichen Online-Tools des Kataloges oder Ergebnisse seiner Anwendung auf bekannte Verfahren wünschenswert wäre, welches es ermöglicht, Tests relativ unkompliziert zu vergleichen und Testeigenschaften als Profile darzustellen. Zur Visualisierung der Testeigenschaften wird eine Profildarstellung empfohlen, die einen unmittelbaren Überblick zu den Ergebnissen liefert. Eine Publikation zur praktischen Anwendung des Kriterienkataloges mit einer entsprechenden Profildarstellung ist in Vorbereitung.

Ob die intendierte Verbreitung und Implementation des Kriterienkataloges zu positiven Veränderungen in Forschung und Praxis führt, müsste im Sinne einer Implementationsevaluation untersucht werden.

Die Entwickler des Kriterienkataloges hoffen, dass der Kriterienkatalog in jedem Fall eine kritische Reflexion über die Auswahl von Instrumenten in Forschung und Praxis anstößt und eine Hilfe bei der evidenzbasierten Auswahl von Verfahren darstellt.

Fazit für die Praxis

Mit dem Kriterienkatalog zur Beurteilung und Auswahl psychodiagnostischer Selbstbeurteilungsinstrumente für die Praxis und Forschung in Psychosomatik und Psychotherapie wird ein praktisches Verfahren zur Verfügung gestellt. Es kann auch zu einem systematischen Vergleich verschiedener Verfahren genutzt werden.

Interessenkonflikt: Die Autoren geben an, dass kein Interessenkonflikt besteht.

Institute

[1] Abteilung für Medizinische Psychologie und Medizinische Soziologie, Universität Leipzig
[2] Institut für Medizinische Psychologie und Medizinische Soziologie, Uniklinik der RWTH Aachen
[3] Klinik und Poliklinik für Psychosomatische Medizin und Psychotherapie, Klinikum rechts der Isar, Technische Universität München
[4] Psychosomatische Medizin und Psychotherapie, Universitätsklinikum Hamburg-Eppendorf & Schön Klinik Hamburg Eilbek
[5] Klinik für Psychosomatische Medizin und Psychotherapie, Klinikum München-Harlaching
[6] Internationale Psychoanalytische Universität, Berlin
[7] Institut und Poliklinik für Medizinische Psychologie, Universitätsklinikum Hamburg-Eppendorf
[8] Institut für Psychologie, Alpen-Adria Universität Klagenfurt, Österreich
[9] Abteilung für Psychosomatische Medizin, Universität Regensburg
[10] Fachbereich Angewandte Humanwissenschaften, Rehabilitationspsychologie, Hochschule Magdeburg-Stendal
[11] Charite-Universitätsmedizin Berlin, Medizinische Klinik mit Schwerpunkt Psychosomatik

Literatur

1 *Wahl I, Meyer B, Loewe B et al.* Die Erfassung der Lebensqualität in der Psychotherapieforschung. Klinische Diagnostik und Evaluation 2010; 4–21
2 *Brähler E, Schumacher J, Strauß B.* Diagnostische Verfahren in der Psychotherapie. Göttingen: Hogrefe 2004
3 *Lindley P, Bartram D, Kennedy N.* EFPA review model for the description and evaluation of psychological tests – Version 3.42. Brussels: EFPA Standing Committee on Tests and Testing 2008
4 *Evers A, Hagemeister C, Hostmaelingen A et al.* EFPA Review Model for the description and evaluation of psychological and educational tests. (Tech. Rep. Version 4.2.6). Brussels: European Federation of Psychology Associations 2013
5 *Föderation der Deutschen Psychologenverbände.* Beschreibung der einzelnen Kriterien für die Testbewertung. Diagnostica 1986; 32: 358–360
6 *Kersting M.* Zur Beurteilung der Qualität von Tests: Resümee und Neubeginn. Psychologische Rundschau 2006; 57: 243–253
7 *Kersting M, Heyse H.* Anforderungen an die Qualität der Verfahren. In: Hornke LF, Winterfeld U.editors. Eignungsbeurteilungen auf dem Prüfstand: DIN 33430 zur Qualitätssicherung. Heidelberg: Spektrum Akademischer Verlag 2004; p 43–54
8 *Testkuratorium der Förderation Deutscher Psychologievereinigungen.* TBS-TK. Testbeurteilungssystem des Testkuratoriums der Föderation Deutscher Psychologenvereinigungen. Stand und Perspektiven. Report Psychologie 2006; 31: 492–500
9 *Löwe B, Rose M, Wahl I et al.* Psychometrie und Psychodiagnostik in Psychosomatik, Psychotherapie und Medizinischer Psychologie – Bericht zum zweiten Arbeitstreffen der DKPM-Arbeitsgruppe „Psychometrie und Psychodiagnostik". Psychother Psychosom Med Psychol 2011; 61: 334–336
10 *Evers A.* The Revised Dutch Rating System for Test Quality. International Journal of Testing 2001; 1: 155–182

11 *Hautzinger M*, *Keller FKC*. Beck-Depressions-Inventar (BDI-II). Revision. Frankfurt/Main: Harcourt Test Services 2006
12 *Herrmann-Lingen C*, *Buss U*, *Snaith RP*. HADS-D. Hospital Anxiety and Depression Scale – Deutsche Version. Göttingen: Hogrefe 1995
13 *Löwe B*, *Grafe K*, *Zipfel S et al.* Diagnosing ICD-10 depressive episodes: Superior criterion validity of the patient health questionnaire. Psychother Psychosom 2004; 73: 386–390
14 *Glaesmer H*, *Braehler E*, *von Lersner U*. Kultursensible Diagnostik in Forschung und Praxis – Stand des Wissens und Entwicklungspotentiale. Psychotherapeut 2012; 57: 22–28
15 *Klinitzke G*, *Romppel M*, *Häuser W et al.* The German Version of the Childhood Trauma Questionnaire (CTQ) – Psychometric Characteristics in a Representative Sample of the General Population. Psychother Psychosom Med Psychol 2012; 62: 47–51
16 *Wingenfeld K*, *Spitzer C*, *Mensebach C et al.* The German Version of the Childhood Trauma Questionnaire (CTQ): Preliminary Psychometric Properties. Psychother Psychosom Med Psychol 2010; 60: 442–450
17 *Embretson SE*, *Reise SP*. Item Response Theory for Psychologists. Mahwah, New Jersey, United States: L. Erlbaum Associates 2000
18 *Forkmann T*, *Gauggel S*, *Spangenberg L et al.* Dimensional assessment of depressive severity in the elderly general population: Psychometric evaluation of the PHQ-9 using Rasch Analysis. J Affect Disord 2013; 148: 323–330
19 *Romppel M*, *Braehler E*, *Roth M et al.* What is the General Health Questionnaire-12 assessing? Dimensionality and psychometric properties of the General Health Questionnaire-12 in a large scale German population sample. Compr Psychiatry 2013; 54: 406–413
20 *Igl W*, *Zwingmann C*, *Faller H*. Änderungssensitivität. Rehabilitation 2005; 44: 100–106
21 *Rouquette A*, *Blanchin M*, *Sebille V et al.* The minimal clinically important difference determined using item response theory models: an attempt to solve the issue of the association with baseline score. J Clin Epidemiol 2014; 67: 433–440
22 *Revicki D*, *Hays RD*, *Cella D et al.* Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol 2008; 61: 102–109
23 *Wild D*, *Grove A*, *Martin M et al.* Principles of good practice for the translation and cultural adaptation process for patient-reported outcomes (PRO) measures: Report of the ISPOR Task Force for Translation and Cultural Adaptation. Value Health 2005; 8: 94–104
24 *Forkmann T*, *Gauggel S*. Freier Zugang zu psychodiagnostischen Instrumenten!. Psychiatrische Praxis 2013; 40: 102–103
25 *Frey A*, *Herzberg PY*. Publishing in Psychology: a description of the current situation in Germany. Psychol Sci Q 2009; 51: 160–166
26 *Eysenbach G*. The open access advantage. J Med Internet Res 2006; 8 doi:10.2196/jmir.8.2.e8
27 *Hardisty DJ*, *Haaga DA*. Diffusion of treatment research: Does open access matter? J Clin Psychol 2008; 64: 821–839
28 *Harnad S*, *Brody T*. Comparing the impact of open access (OA) vs. non-OA articles in the same journals. D-Lib Magazine 2004; 10: doi:10.1045/june2004-harnad
29 *Mey G*, *Mruck K*. Open Access: Auswirkungen einer Informationskrise als Chance für die Information. Journal für Psychologie 2007; 15: 1–17