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

Psychologische Rundschau, 57 (4), 243–253

Zur Beurteilung der Qualität von Tests: Resümee und Neubeginn

Martin Kersting

Zusammenfassung. Mit der steigenden Einsatzhäufigkeit und Akzeptanz von Tests steigt im deutschsprachigen Raum auch der Bedarf nach Informationen über und Beurteilungen von Tests. Das diesbezügliche Angebot kann weder quantitativ noch qualitativ überzeugen und bleibt hinter den in anderen Ländern erreichten Standards zurück. Der Artikel stellt anhand ausgewählter internationaler Beispiele (COTAN, EFPA) verschiedene Testinformations- und Beurteilungssysteme vor und diskutiert deren Vor- und Nachteile. Abschließend wird das neue, dreistufige System zur Information über und zur Beurteilung von Tests des Testkuratoriums dargestellt, bei dem die DIN 33430 Berücksichtigung findet. Auch die Gestaltung der notwendigen Institutionalisierung der Systemanwendung und -kontrolle wird skizziert.
Schlüsselwörter: Testbeurteilung, Testqualität, Qualitätssicherung, DIN 33430

Test evaluation: Summary and restart

Abstract. With rising acceptance and application of tests in the German-speaking countries there is also an increase in the need for information about these instruments. The range of available test information and evaluation is neither quantitatively nor qualitatively convincing and stays behind the standards reached in other countries. This article shows different approaches on the basis of selected international examples for test review systems (COTAN, EFPA) and discusses their pro and cons. Finally a new, three stepped test information and review system is presented, taking the German standard "DIN 33430" into consideration. Furthermore, the implementation of necessary institutionalizing for application and control of the test information and review is outlined.
Key words: test evaluation, test review systems, test use, quality standards, guidelines

Seit ca. 1990 ist die Nachfrage nach Tests[1] in Deutschland deutlich gestiegen (Wottawa, 2002, S. 3). Der Testmarkt ist allerdings weitgehend intransparent. Weder kann der Praktiker rasch und einfach erfahren, welche Tests wo und zu welchen Konditionen angeboten werden, noch kann er ohne weiteres unabhängige Bewertungen der Qualität der angebotenen Tests einsehen. Zwar gibt es eine Vielzahl von Informationen über und Beurteilungen von Tests in Büchern, Fachzeitschriften und Datenbanken (siehe S. 243), diese Informationen sind aber sowohl quantitativ (siehe S. 243) als oft auch qualitativ (siehe S. 245) unbefriedigend. Sie sind häufig unsystematisch aufbereitet und in ihrem Urteil beliebig, überzogen kritisch oder aber uneindeutig in der Bewertung. Die Informationen und Rezensionen werden verstreut publiziert, so dass ein zentraler Zugriff erschwert wird. Außerdem repräsentieren die vorhandenen Testrezensionen nicht die vorhandenen Tests: Trotz der absolut gesehen beachtlichen Anzahl an Testrezensionen liegen zu einigen in der Praxis häufig genutzten Tests keine Informationen und Rezensionen vor (siehe S. 243). Das Wirken einer steuernden, gestaltenden und kontrollierenden Institution ist nicht hinreichend erkennbar. Demgegenüber konnten sich in anderen Ländern, beispielsweise in England und in den Niederlanden, standardisierte Systeme zur Testbeurteilung etablieren (siehe S. 246). Diese Systeme und ihre Vor- und Nachteile werden im vorliegenden Artikel thematisiert. Anschließend wird das neue Testbeurteilungssystem des Testkuratoriums (TK) skizziert (siehe S. 250), bei dem die DIN 33430 (siehe S. 249) den Ausgangspunkt setzt. Der formale Ablauf der Testrezensionserstellung nach dem neuen System wird im vorletzten Abschnitt des Artikels beschrieben, bevor abschließend im letzten Abschnitt des Artikels ein Ausblick auf die Zukunft gewagt wird.

Zur Auffindbarkeit, Quantität und Repräsentativität deutschsprachiger Testrezensionen

Informationen über und Beurteilungen von Tests sind eine klassische Domäne der einschlägigen Fachzeitschriften. 1997 veröffentlichte die „Zeitschrift für Differentielle

[1] Unter dem Begriff „Test" werden im vorliegenden Artikel im Sinne der DIN 33430 (DIN, 2002, S. 23) alle praxiserprobten und wissenschaftlich abgesicherten Erkenntnismittel gefasst, die in standardisierter Weise eingesetzt, Hinweise auf Erleben und Verhalten der getesteten Personen liefern.

DOI: 10.1026/0033-3042.57.4.243

und Diagnostische Psychologie" (ZDDP) beispielsweise ein Sonderheft mit 25 Testrezensionen sowie eine Übersicht über die 229 bis dato erschienenen Testrezensionen (siehe Kubinger, 1997 sowie Punter & Kubinger, 2002).

Kritisch anzumerken ist, dass einige Fachzeitschriften den für Testrezensionen zur Verfügung stehenden Textumfang derart beschränken, dass eine ausgewogene und anregende Darstellung und Kritik kaum möglich ist. Darüber hinaus ist zu bezweifeln, dass in Fachzeitschriften publizierte Testrezensionen die Testanwender erreichen. So wird beispielsweise die „Diagnostica", die sich selbst als Informationsorgan über Tests versteht, nach Steck (1997) de facto nur von 7 % der Praktiker zu diesen Zwecken genutzt.

Nicht zuletzt durch das „Handbuch psychologischer und pädagogischer Tests" von Brickenkamp (Erstauflage: 1975; aktuelle Fassung: Brähler, Holling, Leutner & Petermann, 2002) konnte sich auch die Buchform als Quelle für Informationen über Tests etablieren. Mit der 1996 begonnen Reihe „Tests unter der Lupe" (Fay, 1996, 1999, 2000, 2003, 2005) war das Buchformat nicht länger auf Kurzeinträge beschränkt, sondern stand auch für ausführliche Rezensionen offen. Für einzelne Anwendungsgebiete konnten sich spezifische Überblickswerke etablieren, beispielsweise für den Bereich der Psychotherapie (Brähler, Schuhmacher & Strauß, 2002), für den Bereich der Eignungsdiagnostik (z. B. Dunckel, 1999; Kanning & Holling, 2002; Sarges & Wottawa, 2004) oder für die Klasse der motorischen Tests (Bös, 2001).

Andere Buchpublikationen widmen sich zwar nicht primär Testrezensionen, bieten aber gleichwohl umfassende testbezogene Informationen und Beurteilungen. So konzentrieren sich die Bände der Buchreihe „Kompendien psychologischer Diagnostik" (z. B. Holling, Preckel & Vock, 2004) beispielsweise auf ein Themengebiet wie Intelligenz und präsentieren hierzu sowohl Theorien und Forschungsbefunde als auch einen Überblick über ausgewählte Tests. Andere Autoren verbinden die Testrezensionen mit allgemeinen Ausführungen zur Testtheorie (z. B. Grubitzsch, 1999). Außerdem sind Testbesprechungen fester Bestandteil von Buchreihen (z. B. der Reihe „Tests und Trends"), und Diagnostiklehrbüchern (z. B. Amelang & Schmidt-Atzert, 2006; Kubinger, 2006).

Von hoher Bedeutung für die Auffindbarkeit der Informationen und Rezensionen sind die (zumeist online verfügbaren) Datenbanken, etwa das umfangreiche Informationsangebot der ZPID (http://www.zpid.de/index.php?wahl=products&uwahl=printed&uuwahl=testverzeichnis) oder der ZUMA (ZUMA-Informationssystem (ZIS) (siehe http://www.gesis.org/methodenberatung/zis/)) sowie die Testverzeichnisse einzelner, auf Tests spezialisierter Bibliotheken (z. B. http://www.sulb.uni-saarland.de/fachinfo/ssg/psychotest/system/ oder http://www.bis.uni-oldenburg.de/cgi-bin/tests.pl). Weiterhin zu nennen sind die Informationen der kommerziell orientierten Test-Zentrale (http://www.testzentrale.de) sowie einschlägiger Verlage (z. B. Harcourt (http://www.harcourt.de), Hogrefe (http://www.hogrefe.de) und Schuhfried (http://www.schuhfried.de)). Bei der ZPID (siehe oben) kann ein Verzeichnis von knapp 2000 Rezensionen zu fast 1000 Tests eingesehen werden (ftp://ftp.zpid.de/pub/tests/verz_teil5.pdf).

Dennoch kann das Angebot an Informationen über und Beurteilungen von Tests weder quantitativ noch qualitativ überzeugen. Die wahrgenommenen qualitativen Defizite werden im nächsten Abschnitt dargestellt. Dass die Quantität der Testrezensionen unzureichend ist, wurde von der Deutschen Gesellschaft für Psychologie bereits Ende der neunziger Jahre konstatiert. Infolge dessen wurde eine Initiative zur Anfertigung von Testrezensionen gestartet. Ziel war es, unter der Organisation des TK, zahlreiche Testrezensionen zu erstellen, die sich inhaltlich an der von Häcker, Leutner und Amelang (1998) übersetzten 1985er Fassung der APA Standards orientieren sollten (Amelang, 1999, S. 58). Diese Initiative kam aber nie richtig in Gang. Als Ursache für das Scheitern wird der große Aufwand sowie die Nachrangigkeit von Rezensionen in ihrem Stellenwert als wissenschaftliche Leistung vermutet.

Unbefriedigend ist zudem die Auswahl derjenigen Tests, die eine Rezension erfahren und somit die Repräs-

Tabelle 1. Anwendungshäufigkeiten (Selbstauskunft der Vertreiber) von typologisierenden Tests

	Analysen pro Jahr (2001) (deutschsprachig)	Gesamtzahl der bislang durchgeführten Analysen (Stand 2001)		Quelle
		deutschsprachig	international (inkl. deutschsprachig)	
Biostrukturanalyse®	25.000	600.000	700.000	1, S. 90
DISG-Persönlichkeits-Profil®	70.000	450.000	36 Mio.	1, S. 106
Herrmann-Dominanz-Instrument®	k.A.	über 100.000	ca. 1 Mio.	1, S. 149
INSIGHTS MDI®	k.A.	500.000	4 Mio.	1, S. 172
LIFO-Methode®	12.000	85.000	8 Mio.	1, S. 209
Team-Management-System®	3.500	30.000	700.000	1, S. 258

Anmerkung: Quelle: (1) Schimmel-Schloo, Seiwert und Wagner (2002).

sentativität der Testrezensionen. Das Rezensionswesen konzentriert sich vorwiegend auf Tests mit einem wissenschaftlich-universitären Konstruktionshintergrund. Infolgedessen bleiben Tests, die von rein kommerziell orientierten Organisationen entwickelt und/oder vertrieben werden in der Regel unberücksichtigt, auch wenn diese Tests in der Praxis häufig genutzt werden. Dies kann am Beispiel von typologisierenden Fragebogen veranschaulicht werden, die in der Eignungsdiagnostik zum Einsatz kommen. Tabelle 1 gibt einen Überblick über diese in der Wirtschaft häufig eingesetzten Tests, wobei die Zahlen auf Selbstauskünften der Verfahrensanbieter beruhen. Eine von Klimmer und Neef (2005) durchgeführte unabhängige Befragung der DAX und M-Dax Unternehmen kommt zu teilweise korrespondierenden Ergebnissen. 49 Prozent der teilnehmenden Unternehmen gaben an, in den letzten drei Jahren im Rahmen der Personalarbeit Persönlichkeitstypologien eingesetzt zu haben. Dabei fanden die Tests MBTI® und DISG® mit Abstand die häufigste Verwendung, gefolgt von dem Test zum Teamrollenmodell von Belbin sowie dem HDI®. Wer sich für neutrale Informationen und Beurteilungen zu diesen Tests interessiert, geht in der deutschsprachigen Testrezensionsliteratur weitgehend leer aus. Mit Ausnahme des MBTI®, finden sich in dem oben genannten Verzeichnis der ZPID mit annähernd 2000 Testrezensionen zu den genannten Tests keine Beiträge. So ist es nicht verwunderlich, dass es 73 Prozent der von Göhs und Dick (2001) befragten Unternehmen der deutschen Wirtschaft schwierig fanden, sich unter den Angeboten für einen passenden Test zu entscheiden und nur 28 Prozent der befragten Unternehmen im Rahmen der Personalauswahl Tests einsetzen. Informationen zu eignungsdiagnostischen Tests, die (teilweise) außerhalb des wissenschaftlich-universitären Kontextes entwickelt wurden, finden sich (abgesehen von dem Testhandbuch von Sarges & Wottawa, 2004) nur in verstreuten Buchpublikationen (z.B. Cisek, Schäkel & Scholz, 1989; Erpenbeck & Rosenstiel, 2003; Hossiep, Paschen & Mühlhaus, 2000; Schimmel-Schloo, Seiwert & Wagner, 2002; Wehner & Durchholz, 1980) oder als Einzelbeiträge in Zeitschriften, die keine wissenschaftlichen Fachzeitschriften sind (siehe z.B. Jäger, 2004). Tests, für die zahlreiche Testrezensionen vorliegen, wie z.B. das NEO-FFI, sind der Wirtschaft hingegen weitgehend unbekannt (Klimmer & Neef, 2005). Dies ist möglicherweise eine Folge des Umstands, dass diese Tests fast ausschließlich in wissenschaftsorientierten psychologischen Publikationsorganen besprochen werden, die von der Praxis kaum zur Kenntnis genommen werden. Überspitzt formuliert werden aktuell von wissenschaftlich arbeitenden Psychologen häufig Tests rezensiert, die in der Praxis vergleichsweise selten eingesetzt werden. Dies kann darin begründet sein, dass die Wissenschaftler den Bedarf und die Praxis der Anwender nicht kennen oder sich nicht daran orientieren möchten. Es kann aber auch daran liegen, dass die für eine Rezension notwendigen Informationen zu den in der Praxis häufig angewendeten Tests nicht zur Verfügung stehen (siehe weiter unten, S. 249). Insgesamt zeigt das eignungsdiagnostische Beispiel, dass es neben den benötigten und auch weiterhin willkommenen Initiativen einzelner Personen, Testrezensionen zu verfassen, zusätzlich einer übergeordnetem Steuerungsinstanz bedarf, die Testrezensionen in Auftrag geben kann und aktiv für deren zielgerichtete Verbreitung in Wissenschaft und Praxis sorgt (siehe S. 251). Auch in der aktuellen Debatte um die Beratung und Auswahl von Studienbewerbern, die u.a. testgestützt erfolgen soll (siehe z.B. das Diskussionsforum „Studierendenauswahl" im Heft 2/2005 der Psychologischen Rundschau), werden dringend Hilfestellungen zur Information über und Bewertung von Tests benötigt.

Das System der freien Testbeurteilungen: Vor- und Nachteile

Das deutschsprachige Testrezensionswesen kann als weitgehend frei bezeichnet werden. Der 1986 erschienene „Kriterienkatalog für die Beurteilung von psychologischen Tests" des von der Föderation der Deutschen Psychologenverbände (1986) berufenen TK, stellt zwar einen Orientierungsrahmen dar. Der Katalog beschränkt sich aber auf die bloße Nennung von Aspekten, die bei der Beurteilung eines Tests bedeutsam sind. Diese Aspekte (z.B. der Aspekt „Testdurchführung" mit den Unterpunkten Durchführungsobjektivität, Transparenz, Zumutbarkeit, Verfälschbarkeit und Störanfälligkeit) werden nur äußerst knapp erläutert, es gibt keine Hinweise oder Beispiele zur Bewertung. Ein vergleichbar freies System hat sich in der Schweiz mit den so genannten „Labels" der Diagnostikkommission des Schweizerischen Verbandes für Berufsberatung (SVB) etabliert (siehe www.testraum.ch).

Der Vorteil freier Systeme besteht in ihrer hohen Flexibilität, der Rezensent verfügt über eine große Gestaltungsfreiheit. Er kann entscheiden, welche Aspekte er ausführlich thematisiert und ob und wie er ein Gesamturteil bildet. Dies stellt jedoch zugleich den Nachteil des freien Systems dar. Da es keine verbindlichen Hinweise gibt, wie eine Testbeurteilung zu erstellen ist, fehlt jegliche Grundlage für ein intersubjektiv nachvollziehbares Vorgehen.

Die expliziten oder impliziten Urteile über Tests fallen in den deutschsprachigen Rezensionen häufig entweder unverbindlich oder aber extrem kritisch aus. Beides ist problematisch. Bei der Variante der unverbindlichen Rezension wird ein Test erläutert, seine Vor- und Nachteile werden dargestellt, aber ein verbindliches abschließendes Urteil zu Einzelaspekten oder zum Test insgesamt bleibt aus. Konkrete Anwendungsempfehlungen werden nicht ausgesprochen. Dies führt dazu, dass insbesondere Praktiker kaum Nutzen aus der Lektüre ziehen.

Häufig ist bei deutschsprachigen Testrezensionen aber auch ein überaus kritischer bis eindeutig negativer Ton anzutreffen. In dem oben bereits angesprochenen Rezensionsthemenheft der ZDDP aus dem Jahr 1997 wurden 25 Tests unter die Lupe genommen. In seinem Editorial nimmt Kubinger (1997, S. 3) das Fazit vorweg: Die meisten der 25 einschlägigen Tests seien nicht tauglich. Da die fundamental-kritischen Urteile der Rezensenten nicht regelgeleitet aus Einzelurteilen abgeleitet werden, bleibt es nach der Lektüre solcher Rezensionen häufig bei einem

allgemeinen Unbehagen gegenüber Tests, welches nicht konstruktiv in gezielte Verbesserungsmaßnahmen kanalisiert werden kann. Der kritische Tenor vieler Testrezensionen wirft entweder ein schlechtes Licht auf die Tests und Testautoren oder aber auf die Beurteiler. Da es keine Vorgaben hinsichtlich des Beurteilungsmaßstabs gibt, formuliert jeder Rezensent (leider nur implizit) seinen persönlichen Maßstab, der möglicherweise einseitig und zumindest diskussionsbedürftig ist. Da das System aber keine Transparenz der Maßstäbe vorsieht, findet eine solche Diskussion nicht statt. Die Wirkung solcher negativen Beurteilungen ist auch vor dem Hintergrund der Tatsache zu würdigen, dass überwiegend nur die Tests rezensiert werden, die einen wissenschaftlichen Konstruktionshintergrund aufweisen, während die Tests rein kommerzieller Anbieter (siehe Beispiele laut Tabelle 1) der Rezension entgehen. Daraus ergibt sich der falsche und fatale Eindruck, dass gerade Tests mit wissenschaftlichem Konstruktionshintergrund problematisch seien.

Freien Systemen ermangelt es schließlich an Regelungen dazu, wer überhaupt für die Erstellung von Testrezensionen qualifiziert ist. Da das Verfassen von Testrezensionen als wissenschaftlich nachrangige Leistung gilt, ist es eher schwierig, überhaupt Personal für diese Tätigkeit zu finden. Eine strenge Selektion nach Fachexpertise kann daher vielleicht nicht immer erfolgen und die händeringende Suche nach Rezensenten kann auch als Aufruf zum freien Dilettieren missverstanden werden.

Ausgewählte Systeme anderer Länder

Im Gegensatz zum deutschsprachigen Raum konnten sich in anderen Ländern standardisierte Systeme zur Information über und Beurteilung von Tests etablieren.

Als beispielhaft für die Quantität und Zentralität der Informationen über Tests kann das nordamerikanische BUROS System gelten (siehe z. B. Plake & Impara, 2001; http://www.unl.edu/buros). Eine vorbildliche Datenbank stellt der Educational Testing Service (ETS) zur Verfügung. Die URL http://www.ets.org/testcoll/index.html führt zu Beschreibungen von über 25.000 Tests und Forschungsinstrumenten.

Als Beispiel für ein standardisiertes Testbeurteilungssystem wird in dem vorliegenden Beitrag zunächst das niederländische COTAN System (Evers, 2001b) skizziert, welches vom Committee On Test Affairs Netherlands erarbeitet wurde. Bereits im Jahre 2000 umfasste die „Documentation of Test and Test Research" laut Evers (ebd.) 372 Tests, die einheitlich nach dem COTAN-System analysiert und hinsichtlich vorgegebener Kriterien von jeweils zwei Rezensenten beurteilt worden waren. Die sieben Kriterien, die zunächst separat beurteilt werden, sind in fünf Kategorien eingeteilt: (1) Testkonstruktion (Transparenz des Anwendungsbereiches, des theoretischen Hintergrunds und der Operationalisierung); (2) Qualität (2 a) des Testmaterials sowie (2b) der Verfahrenshinweise; (3) Normen; (4) Reliabilität und (5) Validität, hier (5a) Kriteriumsvalidität und (5 b) Konstruktvalidität. Um eine nachvollziehbare Beurteilung zu gewährleisten müssen alle Rezensenten vorgegebene Fragen beantworten. Zu dem Bereich (2 a) gibt es sechs, zum Bereich (2 b) sieben und zum Bereich (3) acht Fragen. Zu allen übrigen Bereichen liegen jeweils drei Fragen vor. Die Rezensenten haben die Aufgabe, die Fragen auf Grund der ihnen zur Verfügung stehenden Informationen zu beantworten und diese Antworten dann mit Hilfe einer vorgegebenen Beurteilungsskala zu bewerten. Wenn die Informationen, die zur Beantwortung der Fragen notwendig sind, nicht zur Verfügung stehen, führt dieses Informationsdefizit genauso zu einem negativen Urteil wie vorhandene negative Informationen. Die den Beurteilungen pro Frage zu Grunde gelegten Kriterien sind ausführlich operationalisiert (Evers, 2001 a). Das System sieht Beurteilungsregeln vor, die für verschiedene Anwendungsbereiche unterschiedlich streng ausfallen. So muss ein Test, der für wichtige Entscheidungen (z. B. Personalauswahl) über Individuen vorgesehen ist, z. B. eine Reliabilität von mindestens $r = .90$ aufweisen, um die Bewertung „gut" erzielen zu können (siehe Tabelle 2). Auch die Kombination der Beurteilungen der Einzelfragen zu einer Gesamtbeurteilung pro Bereich ist geregelt. Auf diese Art und Weise wird jeder Test von zwei unabhängigen Rezensenten eindeutig als „gut", „ausreichend" oder „unzureichend" beurteilt. Bei Abweichungen in der Beurteilung wird eine Konsensbildung durch Diskussion zwischen den Rezensenten und in Ausnahmefällen durch die Berufung eines dritten Rezensenten erzielt.

Tabelle 2. COTAN System (Evers, 2001a): Beispiel für die Richtlinien zur Bewertung der Reliabilität sowie des Umfangs der Normstichproben

	Reliablität[1]			Umfang Normstichproben		
	Niveau 1	Niveau 2	Niveau 3	Niveau 1	Niveau 2	Niveau 3
unzureichend	<.80	<.70	<.60	<300	<200	<100
ausreichend	.80–.90	.70–.80	.60–.70	300–400	200–300	100–200
gut	>.90	>.80	>.70	>400	> 300	>200

Anmerkungen: Niveaus: (1) Tests für wichtige Entscheidungen auf der individuellen Ebene (z. B. Personalauswahlentscheidung), (2) Tests für weniger bedeutsame Entscheidungen auf der individuellen Ebene (z. B. Fortschrittskontrolle), (3) Tests für Untersuchungen auf Gruppenniveau.
[1] Für Paralleltest-Reliabilität, interne Konsistenz, Test-Retest-Reliabilität und Interrater-Reliabilität.

Mittlerweile hat die European Federation of Psychologists Associations (EFPA) eine Initiative zur Entwicklung eines „common set of European criteria for test reviews" gestartet (Bartram, 2001, S. 180). Dieser Ansatz basiert einerseits auf dem britischen System (http://www. psychtesting.org.uk/viewer.asp?ID=212&sectionid=10&subsection=1) weist andererseits aber auch deutliche Ähnlichkeiten mit dem niederländischen COTAN System auf. Der aktuelle Stand des „EFPA Review Model for the Description and Evaluation of Psychological Tests" ist im Internet verfügbar (http://www.efpa.be).

Vergleichbar zum COTAN System soll auch hier jeder Test von zwei unabhängigen Reviewern begutachtet werden. Das System sieht ebenfalls Beurteilungsrichtlinien und die Anwendung einer Beurteilungsskala vor. Außerdem kann die Bewertung ausbleiben, wenn das Bewertungsmerkmal nicht sinnvoll auf den Test angewendet werden kann oder keine (ausreichenden) Informationen zu dem Merkmal vorliegen. Das EFPA System orientiert seine Beurteilungsrichtlinien ebenfalls an der numerischen Ausprägung von Gütekriterien und legt je nach Anwendungszweck (Entscheidungen über Individuen [z. B. Auswahlentscheidungen] oder Gruppen) unterschiedlich strenge Maßstäbe an. Beispielsweise werden die Kennwerte für die interne Konsistenz auf der vierstufigen Skala entsprechend ihrer Höhe als „inadequate" oder „excellent" klassifiziert. Konkrete Orientierungshilfen finden sich auch hinsichtlich der Frage des Umfangs der Normstichproben (weniger als 150 Personen gelten als „inadequate", mehr als 1000 als „excellent") und dem Umfang der Studien zur Reliabilität und Validität (eine Studie mit weniger als 100 Personen gilt „inadequate", mehrere Studien mit jeweils mehr als 100 Personen gelten als „excellent").

Im Vergleich zum COTAN System legt das EFPA System mehr Wert auf die zu Beginn der Rezension abverlangte, möglichst wertfreie Beschreibung des zu beurteilenden Tests. Es existiert eine umfangreiche Liste der dabei zu berücksichtigenden Merkmale. Fokussiert werden außerdem die möglicherweise angebotenen computergenerierten Berichte. Der EFPA Rezensionsprozess endet mit einer eindeutigen Anwendungsempfehlung, die sechs Empfehlungskategorien unterscheidet. Dieses Gesamturteil lässt sich allerdings nicht arithmetisch aus den Einzelurteilen herleiten. Eine Empfehlung kann z. B. lauten, dass Instrument nur in der Forschung, nicht aber in der Praxis einzusetzen. Andere Empfehlungen schränken die Anwendung der Tests ein, indem lediglich eine Nutzung durch Experten oder (andere Empfehlungskategorie) unter Supervision empfohlen wird. Bei den Tests, die als „geeignet" empfohlen werden, wird zwischen Tests, die durch besonders qualifizierte Testnutzer angewendet werden sollen und Tests, die sich auch für nicht beaufsichtigte Selbsttestungen eignen, unterschieden.

Vor- und Nachteile standardisierter Beurteilungssysteme

Der zentrale Vorteil standardisierter Beurteilungssysteme liegt in der erhöhten Transparenz und Objektivität der Beurteilung. Die Vorgabe von Besprechungsaspekten erhöht die Vergleichbarkeit und entlastet die Rezensenten. Die Wahrscheinlichkeit, dass relevante Aspekte übersehen werden, sinkt. Die Vorgabe einer bestimmten Bewertungsskala zwingt die Rezensenten zu einem eindeutigen Urteil. Da über verschiedene Tests hinweg die gleichen Beurteilungskriterien und Beurteilungsskalen genutzt werden, ist es möglich, eine testübergreifende Betrachtung der Qualität von Tests vorzunehmen, grundsätzlich kritische Bereiche zu identifizieren und somit für Wissensakkumulation zu sorgen. Evers (2001a) konstatiert beispielsweise, dass sich die Qualität der nach dem COTAN System rezensierten Tests über die Jahre hinweg (möglicherweise auf Grund der standardisierten Rezensionen) verbessert hat, dass aber die Bereiche Normen und Kriteriumsvalidität die neuralgischen Punkte der meisten Tests darstellen. Sowohl das COTAN als auch das EFPA System stellen den Rezensenten einen exzellenten und mit hoher Fachexpertise verfassten Erläuterungstext zur Verfügung, der als Beurteilungsrichtlinie fungiert. Positiv hervorzuheben sind schließlich die formellen Regelungen, die einen doppelten Review-Prozess vorsehen und Fragen der Vertraulichkeit von Informationen regeln.

Als Nachteil standardisierter Systeme erweist sich deren Starrheit. Die Beurteilung ist statisch und somit potenziell – sofern das System nicht permanent überarbeitet wird – innovationshemmend. Während sich die Vorstellungen über Qualitätsmerkmale und ihre Bestimmung ändern, schreibt ein standardisiertes System einen einmal erreichten Status fest. Das geschlossene System wird außerdem der Multifunktionalität von Tests nicht gerecht. Die bei COTAN und EFPA realisierte Idee, für unterschiedliche Anwendungsbereiche unterschiedlich strenge Regeln zu formulieren, zeugt von Problembewusstsein, ist aber keine tragfähige Lösung. Die Frage, ob Einzelbeurteilungen zu unterschiedlichen Gütekriterien sinnvoll zu einem Gesamturteil verrechnet werden können, bedarf der grundlegenden Diskussion und des empirischen Belegs. Hier sind zunächst die Widersprüche und Konflikte zu bedenken, die zwischen verschiedenen Gütekriterien (z. B. Reliabilität-Validitäts-Dilemma oder das Dilemma, dass eine Testverlängerung in der Regel die Reliabilität verbessert, die Testökonomie und Akzeptanz aber senkt usw.) sowie zwischen verschiedenen Interessensgruppen herrschen können. Hinsichtlich der Bildung eines Gesamtwerts stellt sich die Frage, ob die einzelnen Aspekte kompensatorisch oder konjunktiv zu einem Gesamturteil verbunden werden sollten und welche inhaltliche Bedeutung ein wie auch immer gebildeter Gesamtwert hat.

Das Hauptproblem des COTAN und EFPA Ansatzes besteht in der direkten Bewertung der numerischen Ausprägung von Kennwerten nach vorgegeben Regeln. So formuliert das COTAN System beispielsweise die Regel, dass eine Reliabilität kleiner als $r = .80$ für Tests (wenn sie für wichtige Entscheidungen auf individueller Ebene eingesetzt werden) als „unzureichend" zu bewerten sei usw. (siehe Tabelle 2). Die Autoren der Systeme sind sich der damit verbundenen Probleme bewusst. Sowohl in den Erläuterungen zum COTAN als auch in den Erläuterungen

zum EFPA System finden sich explizite Distanzierungen von den selbst vorgeschlagenen Mindestgrenzen. Es wird eingeräumt, dass es für die genannten Einteilungen keine schlüssigen wissenschaftlichen Begründungen gäbe und dass es unmöglich sei, klare Kriterien für die Bewertung der technischen Qualitäten eines Instruments festzulegen. Es ist allerdings zu befürchten, dass dieser intelligente Subtext in der Praxis verloren geht und die veröffentlichten Einteilungen der numerischen Ausprägung der Koeffizienten in „gut" und „böse" ohne Verstand, aber mit dem Taschenrechner gehorsam abgehakt, befolgt und verfolgt werden. Schon jetzt existieren von einigen Anbietern Beurteilungssysteme für Tests, die nach diesem nicht begründbaren, aber einfachen Bewertungsschema arbeiten. Die Testbeurteilung nach Zahlen kommt dem Ruf der Praxis nach einfachen Beurteilungssystemen entgegen. Eine solche Beurteilung ist objektiv, aber scheingenau. Es handelt sich um überkohärente und pseudorationale Beurteilungssysteme.

Eine Bewertung der Güte eines Tests auf Grund der numerischen Ausprägung von Kennwerten wäre nur dann möglich, wenn die Kennwerte vollständig vom Messinstrument selbst dominiert würden. De facto charakterisieren Testgütekennwerte nach der klassischen Testtheorie, wie Fischer bereits (1968, S. 133) ausführte, jedoch die jeweils realisierte Kombination aus einem Test einerseits und einer Untersuchungsgruppe andererseits. Die Ausprägung der Gütekriterien hängt wesentlich von der Verteilung der Parameter in der Referenzpopulation ab. Die Reliabilität eines Tests variiert also beispielsweise in Abhängigkeit davon, welche Referenzpopulation herangezogen wird. Die Reliabilität charakterisiert somit nicht die Genauigkeit eines Messinstruments, sondern seine Genauigkeit in Bezug auf eine bestimmte Population. Entscheidend für die aus der Kombination zwischen dem Test und der Untersuchungsgruppe resultierenden Kennwerte sind z.B. die Streuungen der Werte in der jeweiligen Gruppe. Gerade in der angewandten psychologischen Diagnostik hat man es z.B. häufig mit Streuungseinschränkungen zu tun, wenn etwa im Kontext der Studierendenauswahl nur Personen mit Abitur zu einem Auswahltest zugelassen werden. Würde man in einer solchen Situation die Kennwerte eines Tests bestimmen, so lassen sich diese nicht ohne weiteres mit den Kennwerten vergleichen, die an einer bildungsrepräsentativen und somit heterogenen Gesamtgruppe erhoben wurden. Für eine Testrezension sind diese Umstände der Kennwertberechnung zu berücksichtigen.

Darüber hinaus handelt es sich bei Kennwerten lediglich um Einzelschätzungen. Schmidt (1992) verdeutlicht die Relativität solcher Einzelschätzungen anhand einer Studie zur Kriteriumsvalidität eines eignungsdiagnostischen Tests. Die Studie wurde mit 1428 Personen durchgeführt und ergab für den Test eine Kriteriumsvalidität von $r = .22$. Schmidt (ebd.) zerlegte die Gesamtgruppe per Zufall in 21 Teilgruppen mit je 68 Personen (dem Medianwert der Stichprobengröße von Studien zur Kriteriumsvalidität eignungsdiagnostischer Tests) und bestimmte für jede der Teilgruppen die Kriteriumsvalidität. Die resultierenden Kriteriumsvaliditäten lagen im Bereich von $r = .02$ bis $r = .39$. Nach dem EFPA System würde die Kriteriumsvalidität ein und desselben Tests also mal als „nicht ausreichend", mal aber als „gut" gewertet. Diese Art des statistischen Schließens gilt seit der Etablierung der Meta-Analyse als überwunden. Mit fixen Beurteilungskategorien nach Art des COTAN und EFPA Systems würde man zu dem Primat der Situationsspezifität regredieren. Erst die Abkehr von der Überbewertung der Einzelstudien und die Einführung von Meta-Analysen (z.B. Hunter & Hunter, 1984) zeigte, dass Validitätskoeffizienten in hohem Maße generalisierbar sind und die echte, nicht zu Lasten statistischer Artefakte gehende Varianz kaum Raum für Moderatoreffekte lässt. Die Heterogenität der Befunde einschlägiger Studien ließ sich überwiegend auf drei Artefakte zurückführen: (a) unterschiedlich reliable Prädiktoren und Kriterien, (b) Streuungsdifferenzen zwischen den untersuchten Stichproben und (c) zufallsbedingte Variation der Korrelations- bzw. Regressionskoeffizienten. Testbeurteilungen, die sich allein an der numerischen Ausprägung von Koeffizienten orientieren, sind dem Einfluss dieser hier genannten Artefakte weitgehend hilflos ausgeliefert.

Fataler noch ist die Tatsache, dass eine Fixierung auf die numerische Ausprägung von Kennwerten einseitig gestaltete Untersuchungspläne provozieren würde. Falls hohe Koeffizienten wider besseren Wissens als Qualitätsmerkmal eines Tests geadelt würden, würde man in Zukunft die Untersuchungen so gestalten, dass nicht die Wahrscheinlichkeit für Erkenntnisse, sondern die Wahrscheinlichkeit für hohe Koeffizienten steigt. Da die Kennwerte eine Kombination aus Test und Untersuchungsgruppe sind, kann man die erforderlichen Qualitätswerte für schlechte Tests mit „besonderen" Untersuchungsgruppen erzielen. Hierzu ist es z.B. förderlich, heterogene Stichproben zu untersuchen, die breite Streuungen aufweisen. Für die Gültigkeitsprüfung kann man Kriterien heranziehen, die vor allem reliabel sind und bei denen z.B. eine Prädiktor-Kriteriums Kontamination vorliegt usw. Bei der Bestimmung der Retest-Reliabilität wirkt sich in der Regel die Wahl eines sehr kurzen Zeitintervalls positiv auf den Kennwert (aber nicht auf die Erkenntnis) aus. Ein weiteres und letztes Beispiel kann anhand von Kriteriumsvaliditäten aufgezeigt werden. Häufig werden über ein Dutzend Skalen eines Tests mit zahlreichen Kriterien korreliert. Aus den möglicherweise über 100 Korrelationen suchen sich die Testautoren dann einige wenige „hohe" Korrelationen als Nachweis der Kriteriumsvalidität heraus. Hier kann es in einer Testrezension nicht darum gehen, die Höhe dieser Korrelationen zu klassifizieren. Notwendig ist vielmehr, dass die Testrezensenten über ein vertieftes Verständnis von Statistik und methodisch angemessenen Vorgehensweisen verfügen (siehe Hager, 2005). Bei der simultanen Bestimmung zahlreicher Korrelationskoeffizienten kumulieren die statistischen Fehlerwahrscheinlichkeiten, so dass es in diesem Fall Aufgabe der Rezensenten ist, zu prüfen, ob eine entsprechende Kontrolle durchgeführt wurde. Vorab ist zu bewerten, ob die gewählten Kriterien theoretisch angemessen sind.

Da im nächsten Abschnitt auf die DIN 33430 (DIN, 2002) eingegangen wird, sei an dieser Stelle erwähnt, dass

die DIN Kommission sich aus den hier ausgeführten Gründen gegen eine Interpretation der numerischen Ausprägungen von Kennwerten als Qualitätsmerkmal entschieden hat. Dies wird oft verkannt, da in dem so genannten „Anhang B" der DIN 33430 zur Reliabilität ausgeführt wird: „Erfahrungsgemäß ergeben sich bei Zuverlässigkeitsuntersuchungen je nach gewählter Verfahrensklasse und Art der Zuverlässigkeit Werte zwischen $r = 0.70 - 0.85$" (DIN, 2002, S. 24). Abgesehen davon, dass es sich hierbei um eine beschreibende und nicht um eine wertende Aussage handelt, gilt es zu beachten, dass der „Anhang B" nicht normativ (sondern informativ) ist. Er hat (ebenso wie alle „Anmerkungen" der DIN 33430 und alle „Leitsätze") keinerlei normative Kraft. Im normativen Teil der DIN wird zur Reliabilität (DIN, 2002, S. 6) und Validität (ebd. S. 5) explizit ausgeführt, dass neben der numerischen Höhe der Koeffizienten auch die Qualität der Untersuchungen zu bewerten ist, mit denen die Koeffizienten bestimmt wurden. Als Qualitätsmerkmale werden die Angemessenheit des Untersuchungsansatzes für das zu messende Merkmal sowie die Größe, Repräsentativität (für die Zielgruppe) und Aktualität der Untersuchungsgruppe benannt. Als Qualitätsmerkmal gilt darüber hinaus vor allem das Vorliegen unabhängiger Vergleichs- und Wiederholungsuntersuchungen (DIN, 2002, S. 6).

Die in COTAN und EFPA genannten Empfehlungen zu den Stichprobengrößen sind ebenfalls auf Grund der Unterschiedlichkeit der Anwendungszwecke inhaltlich/methodisch nicht nachvollziehbar. Angemessen ist vielmehr eine Betrachtung der jeweils notwendigen Teststärke, die in Abhängigkeit von den zu erwartenden Effekten festgesetzt wird.

Testbeurteilungen und die DIN 33430

Im Juni 2002 wurde die DIN 33430 verabschiedet (DIN, 2002; Heyse & Kersting, 2004; Kersting & Heyse, 2004), die „Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen" formuliert. Es besteht ein enger inhaltlicher Zusammenhang zwischen den Themen „DIN 33430" und „Testbeurteilungen", dieser Zusammenhang erschließt sich aber auf Grund der nachfolgend genannten drei Gründe nicht auf den ersten Blick: (1) Die DIN 33430 bezieht sich allein auf berufsbezogene Eignungsbeurteilungen; (2) sie ist eine Prozess- und keine Produktnorm; (3) sie bezieht sich nicht explizit auf Tests, sondern auf alle Verfahren der Eignungsbeurteilung (also z. B. auch auf Interviews und Assessment Center). Welche Bedeutung kann der DIN 33430 unter diesen Umständen für allgemeine Testrezensionen zukommen? Kersting (2006) hat eine Checkliste zur DIN 33430 erstellt und dafür den Text der DIN 33430 in 318 Einzelaussagen zerlegt sowie thematisch geordnet. Dabei stellte sich heraus: 140 Aussagen (44% aller Aussagen) der DIN 33430 formulieren Anforderungen an Verfahrenshinweise (auch „Handanweisungen" oder in Bezug auf Tests „Testmanuale" genannt). Diese Teilmenge der DIN 33430 wurde von Kersting (ebd.) zu einer eigenständigen Checkliste „Anforderungen an Verfahrenshinweise" zusammengefasst, die auch online verfügbar ist (http://www.kersting-internet.de/DIN-Screen.html). Die DIN 33430 fordert von den Testautoren und/oder -vertreibern umfassende Informationen zur Konstruktion und empirischen Überprüfung sowie zur Anwendung, Auswertung und Interpretation der Verfahren ein. Im Detail geht es von der Forderung nach einer transparenten Informationspolitik über die Forderung nach umfassenden Informationen zu den empirischen Untersuchungen, aus denen die Verfahrenskennwerte abgeleitet wurden, bis hin zu der Pflicht, über den empfohlenen Umgang mit nicht bearbeiteten Testaufgaben oder Items zu informieren. Sofern Merkmale erfasst werden sollen, für die eine Zeit- und Situationsstabilität angenommen wird, müssen Angaben zur Retest-Reliabilität berichtet werden usw.

Verfahrenshinweise, die diesen umfangreichen Anforderungen nicht gerecht werden, entsprechen nicht der DIN 33430. Damit kann ein Prozess, bei dem ein Test mit derartig unzureichenden Verfahrenshinweisen eingesetzt wird, insgesamt auch nicht der DIN 33430 entsprechen. Man kann somit zwar niemals formulieren, dass ein Test der DIN 33430 entspricht, gleichwohl kann man aber konstatieren, dass ein Prozess der Eignungsbeurteilung *nicht* der DIN 33430 entspricht, weil bereits ein Prozesselement nicht den DIN 33430 Anforderungen genügt.

Während sich zahlreiche Aussagen der DIN 33430 explizit nur auf die Eignungsbeurteilung beziehen, kann die hier in Frage stehende Teilmenge der DIN 33430 im Umfang von 140 ausgewählten, verfahrensbezogenen Aussagen auf Tests aus allen Anwendungsbereichen angewendet werden. Mit der Nutzung einer Teilmenge der DIN 33430 im Kontext von Testbeurteilungen für alle Testanwendungsbereiche sind Vorzüge verbunden, die sich aus dem Charakter der „Norm" ergeben. Zwar sind DIN Normen nicht rechtsverbindlich, gleichwohl können sie, anders als berufsständische Regeln, in Rechtsstreitigkeiten als eine Art Beweismittel Bedeutung erlangen (siehe Kersting & Püttner, 2006). Unbestimmte Begriffe wie „allgemein anerkannte Regeln der Technik" können mit den Inhalten der DIN ausgefüllt werden. Unter anderem aus diesem Grunde ist es mit der DIN 33430 erstmals gelungen, allen im Bereich des psychologischen Testens Tätigen (und das sind zu fast 90% Nicht-Psychologen; siehe Bartram, 2001) – ungeachtet ihrer Berufsgruppenzugehörigkeit – Qualitätsstandards zu setzen. Gerade in dieser Hinsicht unterscheidet sich der DIN Ansatz von den Qualitätsinitiativen der Vergangenheit. Diese wurden innerhalb psychologischer Berufsverbände formuliert und entfalteten bestenfalls bei den Mitgliedern dieser Gruppe ihre Wirkung (Kersting & Hornke, 2003). Ein über die Berufsgruppe der Psychologen hinausgehender Effekt ist insbesondere in Bezug auf Tests notwendig, da eben Tests häufig von Nicht-Psychologen konstruiert und/oder vertrieben und/oder angewendet werden (siehe Tabelle 1). Ein wesentliches Problem der bisherigen Beurteilung von Tests besteht darin, dass sich etliche Anbieter einer Bewertung ihrer Tests entziehen, indem wesentliche Infor-

mationen zum Test vorenthalten werden. Auch Experten können angesichts eines Informationsmangels nicht sagen, ob dieser und jener Test eine mangelhafte Qualität aufweist. Mit diesem Informationsdefizit räumt die DIN 33430 auf, indem sie umfangreiche Informationen zum Verfahren einfordert. Der DIN 33430 liegt somit ein Informationsansatz zu Grunde. Auf der Basis der geforderten Informationen kann ein sachverständiger Rezensent auf der nächsten Stufe eines Bewertungsverfahrens (siehe unten) eine Testbeurteilung vornehmen. Explizit formuliert: Die Testbeurteilung erfolgt nicht nach der DIN 33430, sondern die Testbeurteilung erfolgt auf Grund der Informationen, die nach der DIN 33430 zu einem Test vorliegen müssen. Sind zentrale Informationen aber erst gar nicht vorhanden (was häufig der Fall ist), führt dies automatisch zu einer negativen Beurteilung. Somit eignet sich die DIN 33430 für ein rasches Negativ-Screening von Tests und als Vorbereitungsstufe einer Testbeurteilung, ohne selbst ein Testbeurteilungssystem zu sein. Die Grundidee, bewährte Standards für die Beurteilung von Tests zu nutzen, wurde auch von Braden und Niebling (2005) verfolgt. Diese Autoren nutzen die APA Standards zur Prüfung von Tests.

Die Bewertung eines Tests, unabhängig von dessen Einsatz, z. B. nach Art eines Testgütesiegels, ist nach DIN 33430 nicht möglich, da der jeweilige diagnostische Auftrag, die Rahmenbedingungen, das Verfahren und die mitwirkenden Personen simultan zu betrachten sind. Die DIN 33430 zielt auf eine Prozesslenkung, nicht auf eine Prüfung am Ende der Produktionskette. Nach der DIN 33430 ist nicht ein Test an sich problematisch, sondern der Gebrauch, der von einem Test gemacht wird, kann problematisch sein. In diesem Sinne sind die Verfahrenshinweise auch eine Gebrauchsanweisung, die über das Produkt und seinen Gebrauchswert informiert. Für die Verfahrenshinweise sollen im Sinne eines „Reporting" diejenigen Informationen erarbeitet und systematisch sowie empfängerorientiert aufbereitet werden, die für eine Entscheidung über den Testeinsatz sowie für den Testeinsatz selbst notwendig sind.

Das neue TK-System zur Information über und Beurteilung von Tests

Es liegt nahe, aus den Vorzügen der vorhandenen Systeme (einschließlich des DIN 33430 Ansatzes) ein neues System zu formieren, welches das Beste aus verschiedenen Welten vereint und die systemspezifischen Nachteile minimiert. Das Testkuratorium (TK) (in Vorb.) hat ein entsprechendes System zur Information über und Beurteilung von Tests entwickelt und verabschiedet, das an dieser Stelle skizziert und kommentiert werden soll. Das System umfasst die folgenden drei Stufen:

- Stufe 1: Prüfung der Informationsgrundlage nach DIN 33430
- Stufe 2: Testbeschreibung nach einem vorgegebenen Raster und unter Berücksichtigung der Datenbankkompatibilität
- Stufe 3: Testbeurteilung und -rezension durch zwei unabhängige Rezensenten auf der Grundlage einer Beurteilungsrichtlinie

Für die erste Stufe sind nur diejenigen Aussagen der DIN 33430 relevant, die sich auf die Verfahrenshinweise beziehen und die über den Anwendungsbereich Eignungsbeurteilung hinweg Geltung beanspruchen können. Grundlage der Prüfung auf Stufe 1 ist die „Checkliste 1" der Publikation „DIN SCREEN" (Kersting, 2006), die offiziell als „Standard zur Information und Dokumentation von Instrumenten zur Erfassung menschlichen Erlebens und Verhaltens des Testkuratoriums der Föderation Deutscher Psychologenvereinigungen" gilt. Diese Checkliste sollte bereits von den Testanbietern ausgefüllt sein, die Angaben werden von den Rezensenten geprüft. Der Bericht über die Prüfung nach Stufe 1 kann sich auf die Aspekte beschränken, zu denen die nach DIN 33430 geforderten Informationen in den Verfahrenshinweisen *fehlen*. Derartige Beanstandungen sollten mit zunehmender Etablierung des neuen Systems seltener werden, da im Sinne eines pro-aktiven Vorgehens auch den Testautoren und Vertriebsorganisationen die Checkliste als Qualitätsleitfaden zur Verfügung steht. Die erste Prüfstufe endet mit der Feststellung, ob der Test nach Ansicht der Rezensenten „prüffähig" ist. Ein Test, der in diesem Sinne nicht prüffähig ist, erhält ohne weitere Begutachtungen eine negative Gesamtbeurteilung. Das Urteil zur „Prüffähigkeit" ist ein qualitatives Werturteil der Rezensenten und wird nicht allein auf Grund von Auszählungen der mit „trifft zu" und „trifft nicht zu" beantworteten Aussagen in der Checkliste bestimmt. Dies liegt vor allem darin begründet, dass sich die Informationsanforderungen der DIN 33430 in ihrer Bedeutsamkeit deutlich unterscheiden. Welche Anforderungen als besonders bedeutsam oder weniger bedeutsam gelten, variiert zusätzlich in Abhängigkeit von der (vom Testautor formulierten) diagnostischen Zielstellung des Tests.

Auf der Stufe 2 erfolgt eine Testkategorisierung und es werden formale Datenbankangaben getroffen. Hierzu wird das Kategorisierungssystem der ZPID (http://www.zpid.de/index.php?wahl=products& uwahl=fee&uuwahl=ptfeld1) genutzt. Zur Wahrung internationaler Kompatibilität werden zusätzlich noch datenbankfähige Informationen im Sinne des EFPA Systems verlangt.

Die eigentliche Testbeurteilung stellt die dritte Stufe des Systems dar. Hierzu hat das TK (in Vorb.) eine Beurteilungsrichtlinie verfasst. Durch die Richtlinie soll die Objektivität und Vergleichbarkeit der Rezensionen erhöht werden, obwohl die Beurteilung ein subjektives Werturteil der Rezensenten bleibt. Eine Bewertung allein auf Grund der nominellen Ausprägung der Testgütekriterien, die weiter oben mit Bezug auf das COTAN und EFPA System kritisiert wurde, ist im TK System nicht vorgesehen. In der Richtlinie werden sieben Beurteilungskategorien vorgegeben (siehe Tabelle 3). In vier Fällen (z. B. für „Reliabilität" und „Validität") müssen die Rezensenten zusätzlich zum Freitext ihre Beurteilung auf einer Skala ausdrücken. Diese Beurteilungsskala sieht vier Abstufungen vor: „Der

Tabelle 3. Besprechungs- und Beurteilungskategorien des neuen TK Systems (TK, in Vorb.)

	Bewertung (*)	max. Zeichenzahl (inkl. Leerzeichen) für die freie Bewertung
1. Allgemeine Informationen über den Test, Beschreibung des Tests und seiner diagnostischen Zielsetzung	frei und formalisiert	1000
2. Theoretische Grundlagen als Ausgangspunkt der Testkonstruktion	frei	1000
3. Objektivität	frei und formalisiert	1000
4. Normierung (Eichung)	frei	1000
5. Zuverlässigkeit (Reliabilität, Messgenauigkeit)	frei und formalisiert	1000
6. Gültigkeit (Validität)	frei und formalisiert, auch unter Berücksichtigung der Fairness (soweit in Anspruch genommen)	1000
7. Weitere Gütekriterien (Störanfälligkeit, Unverfälschbarkeit und Skalierung)	frei	1000
8 Abschlussbewertung/Empfehlung	frei	2000
		Gesamt: max. 9.000

Anmerkungen: (*) zur formalisierten Bewertung ist eine vierstufige Skala vorgesehen: „Der Test erfüllt die Anforderungen (...) (a) >voll<, (b) >weitgehend<, (c) >teilweise< und (d) >nicht<.

Test erfüllt die Anforderungen (...) (a) >voll<, (b) >weitgehend<, (c) >teilweise< und (d) >nicht<. Der Umfang des Rezensionstextes pro Kategorie ist ebenfalls geregelt (siehe Tabelle 3).

Alle nach dem neuen System erarbeiteten testbezogenen Informationen und Beurteilungen werden in „Report Psychologie" veröffentlicht. (Der Nachdruck in anderen Zeitschriften ist möglich und erwünscht.) Sie sollten zusätzlich online zugänglich sein. Ideal wäre eine unbeschränkte und freie online Zugänglichkeit als open-access Verfahren (siehe Bierhoff, Funke, Reips & Weichselgartner, 2005).

Institutionalisierung

Verantwortlich für die Gestaltung und Umsetzung des neuen Systems ist das TK. Eine zentrale Organisation ist notwendig, um die Aktivitäten zu überblicken und zu koordinieren. Weitere Funktionen liegen darin, dass diese Organisation „nach außen" hin als Ansprechpartner sichtbar und erreichbar ist sowie in allen Fragen rund um das Thema „Test" Fachkompetenz demonstriert.

Der organisatorische Ablauf für eine konkrete Testrezension nach dem TK System sieht wie folgt aus. Anstelle von Gelegenheitstestrezensionen steht zu Beginn eine aktive Auswahl der zu rezensierenden Tests durch das TK. Ergänzend ist es möglich, dass die Fachgemeinschaft einzelne Tests zur Besprechung vorschlägt. Für jeden zu beurteilenden Test beauftragt das TK gegen Honorar zwei Reviewer. Das TK bürgt für die einschlägige Expertise und Unabhängigkeit sowie Unvoreingenommenheit der Reviewer. Die beiden Reviewer arbeiten zunächst in Unkenntnis und unabhängig voneinander ihre Beurteilungen für die vorgegebenen Beurteilungsaspekte aus, fixieren diese Urteile auf der vorgegebenen Beurteilungsskala und übermitteln sie an das TK. Erst wenn die Urteile beider Rezensenten vorliegen, werden die beiden Reviewer aufgefordert, im zweiten Schritt gemeinsam zu arbeiten und Konsensurteile zu finden sowie eine endgültige Rezension zu verfassen. Sollten trotz eines Einigungsversuches abweichende Beurteilungen fortbestehen, werden zu den strittigen Passagen beide Beurteilungen veröffentlicht. Die von beiden Reviewern gemeinsam erstellte Rezension wird vom TK in anonymisierter Form an die Testautoren geschickt, um ihnen Gelegenheit einzuräumen, innerhalb einer gesetzten Frist Stellung zu beziehen. Im Falle einer solchen Stellungnahme entscheidet das TK, ob es die beiden Rezensenten bittet, auf Grund der Stellungnahme eine Modifikation der Testrezension vorzunehmen. Jeder Rezensent entscheidet für sich, ob er bei der Publikation der Rezension namentlich genannt werden oder ob er anonym

bleiben will. Die Rezensenten werden auf Wunsch am Ende eines Jahres in „Report Psychologie" als Mitglieder des „TK-Expertenpools" genannt. Im Falle von „confidential tests" sichert das TK den Testanbietern die Vertraulichkeit bestimmter Informationen zu. Diese Zusicherung ist notwendig, weil sich einige Testanbieter aktuell einer Rezension ihrer Tests mit der Begründung entziehen, sie könnten ihr Kapital (die Testentwicklung) und/oder vertrauliche Kundenbeziehungen (etwa in Form von Angaben zu den Normgruppen usw.) nicht bekannt geben.

Das TK ist für die Analyse und Publikation der testübergreifenden Erkenntnisse (vgl. Evers, 2001 a) sowie für die kontinuierliche Evaluation und Modifikation des Systems verantwortlich. Außerdem obliegt ihm die Dokumentation und gezielte Publikation der Testrezensionen. Neben der fortbestehenden Nutzung von psychologieinternen Medien, ist für eine aktive Ansprache fachfremder Medien zu sorgen. Auch die Etablierung eines Internetportals ist anzustreben.

Dem mit dem neuen System verbundenen hohen Aufwand sind potenzielle Einnahmen gegenüberzustellen: Möglich sind direkte Einnahmen durch Testanbieter, die bereits in der Vergangenheit eine Sockelfinanzierung für Testrezensionen zur Verfügung gestellt haben (Amelang, 1999). Darüber hinaus könnte, unter Berücksichtigung des Rechtsstatus der DGPs, der Ausbau des TK zu einer Organisation erwogen werden, die ihre Kosten durch die neutrale Beratung in allen Fragen rund um das Thema „Testen" sowie durch Evaluationsstudien und/oder Zertifizierungen deckt, sofern keine der Dienstleistungen die Unabhängigkeit und Unparteilichkeit gefährdet. Denkbar ist auch, dass diese Organisation Datenbanken zu Tests aufbaut (z. B. für Normierungen oder Re-Analysen zu den Gütekriterien) und als clearing Stelle/Ombudsmann für Testentwickler und Anwender fungiert. Langfristig wäre damit der Grundstein für eine Organisation gelegt, die in Deutschland vergleichbare Aufgaben wahrnehmen kann wie der ETS in Nordamerika (http://www.ets.org) oder die NFER in Großbritannien (http://www.nfer.ac.uk). Eine derartige Organisation stellt jedoch selbst bei optimistischer Betrachtung bestenfalls den Endpunkt einer sehr langfristigen Entwicklung dar.

Fazit

Ausgehend von den Vorzügen und Nachteilen vorhandener nationaler und internationaler Testinformations- und Beurteilungssysteme sowie vor dem Hintergrund der DIN 33430, wurde in dem vorliegenden Artikel die Notwendigkeit eines Neubeginns bei der Information über Tests und der Beurteilung der Qualität von Tests aufgezeigt. Mit dem neuen TK System steht ein tragfähiger Ansatz zur Verfügung, der allerdings nicht risikofrei ist. Es wird nun vor allem darauf ankommen, dass das TK in kurzer Zeit genügend hochqualifizierte Rezensenten findet, die sich dem Aufwand der Testrezensionserstellung nach dem neuen System unterziehen. Förderlich hierfür wäre, neben der Honorierung, vor allem die bislang versagte Anerkennung von Testrezensionen als hochrangige wissenschaftliche Leistung. Diese Anerkennung in der Wissenschaft sollte aber nicht über die Zielgruppe der Testbeurteilungen hinweg täuschen: Die Testbeurteilungen sollten so verfasst und publiziert werden, dass sie auch die für quantitativ bedeutsame Testanwendungen entscheidungsverantwortlichen Praktiker erreichen. Das Ziel muss es außerdem sein, dass die Anzahl hochwertiger Testbeurteilungen mit der Zeit steigt und nicht sinkt. Von Bedeutung ist in diesem Kontext auch, inwieweit parallel zu dem TK System andere Testrezensionen weiter bestehen und inwieweit sich diese anderen Testrezensionen (z. B. in Fachzeitschriften) an dem neuen TK System orientieren werden. Auch muss das neue deutsche System in die internationalen Initiativen (siehe Kersting & Hornke, 2006) integriert werden. Abzuwarten bleibt, welche Konsequenzen eine „offizielle" negative Testbeurteilung nach sich zieht. Zu hoffen ist, dass in Folge von Anerkennung und Kritik die Testqualität insgesamt langfristig steigt. Zu befürchten sind (auch rechtlich ausgetragene) Streitigkeiten. Bei allen zweifellos vorhandenen Bedenken kommt es nun aber darauf an, den Anfang zu wagen und das neue System nicht fachintern zu zerreden, noch bevor es fachübergreifend Wirkung erzielen kann. Das neue Testbeurteilungssystem verdient zumindest eine Chance: Die, sich zu bewähren und sich kontinuierlich zu verbessern.

Literatur

Amelang, M. (1999). Rechenschaftsbericht des Präsidenten der Deutsche Gesellschaft für Psychologie Prof. Dr. Manfred Amelang über die Amtsperiode 1996–1998. *Psychologische Rundschau, 50*, 40–58.

Amelang, M. & Schmidt-Atzert, L. (2006). *Psychologische Diagnostik und Intervention* (4., Aufl.). Heidelberg: Springer.

Bartram, D. (2001). Guidelines for test users: A review of national and international initiatives. *European Journal of Psychological Assessment, 17*, 173–186.

Bierhoff, H.-W., Funke, J., Reips, U.-D. & Weichselgartner, E. (2005). Information und Kommunikation 2005. Ein Lagebericht und einige Zukunftsperspektiven. *Psychologische Rundschau, 56*, 212–219.

Braden, J. P. & Niebling, B. C. (2005). Using the Joint Test Standards to Evaluate the Validity Evidence for Intelligence Tests. In D. P. Flanagan & P. L. Harrison (Ed.), *Contemporary Intellectual Assessment. Theories, Tests, and Issues.* (pp. 615–630). New York: The Guilford Press.

Bös, K. (Hrsg.). (2001). *Handbuch Motorische Tests* (2., vollständig überarbeitete und erweiterte Auflage). Göttingen: Hogrefe.

Brähler, E., Holling, H., Leutner, D. & Petermann, F. (2002). *Brickenkamp Handbuch psychologischer und pädagogischer Tests* (3., vollständig überarbeitete und erweiterte Auflage, Band 1 und 2). Göttingen: Hogrefe.

Brähler, E., Schuhmacher, J. & Strauß, B. (Hrsg.). (2002). *Diagnostische Verfahren in der Psychotherapie* (Band 1). Göttingen: Hogrefe.

Cisek, G., Schäkel, U. & Scholz, J. (Hrsg.). (1989). *Instrumente der Personalentwicklung auf dem Prüfstand.* Hamburg: Windmühle.

DIN (2002). *DIN 33430: Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen.* Berlin: Beuth.

Dunckel, H. (Hrsg.) (1999). *Handbuch psychologischer Arbeitsanalyseverfahren.* Zürich: vdf, Hochschulverlag an der ETH Zürich.

Erpenbeck, J. & Rosenstiel, L. v. (Hrsg.). (2003). *Handbuch Kompetenzmessung.* Stuttgart: Schäffer-Pöschel.

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

Evers, A. (2001 b). The Revised Dutch Rating System for Test Quality. *International Journal of Testing, 1,* 155–182.

Fay, E. (1996). *Tests unter der Lupe, Band 1.* Heidelberg: Asanger.

Fay, E. (Hrsg.). (1999). *Tests unter der Lupe, Band 2.* Lengerich: Pabst.

Fay, E. (Hrsg.). (2000). *Tests unter der Lupe, Band 3.* Lengerich: Pabst.

Fay, E. (Hrsg.). (2003). *Tests unter der Lupe, Band 4.* Göttingen: Vandenhoeck & Ruprecht.

Fay, E. (Hrsg.). (2005). *Tests unter der Lupe, Band 5.* Göttingen: Vandenhoeck & Ruprecht.

Fischer, G. H. (1968). *Einführung in die Theorie psychologischer Tests: Grundlagen und Anwendungen.* Bern: Huber.

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

Göhs, N. & Dick, M. (2001). Testverfahren bei der Personalauswahl. Qualitätssuche im intransparenten Markt. *Personal, Heft 1,* 46–48.

Grubitzsch, S. (Hrsg.). (1999). *Testtheorie – Testpraxis. Psychologische Tests und Prüfverfahren im kritischen Überblick.* (2. Auflage der vollständig überarbeiteten und erweiterten Neuausgabe 1991). Eschborn: Klotz.

Häcker, H., Leutner, D. & Amelang, M. (1998). *Standards für pädagogisches und psychologisches Testen.* Göttingen: Hogrefe.

Hager, W. (2005). Vorgehensweisen in der deutschsprachigen psychologischen Forschung. Eine Analyse empirischer Arbeiten der Jahre 2001 und 2002. *Psychologische Rundschau, 56,* 191–200.

Heyse, H. & Kersting, M. (2004). Anforderungen an den Prozess der Eignungsbeurteilung. In L. F. Hornke & U. Winterfeld (Hrsg.), *Eignungsbeurteilungen auf dem Prüfstand: DIN 33430 zur Qualitätssicherung* (S. 29–41). Heidelberg: Spektrum Akademischer Verlag.

Holling, H., Preckel, F. & Vock, M. (2004). *Intelligenzdiagnostik.* Göttingen: Hogrefe.

Hossiep, R., Paschen, M. & Mühlhaus, O. (2000). *Persönlichkeitstests im Personalmanagement.* Göttingen: Hogrefe.

Hunter, J. E. & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96,* 72–98.

Jäger, R. S. (2004). Test im Test. Insights MDI – Wissenschaftlich betrachtet. *Personal Magazin, 1,* 22.

Kanning, U. P. & Holling, H. (Hrsg.). (2002). *Handbuch personaldiagnostischer Instrumente.* Göttingen: Hogrefe.

Kersting, M. (2006). *„DIN Screen" – Leitfaden zur Kontrolle und Optimierung der Qualität von Verfahren und deren Einsatz bei beruflichen Eignungsbeurteilungen.* Lengerich: Pabst Science Publishers.

Kersting, M. & Heyse, H. (2004). Anforderungen an die Qualität der Verfahren. In L. F. Hornke & U. Winterfeld (Hrsg.), *Eignungsbeurteilungen auf dem Prüfstand: DIN 33430 zur Qualitätssicherung* (S. 43–54). Heidelberg: Spektrum Akademischer Verlag.

Kersting, M. & Hornke, L. F. (2003). Qualitätssicherung und -optimierung in der Diagnostik: die DIN 33430 und notwendige Begleit- und Folgeinitiativen. *Psychologische Rundschau, 54,* 175–178.

Kersting, M. & Hornke, L. F. (2006). Improving the Quality for Proficiency Assessment: The German Standardization Approach. *Psychology Science, 48,* 85–98.

Kersting, M. & Püttner, I. (2006). Personalauswahl: Qualitätsstandards und rechtliche Aspekte. In H. Schuler (Hrsg.), *Lehrbuch der Personalpsychologie* (2. neubearbeitete Auflage) (S. 841–861). Göttingen: Hogrefe.

Klimmer, M. & Neef, M. (2005). Einsatz von Persönlichkeitstypologien in der deutschen Wirtschaft. *Wirtschaftspsychologie aktuell, 12,* 31–34.

Kubinger, K. D. (2006). *Psychologische Diagnostik.* Göttingen: Hogrefe.

Kubinger, K. D. (1997). Editorial zum Themenheft >Testrezensionen: 25 einschlägige Verfahren<. *Zeitschrift für Differentielle und Diagnostische Psychologie, 18,* 13.

Plake, B. S. & Impara, J. C. (2001). *The fourteenth mental measurements yearbook.* Lincoln, NB: The Buros Institute of Mental Measurement.

Punter, J. F. & Kubinger, K. D. (2002). Was ist aus der Kritik der „Testrezensionen: 25 einschlägige Verfahren" (Zeitschrift für Differentielle Psychologie und Diagnostische Psychologie, 18, Heft 1–2) geworden? *Psychologie in Österreich, Heft 2–3,* 24–33.

Sarges, W. & Wottawa, H. (2004). *Handbuch wirtschaftspsychologischer Testverfahren* (2., überarbeitete und erweiterte Auflage). Lengerich: Pabst.

Schimmel-Schloo, M., Seiwert, L. J. & Wagner, H. (Hrsg.). (2002). *PersönlichkeitsModelle.* Offenbach: Gabal.

Schmidt, F. L. (1992). What do data really mean? Research findings, Meta-Analysis, and cumulative knowledge in Psychology. *American Psychologist, 47,* 1173–1181.

Steck, P. (1997). Psychologische Testverfahren in der Praxis. *Diagnostica, 43,* 267–284.

Testkuratorium der Föderation Deutscher Psychologenverbände (in Vorbereitung). *Richtlinien des Testkuratoriums für die Beurteilung von Tests zur Erfassung menschlichen Erlebens und Verhaltens.*

Wehner, E. G. & Durchholz, E. (1980). *Persönlichkeits- und Einstellungstests.* Stuttgart: Kohlhammer.

Wottawa, H. (2002). Einige wichtige Entwicklungen der Psychologischen Diagnostik im letzten Jahrzehnt. *Psychologie in Österreich, 2–3,* 3–5.

Dr. Martin Kersting

Institut für Psychologie der Rheinisch-Westfälischen
Technischen Hochschule Aachen
Jägerstraße 17–19
52056 Aachen
E-Mail: martin@kersting-internet.de