Kersting, M. (2010). Tests und Persönlichkeitsfragebogen in der Personalarbeit. Personalführung, 10/2010, 20-31.

Tests und Persönlichkeitsfragebogen in der Personalarbeit

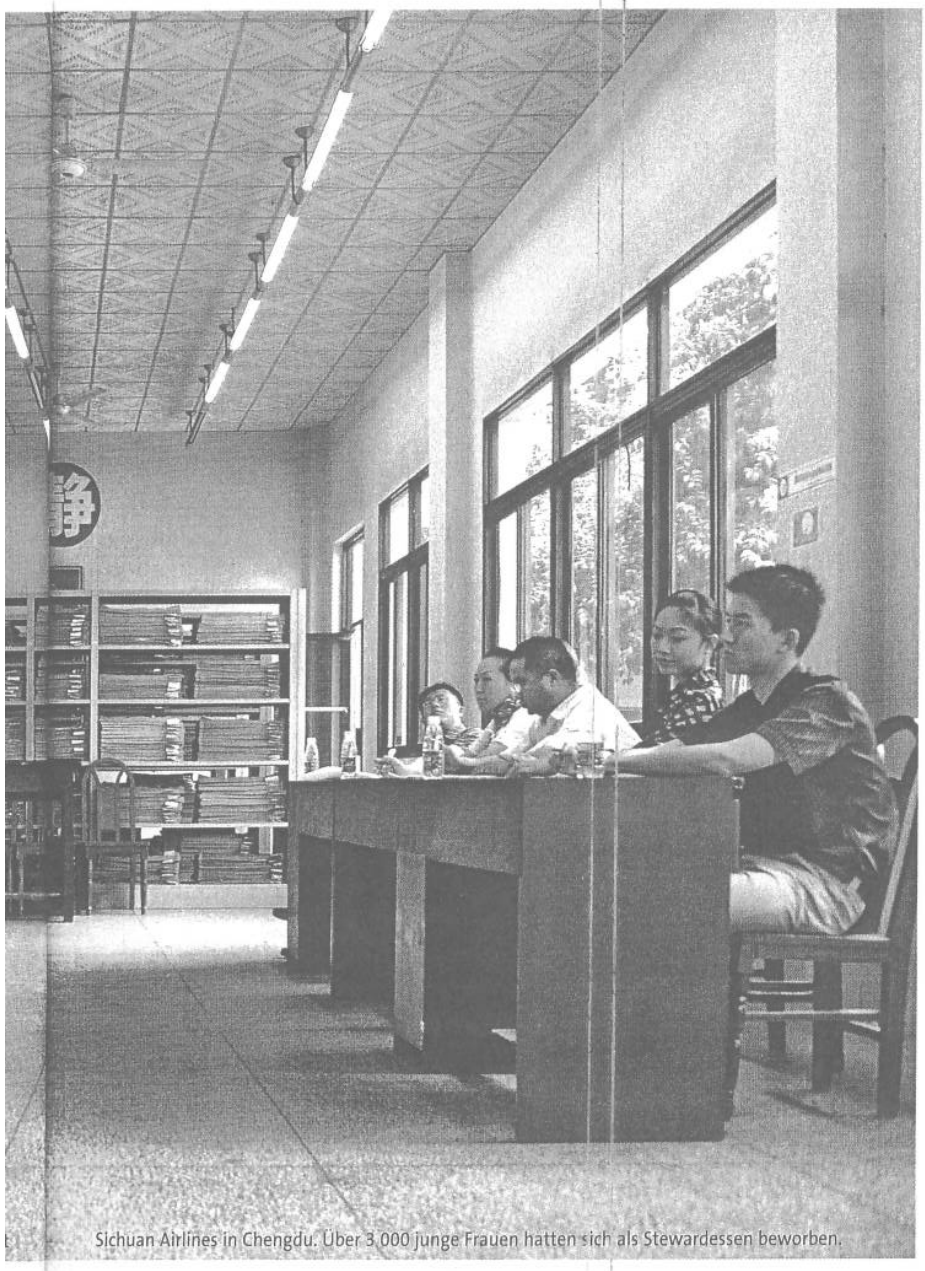
Die Qual der Wahl –

Erfolg oder Misserfolg von Organisationen werden in Zukunft in noch höherem Maße als bisher schon von der Qualität der Personaldiagnostik abhängen. Je mehr Unternehmen Kandidaten aus sogenannten Randgruppen auswählen müssen – Bewerber mit ungewöhnlichen Berufsbiografien oder solche, deren Eignung sich nicht auf den ersten Blick offenbart –, umso sorgfältiger müssen Auswahlverfahren gestaltet werden. Tests dürften daher in Zukunft eine größere Rolle spielen als bisher. Der Autor erläutert zunächst grundsätzliche Kriterien der Testbeurteilung und beschreibt Indikatoren für gute und mangelhafte Testqualität. Anschließend stellt er eine praktische Checkliste und ein neues Testbeurteilungssystem vor.

Aufstellen zum Gruppeninterview hieß es letztes Jahr bei der regionalen Fluggesellschaft.

ZUR PERSON Professor Dr. Martin Kersting ist Diplom-Psychologe und war über zehn Jahre als Berater und Experte für Eignungsdiagnostik tätig. Acht Jahre lang arbeitete Kersting, der Mitglied des Arbeitskreises Assessment Center, der DIN Kommission sowie des Testkuratoriums ist, an der RWTH Aachen. Seit 2008 ist er Professor an der Fachhochschule des Bundes in Münster. Kersting ist Mitautor der DIN 33430 und Autor verschiedener Tests zu kognitiven und sozialen Kompetenzen sowie zahlreicher Publikationen zu wirtschafts- und personalpsychologischen Themen.

Kriterien für eine gute Praxis



Sichuan Airlines in Chengdu. Über 3.000 junge Frauen hatten sich als Stewardessen beworben.

Es gab einmal eine Zeit, in der Organisationen den Personalbedarf durch Neueinstellungen bedienen konnten und dabei unter einer Vielzahl a priori geeigneter Personen wählen konnten. Damals reichte vielleicht ein freies Vorstellungsgespräch noch aus. Diese Zeiten sind vorbei. Der Teich, in denen nur High Potentials schwimmen, ist weitgehend leer gefischt. Nun gilt es, die reichlich vorhandenen, bislang aber vernachlässigten Potenziale auszuschöpfen, zum Beispiel Personen mit ungewöhnlichen oder diskontinuierlichen Erwerbsbiografien, Frauen oder Deutsche mit Migrationshintergrund, Ausländer sowie Personen mit unentdecktem Potenzial in der eigenen Belegschaft. Hier reicht das Schmidtsucht-Schmidtchen-Prinzip, das freie Vorstellungsgespräch, das vor allem stereotype Voreinstellungen bestätigt, nicht mehr aus.

Für die Personalauswahl und Potenzialbeurteilung bieten Tests und Persönlichkeitsfragebogen (im Folgenden zusammenfassend unter dem Begriff Tests behandelt) eine wertvolle Ergänzung. Einige Tests haben eine sehr hohe Aussagekraft. Die Gültigkeit (Validität) eines Tests zur kognitiven Kompetenz bezüglich der Vorher-

sage des Berufserfolgs ist beispielsweise mit r = .66 (Kramer 2009) deutlich höher als die Gültigkeit von strukturierten Eignungsinterviews mit r = .44 (McDaniel et al. 1994) und Assessment-Centern mit r = .41 (Holzenkamp et al. 2010). Tests können unter anderem zur Förderung der Selbsterkenntnis, zur Personalauswahl, Potenzialanalyse und Personalentwicklung genutzt werden.

Es ist daher grundsätzlich zu begrüßen, dass die Nachfrage nach Tests in Deutschland seit etwa 1990 deutlich gestiegen ist (Wottawa 2002). Bedenken gegen die angeblich mangelhafte Akzeptanz von Leistungstests sowie gegen die angeblich verfälschungsbedingte mangelhafte Aussagekraft von Persönlichkeitsfragebogen haben sich als weitgehend unzutreffend erwiesen (vgl. Kersting 2008b; Marcus 2003). Der Testmarkt ist allerdings intransparent. Weder kann man rasch und einfach erfahren, welche Tests wo zu welchen Konditionen angeboten werden, noch finden sich ohne Weiteres unabhängige Bewertungen der Qualität der angebotenen Tests. Die Qualität der Tests ist sehr heterogen und variiert darüber hinaus in Abhängigkeit von den jeweiligen Einsatzzielen.

Qualitätsmerkmale

Ein Test muss unter anderem 1.) objektiv, 2.) zuverlässig und 3.) gültig sein. Ein weiterer Qualitätsaspekt ist 4.) die Normbasis. Keines der genannten Kriterien kann für sich allein betrachtet werden, sondern die Beurteilung muss vor dem Hintergrund des jeweiligen Einsatzzwecks erfolgen. Aus diesem Grund basiert die Testauswahl notwendigerweise auf der Anforderungsanalyse beziehungsweise dem Kompetenzmodell der Organisation. Der Test muss solche Dimensionen erfassen, die anforderungsrelevant sind. Die häufig gestellte Frage, welcher Test ‚gut' sei, offenbart mangelhafte Sachkenntnis: Ein Test ist nicht an sich gut, sondern nur gut für etwas, und dieses Etwas muss zunächst bestimmt werden. Erst wenn man sich über die Ziele der Testung im Klaren ist, kann für die Verfahren, die potenziell zu den Zielen beitragen können, die Qualitätsbetrachtung beginnen.

Objektivität Das Testergebnis muss unabhängig von den Personen sein, die den Test a) durchführen, b) auswerten und das Testergebnis c) interpretieren. Während die Durchführungs- und Auswertungsobjektivität bei den meisten Testverfahren gegeben sind, hapert es häufig bei der Interpretationsobjektivität. Einem Test, dessen Ergebnis von unterschiedlichen Personen unterschiedlich gedeutet oder gelesen wird, mangelt es an Interpretationsobjektivität. Gerade die bei vielen Anwendern beliebten ‚freien' Interpretationen der Testergebnisse sind zu vermeiden.

Zuverlässigkeit Bei der Zuverlässigkeit (Reliabilität) geht es um die Frage nach Messfehlern. Jede Messung ist mit einem Fehler behaftet. Je kleiner der Messfehler ist, desto größer ist die Zuverlässigkeit. Um die Zuverlässigkeit eines Tests zu bestimmen, wird der Zusammenhang zwischen zwei oder mehreren Testwerten, die dasselbe erfassen, berechnet: zum Beispiel a) zwischen zwei Messungen, die zu unterschiedlichen Zeitpunkten durchgeführt wurden (Retest-Reliabilität), b) zwischen zwei Formen des gleichen Tests (Paralleltest-Reliabilität) oder c) zwischen einzelnen Fragen des Tests (split-half bzw. interne Konsistenz). Seriöse Tests täuschen kein punktgenaues Ergebnis vor, sondern nennen einen Wertebereich, in dem der ‚wahre' Wert mit einer gewissen Wahrscheinlichkeit (z. B. 95 %) liegt (das sog. Vertrauensintervall).

Gültigkeit Die Gültigkeit (Validität) bezieht sich genau genommen nicht auf den Test, sondern thematisiert die Aussagekraft der aus dem Test abgeleiteten Schlussfolgerungen. Repräsentieren die Fragen und Aufgaben des Tests das interessierende Merkmal (Inhaltsvalidität)? Erfasst der Test das, was er erfassen soll, und können die Testergebnisse im Sinne einer überzeugenden Theorie erklärt werden (Konstruktvalidität)? Schließlich: Kann man mit dem Test irgendein relevantes Verhalten in der künftigen Realität vorhersagen (Kriteriumsvalidität)?

Ein Test ist untauglich, wenn er nicht das Merkmal erfasst, was er zu erfassen vorgibt. Er ist auch untauglich, wenn die zugrunde gelegte Theorie nicht (mehr) dem aktuellen Stand der Erkenntnis entspricht. In der Praxis erfreuen sich beispielsweise sogenannte Typentests (z. B. D-I-S-G®, GPOP®, INSIGHTS MDI® oder MBTI®) einer hohen Beliebtheit. Einige Typentests gehen auf hoffnungslos veraltete Theorien (z. B. Carl Gustav Jung oder Wil-

liam Moulton Marston) zurück, für deren Gültigkeit bis heute keine empirischen Belege erbracht werden konnten (Jäger 2004). Während mit Typentests, entsprechend dem vorempirischen Stand der damaligen Wissenschaft, häufig lediglich die grobe (bipolare) Zugehörigkeit einer Person zu einem nur aus Plausibiliätsgründen konstruierten Typ festgelegt wird, ist es mit Persönlichkeitsfragebogen zu aktuellen Theorien (wie z. B. dem „Fünf Faktoren Modell der Persönlichkeit") möglich, die Ausprägung einer Persönlichkeitseigenschaft auf verschiedenen, mehr oder minder unabhängigen kontinuierlichen Persönlichkeitsdimensionen zu bestimmen, deren Zusammenhang mit berufsrelevanten Erfolgsfaktoren empirisch nachgewiesen wurde.

Andere Tests wie das HBDI® oder das Struktogramm (Biostrukturanalyse®) legen ihrer Interpretation Annahmen über den Zusammenhang von Hirnarealen und Denkmustern zugrunde. Auch hier müsste natürlich die Theorie dem aktuellen Stand der Forschung entsprechen, und es müsste gewährleistet sein, dass man mit einem Fragebogen überhaupt Indikatoren für die Physiologie des Gehirns gewinnen kann. Während kaum eine Organisation auf die Idee kommen würde, anstelle moderner Kommunikationstechnologien die Schreibmaschine und die Rohrpost zu nutzen, gibt es dennoch Organisationen, die in der Personaldiagnostik auf die antiquierte Typenlehre oder auf völlig veraltete Annahmen über die Funktionsweise des Gehirns setzen. Wer im Industriezeitalter einen Rückgriff auf Theorien vornimmt, die sich wissenschaftlich als überholt und unzutreffend erwiesen haben, demonstriert den getesteten Personen (internen und externen Kandidaten), dass die Personalarbeit der Organisation in der Vergangenheit stecken geblieben ist.

Vorstellung in der Shopping-Mall – diesen ungewöhnlichen Ort für Interviews mit Bewerberinnen wählte die China Southern Airlines Mitte letzten Jahres in Shenyang. Rund 80 Stewardessen wollte das Unternehmen rekrutieren.

Qualitätsmerkmale im Kontext

Die drei Gütekriterien Objektivität, Zuverlässigkeit und Gültigkeit stehen in einer Abhängigkeitsbeziehung zueinander (Abb. 1). Ein Test, der keine Objektivität aufweist, kann nicht zuverlässig und gültig sein. Die Zuverlässigkeit wiederum ist eine notwendige Voraussetzung für die Gültigkeit. Insgesamt können Tests bezüglich aller drei genannten Gütekriterien hervorragende Werte erzielen. Gerade die hohe Objektivität vieler Tests ist hier von entscheidender Bedeutung.

Es erscheint zunächst kontraintuitiv, dass ein ‚simpler' Test zum Ankreuzen so viel treffsicherere Aussagen ermöglicht als ein aufwendiges Interview oder AC. Ein wesentlicher Grund ist die Objektivität. Egal wer den Test durchführt, auswertet und interpretiert: Alle kommen übereinstimmend zu dem gleichen Ergebnis. Zeichnet man hingegen ein Interview oder eine AC-Übung mit einem Kandidaten auf Video auf und zeigt dieses mehreren Beobachtern, wird die gleiche Leistung von unterschiedlichen Personen ganz uneinheitlich bewertet werden, selbst wenn die Beobachter vorher gut geschult wurden. Diese Verfahren sind weniger objektiv als Tests. Fehlt aber die Objektivität, schränkt dies notwendigerweise auch die Zuverlässigkeit und Gültigkeit massiv ein. Gerade der Aspekt, der Test-Kritikern reduktionistisch vorkommt, das stupide ‚Kreuzchen-Zählen' bei Tests, ist ein wichtiger Grundpfeiler der Leistungsfähigkeit.

Pseudorationale Testbeurteilungen

Die Ausprägung der Zuverlässigkeit und Gültigkeit wird in der Regel als Koeffizient beziffert, der Werte zwischen 0,00 und 1,00 annehmen kann. Es ist aber nicht sinnvoll, die Güte eines Tests allein aufgrund der numerischen Ausprägung von Kennwerten zu beurteilen oder zwei Tests allein anhand dieser Zahlen miteinander zu vergleichen. Die erzielten Kennwerte hängen nicht nur vom Test, sondern auch von den Umständen der jeweiligen Untersuchung ab. Ganz wesentlich ist zum Beispiel, dass Testwerte möglichst stark variieren. Setzt man zum Beispiel einen Gedächtnistest bei 100 zufällig ausgewählten Personen ein und einen anderen Gedächtnistest bei 100 Gedächtniskünstlern, so wird die letztgenannte Gruppe kaum Fehler im Test aufweisen, das heißt, die Testwerte variieren nicht, und folglich fallen die Koeffizienten der Testgütekriterien gering aus. Dies liegt aber nicht am Test – ein anderer Test hätte in dieser Gruppe auch versagt.

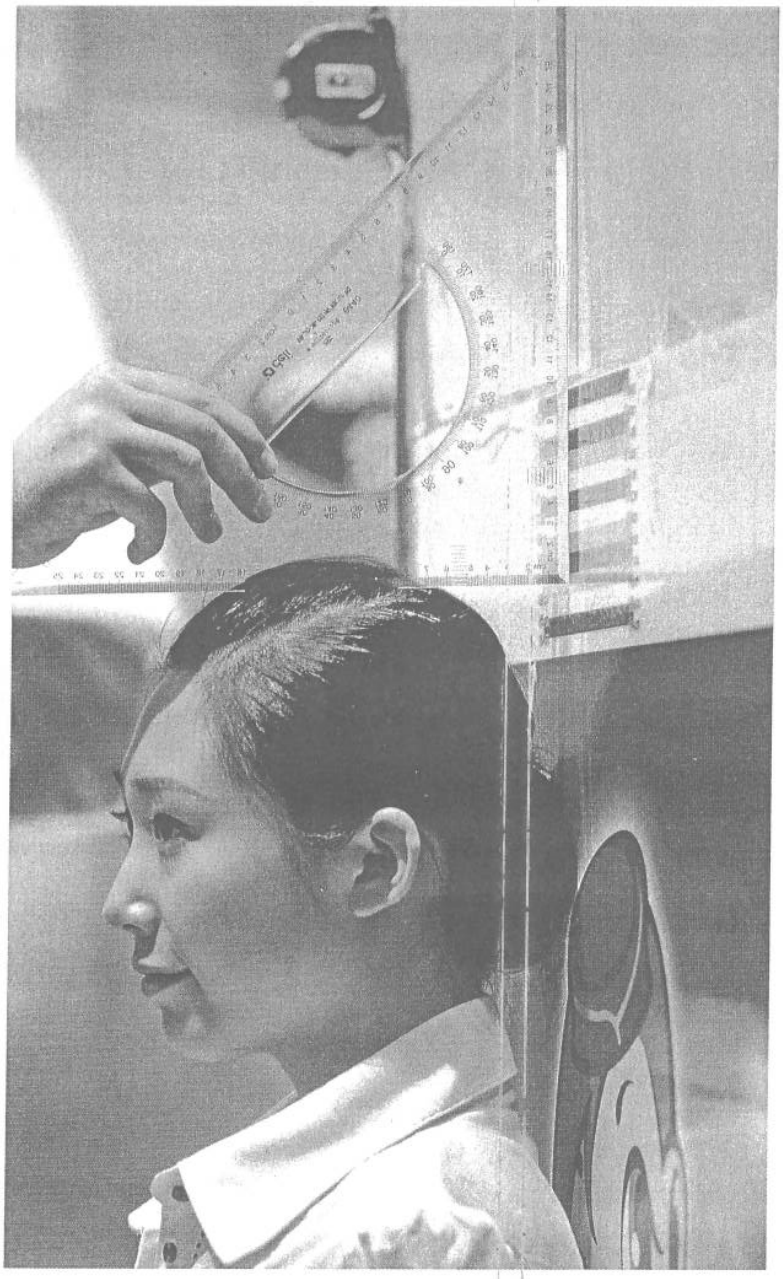
Die Kriterien sind also sachgerecht, nicht schematisch anzuwenden. Darüber hinaus handelt es sich bei Kennwerten lediglich um Einzelschätzungen. Es gibt mechanische Bewertungssysteme, die leider auch Eingang in wissenschaftliche Lehrbücher gefunden haben, die zum Beispiel behaupten, ein guter Test müsse eine Zuverlässigkeit von mindestens r = .80 aufweisen. Derartige mechanische Systeme sind pseudorational und verkennen das Zusammenspiel zwischen dem Test und seinen Einsatzbedingungen. Ein Test hat nicht einen bestimmten Kennwert, sondern in unterschiedlichen Situationen fallen die Gütekriterien desselben Tests unterschiedlich aus (vgl. hierzu auch Kersting 2006).

Normen

Die Auswertung und Interpretation der meisten Tests beruht auf einem normativen Ansatz, das heißt, man vergleicht den gemessenen Wert mit den Werten anderer Personen. Daher muss man wissen, welche Vergleichswerte für die Auswertung und Interpretation herangezogen werden können. Gut ist ein Test unter diesem Gesichtspunkt, wenn die Normdaten in jüngerer Zeit an einer sehr großen Gruppe von Personen erhoben wurden, die der Zielgruppe des Verfahrens entsprechen.

Beurteilungsbasis: Zahlen, Daten und Fakten

Um einen Test beurteilen zu können, bedarf es Informationen darüber, in welchem Umfang die genannten sowie weitere Gütekriterien (z. B. Ökonomie, Akzeptanz) erfüllt sind. Ein Großteil dieser Informationen besteht notwendigerweise aus Berichten über die Ergebnisse empirischer Studien. Es ist nicht möglich, einen Test allein aufgrund seiner Aufgaben oder Fragen zu beurteilen. Für zahlreiche Tests fehlen die notwendigen Informationen. Die Situation auf dem Testmarkt ist vergleichbar damit, dass Kunden einen Pkw kaufen und nicht erfahren, ob er Benzin oder Diesel verbrennt, wie viel Kraftstoff er verbraucht oder wie stark der Motor ist. Stattdessen werden potenzielle Kunden in Hochglanzbroschüren mit inhaltsleeren Aussagen abgefertigt; häufig steht „Testinformation" drauf, wo nur Allgemeinplätze drinstecken. Die Mitteilung, mit



Eine eindeutige Angelegenheit ist das Messen der Körpergröße, der sich eine Bewerberin bei der China Southern Airlines unterzieht. Bei vielen psychologischen Tests ist fraglich, ob sie überhaupt messen, was sie vorgeben, zu messen.

einem Test sei für „zwei Millionen Menschen weltweit" oder für „mehrere Tausend Topmanager" eine Analyse durchgeführt worden, sagt nichts über die tatsächlich vorliegende Normbasis aus, denn die Daten einer Testdurchführung fließen nicht automatisch in die Normgruppe ein. Auch Aussagen über die Treffsicherheit und Gültigkeit des Verfahrens sind wertlos, solange sie nicht empirisch belegt sind.

Zirkustricks statt Gültigkeitsbelege?

Leider verlassen sich einige Anwender bei der Testauswahl auf ihr subjektives ‚Evidenzgefühl'. Das kommt so zustande: Man füllt den Test selbst aus, bittet Vertraute, den Test auszufüllen, und befindet dann, ob das erstellte Gutachten ‚passt'. Tatsächlich sind fast alle Anwender überzeugt, dass das testbasierte Gutachten wirklich zutreffend ist. Wer ein derart ‚stimmiges' Gutachten liest, stellt keine Fragen mehr nach der empirischen Basis. Schon vor über 60 Jahren wurde nachgewiesen, dass Menschen davon überzeugt sind, ein vermeintliches Test-Gutachten würde ihre individuelle Persönlichkeit überaus zutreffend beschreiben, auch wenn ihnen in Wirklichkeit ein Standardtext vorgelegt wurde, der nichts mit ihren individuellen Testergebnissen zu tun hatte. Der Effekt wurde nach dem berühmten Zirkusgründer „Barnum-Effekt" genannt. Dies funktioniert, indem das Gutachten allgemeingültige, unverbindliche und selbstwertdienliche (also positive) Phrasen nutzt, in denen sich die meisten Menschen wiederfinden (z. B.: ‚Sie sind mit viel Begeisterung und Elan bei der Sache, aber verfügen auch über das notwendige Feingefühl und die Geduld für die schwierigen Fälle.'). Die Vorschläge, die aus solchen ‚Persönlichkeitstests' abgeleitet werden, sind ebenso sinnfrei (z. B.: ‚Auch wenn Sie immer neue Projekte auftun, sollten Sie nicht vergessen, die Deadlines der laufende Projekte einzuhalten.')

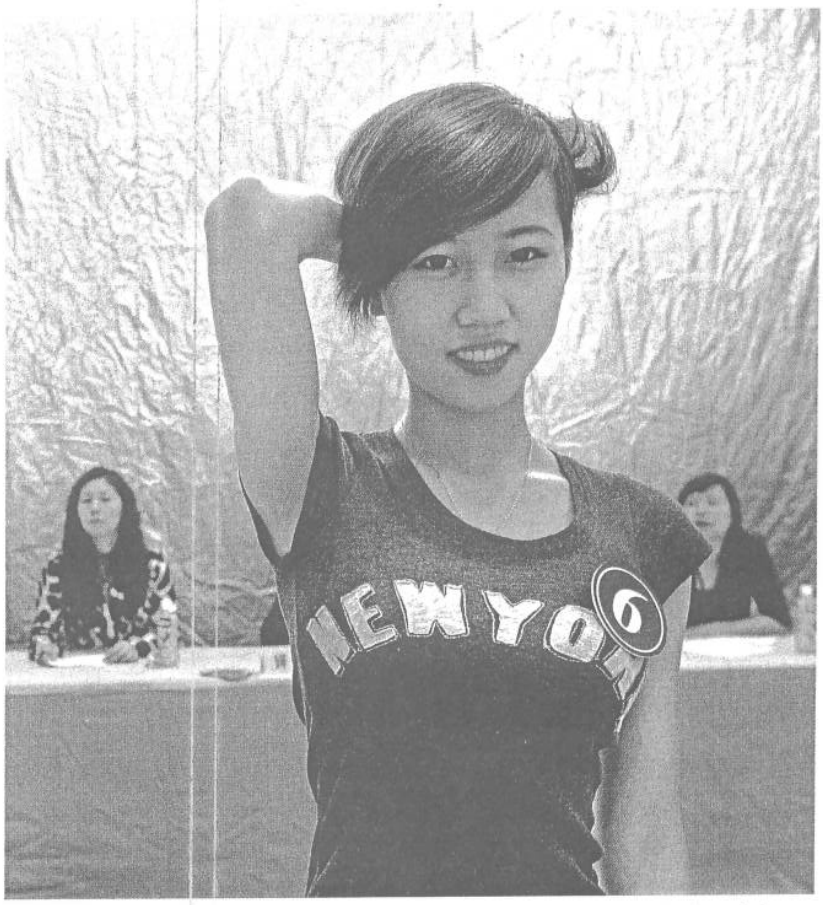
Prüfung der Aussagekraft

Mit modernen Techniken kann man Anwender noch erheblich besser grundlos in helle Freude versetzen, indem sich im Gutachten einfach Umformulierungen der Aussagen wiederfinden, denen die getestete Person zugestimmt hat. Stimmt die getestete Person der Testaussage „Ich bin nicht unbedingt daran interessiert, eine leitende Position zu haben" zu, generiert der Computer ein Gutachten mit dem Satz: „Eine Karriere als Führungskraft ist Ihnen nicht so wichtig" – und der Teilnehmer staunt, wie gut das Gutachten seine wahre Persönlichkeit wiedergibt. Auf das subjektive Evidenzgefühl ist also kein Verlass, entscheidend ist die Frage, ob die aus dem Test abgeleiteten Schlussfolgerun-

gen geeignet sind, Vorhersagen über das Verhalten in der Zukunft zu treffen.

Kann man mit einem Test zum Beispiel vorhersagen, welche Personen im Vertrieb gemessen an harten Daten erfolgreicher sein werden als andere? Führt die empfohlene Zusammensetzung von Teams mit unterschiedlichen „Typen" nachweislich zu besseren Ergebnissen als die bis dato praktizierte Teamzusammenstellung? Dies lässt sich in empirischen Studien prüfen, gute Tests berichten detailliert über die Ergebnisse dieser Studien. Das sollte nicht pauschal („in zahlreichen Studien hat sich der Test bewährt…"), sondern konkret erfolgen: Im Jahr X wurde eine Studie mit Z Personen im Alter von… bis… durchgeführt, als Kriterium für den Berufserfolg galt Y (usw.). Die Basis für die Testbeurteilung sind Zahlen, Daten und Fakten.

Notwendig sind detaillierte Berichte über die Ergebnisse dieser (im Idealfall extern durchgeführten) empirischen Studien. Will man sich vor Scharlatanerie schützen, muss man in der Personaldiagnostik wie in der Medizin evidenzbasiert vorgehen, also nur Aussagen vertrauen, die sich auf belastbares empirisches Beweismaterial stützen.



Alles eine Frage der Haltung: Bewerberin bei der China Eastern Airlines in Wuhan, Provinz Hubai.

Wichtige Fragen an Testanbieter

Welche Zahlen, Daten und Fakten werden konkret benötigt, um einen Test beurteilen zu können? Dieser Frage hat sich das Testkuratorium gewidmet, dessen Aufgabe es ist, die Öffentlichkeit vor unzureichenden diagnostischen Verfahren und vor unqualifizierter Anwendung diagnostischer Verfahren zu schützen. Das Testkuratorium hat einen Katalog von Informations-Mindestanforderungen identifiziert. Es handelt sich um eine Menge an definierten Informationen, die Testanbieter beantworten müssen, wenn sie ihren Kunden eine Bewertung des angebotenen Produkts ermöglichen wollen. Bei der Formulierung des Katalogs hat das Testkuratorium die DIN 33430 genutzt, eine Norm zur Personalauswahl, die seit 2002 in Kraft ist und die „Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen" formuliert (Kersting 2008a).

Die Anwendung dieser Norm erscheint auf den ersten Blick unpassend, denn die Norm bezieht sich zum einen auf alle Verfahren der Eignungsbeurteilung (also nicht nur auf Tests, sondern z. B. auch auf Interviews und AC), und zum anderen handelt es sich überhaupt nicht um eine Produkt-, sondern um eine Prozess-Norm. Dennoch formulieren 44 Prozent der in der DIN 33430 getroffenen Aussagen (der insgesamt 318 Aussagen) Inform onsanforderungen an Verfahrenshinw (auch „Handanweisungen" oder in Be auf Tests „Testmanuale" genannt).

Diese Teilmenge der DIN 33430 140 Informations-Anforderungen an T wurde von Kersting (2008a) zu einer genständigen Checkliste „Anforderun an Verfahrenshinweise" zusammengef Die Checkliste gilt offiziell als „Stanc zur Information und Dokumentation Instrumenten zur Erfassung menschlic Erlebens und Verhaltens des Testkur riums der Föderation Deutscher Psyc logenvereinigungen". Beispiele für die gen der Checkliste finden sich in Ab dung 2.[1] Verfahrenshinweise, die di

umfangreichen Informationsanforderungen nicht gerecht werden, entsprechen nicht der DIN 33430.

Schritte zu mehr Transparenz

Ein wesentliches Problem bei der Beurteilung von Tests besteht darin, dass sich etliche Anbieter einer Bewertung ihrer Tests entziehen, indem sie relevante Informationen vorenthalten. Auch Experten können angesichts eines Informationsmangels nicht sagen, welche Qualität dieser und jener Test aufweist. Mit diesem Informationsdefizit räumt die DIN 33430 auf, indem sie umfangreiche Informationen zum Verfahren einfordert. Jede Form von Geheimniskrämerei seitens eines Testanbieters ist ein Alarmsignal. Informationen zum Test sollten frei zugänglich sein, die Grundaussagen bedürfen der Schriftform. Informationen zu einem Test nur in kostenpflichtigen Seminaren zu präsentieren, ist inakzeptabel. Transparenz ist auch notwendig, um den Testteilnehmern die Ergebnisse erklären zu können und beispielsweise die Vergleichsgruppe (Normgruppe) und den Messfehler des Tests zu erläutern.

Die Anzahl von 140 Informationsforderungen erscheint zunächst hoch. Es geht aber nicht darum, dass alle 140 Informationen eingeholt werden, man kann auch Schwerpunkte setzen. Wichtig ist es, überhaupt Fragen (z. B. nach den Ergebnissen empirischer Studien) zu stellen. Testanbieter, die vorgeben, die offizielle Checkliste nicht zu kennen, oder die die Relevanz der geforderten Informationen in Abrede stellen, sollte man meiden. Die Qual der Wahl stellt sich nur mit Blick auf die Anbieter, die die geforderten Informationen beibringen, denn die Informationen allein reichen nicht aus. Sie sind eine notwendige, aber keine hinreichende Bedingung für Qualität. Wenn die Informationen vorliegen, kann man im nächsten Schritt die Qualität beurteilen.

Neues Testbeurteilungssystem

Das Testkuratorium (2010) hat ein neues System zur Information über und Beurteilung von Tests entwickelt. In einem ersten Schritt wird geprüft, ob überhaupt genügend Informationen vorliegen, um den Test zu beurteilen. Grundlage dafür ist die oben beschriebene Checkliste. Ein Test, der in diesem Sinne nicht prüffähig ist, erhält ohne weitere Begutachtungen eine ne-

gative Gesamtbeurteilung. Erst wenn diese Informationen vorliegen, wird eine systematische Erfassung des Tests in einer Datenbank vorgenommen, und die eigentliche Testbeurteilung beginnt. Diese erfolgt auf der Basis einer neu geschaffenen Beurteilungsrichtlinie. In der Richtlinie werden insgesamt sieben Beurteilungskategorien vorgegeben, für vier der sieben Kategorien werden die Urteile zusätzlich zum Freitext auf einer Skala formuliert (vgl. Abb. 3). Diese Beurteilungsskala sieht vier Abstufungen vor.

Unabhängige Begutachtung

Eine weitere Besonderheit des Systems besteht darin, dass jeder Test von zwei Experten beurteilt wird, die zunächst in Unkenntnis und unabhängig voneinander ihre Beurteilungen für die vorgegebenen Beurteilungsaspekte ausarbeiten. Erst wenn die Urteile beider Rezensenten dem Testkuratorium vorliegen, werden die beiden Experten aufgefordert, im zweiten Schritt gemeinsam zu arbeiten, Konsensurteile zu finden und eine endgültige Rezension zu verfassen. Die von den beiden Gutachtern gemeinsam erstellte Rezension wird vom Testkuratorium in anonymisierter Form an die Testautoren geschickt, um ihnen Gelegenheit einzuräumen, innerhalb einer gesetzten Frist Stellung zu beziehen. Das Testkuratorium sichert den Testanbietern die Vertraulichkeit bestimmter Informationen zu. Somit kann sich niemand der Testbeurteilung mit der Begründung entziehen, alle Informationen zum Test fielen unter das zu wahrende „Betriebsgeheimnis". Derartigen Ausreden ist allerdings grundsätzlich mit Skepsis zu begegnen.

Alle nach dem neuen System erarbeiteten testbezogenen Informationen können kostenfrei unter www.zpid.de/Testkuratorium eingesehen werden, hier steht auch der Text des Beurteilungssystems zum Down-

Qualitätskriterien für Tests – die Hierarchie der Gütekriterien

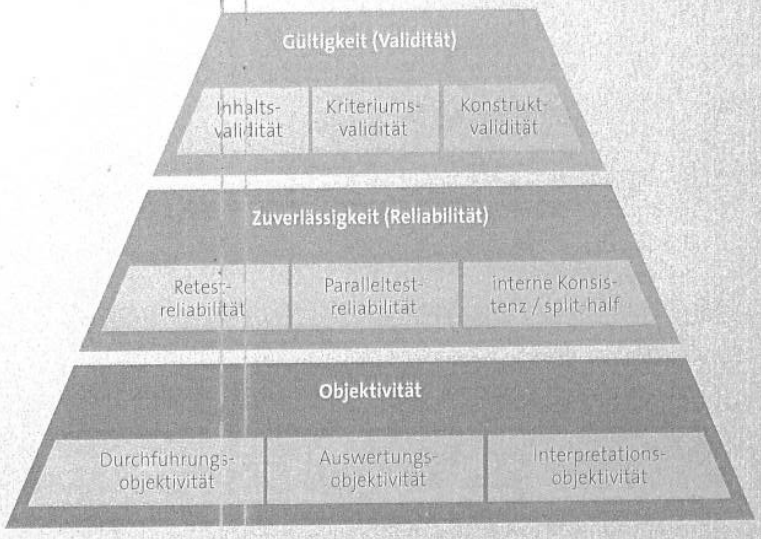


Abb. 1

Beispielhafte Aussagen der DIN-Screen-Checkliste 1 zu Informationsanforderungen an Tests

Nr.	Aussage	ja	nein	nicht zu bewerten	Anmerkungen	Quelle (Seite)
1	Für jedes eingesetzte Verfahren liegen Verfahrenshinweise (Handhabungshinweise) vor. (DIN, S. 6)	○	○	○	○	
33	In den Verfahrenshinweisen werden die Ergebnisse einer (oder mehrerer) empirischen(r) Untersuchung(en) berichtet. (DIN, S. 6)	○	○	○	○	
44	In den Verfahrenshinweisen werden Regeln aufgestellt, wie bei der Auswertung mit nicht bearbeiteten Fragen bzw. (Teil-)Aufgaben umgegangen wird. (DIN, S. 8)	○	○	○	○	
57	Die herangezogenen Normwerte entsprechen der Referenzgruppe der Zielgruppe. (DIN, S. 7)	○	○	○	○	
58	Die Angemessenheit der Normwerte wurde in den letzten acht Jahren überprüft. (DIN, S. 7)	○	○	○	○	
74	Die Zuverlässigkeit wurde über die Retest-Methode bestimmt, oder die Retest-Reliabilität wurde durch einen geeigneten Untersuchungsplan geschätzt. (DIN, S. 15)	○	○	○	○	
80	Aus den Verfahrenshinweisen wird deutlich, welche empirischen Nachweise der Inhalts- und / oder Kriteriums- und / oder Konstruktgültigkeit eine Anwendung des Verfahrens für den intendierten Anwendungszweck rechtfertigen. (DIN, S. 16)	○	○	○	○	

Quelle: Kersting 2008a

Abb. 2

load bereit. Abbildung 3 zeigt beispielhaft die im Auftrag des Testkuratoriums von Höft und Muck (2009) erarbeitete Bewertung für den „Golden Profiler of Personality" (GPOP), einen Persönlichkeitsfragebogen auf der Basis eines Typenmodells. Bezüglich der Gültigkeit (Validität) erfüllt der Test die Anforderungen nicht, da für diesen Test – wie für zahlreiche andere Tests – empirische Studien zur Konstrukt- und Kriteriumsvalidität fehlen.

Beurteilung des Golden Profiler of Personality (GPOP) nach dem Testbeurteilungssystem des Testkuratoriums 2010*

Golden Profiler of Personality (GPOP) Deutsche Adaptation des Golden Personality Type Profiler von John P. Golden	Die TBS-TK-Anforderungen sind erfüllt			
	voll	weitgehend	teilweise	nicht
Allgemeine Informationen, Beschreibung und diagnostische Zielsetzung			X	
Objektivität			X	
Zuverlässigkeit			X	
Validität				X

*Autoren der Test-Rezension: Höft / Muck 2009

Abb. 3

Übertragung auf Einzelfälle

Bislang liegen nur sehr wenige Rezensionen nach dem neuen System vor, und es ist nicht zu erwarten, dass auf absehbare Zeit zu jedem Test eine unabhängige Begutachtung vorliegen wird. Dazu gibt es zu viele Tests, allein in Deutschland sind es mehrere Tausend. Ein Blick in die Beurteilungsrichtlinie und in beispielhafte Rezensionen lohnt dennoch, da dadurch das Prinzip der Testbewertung verständlich und für den eigenen Anwendungsfall analogisierbar wird. Vor allem die Checkliste mit den Fragen an Testanbieter kann und sollte in Bezug auf jeden Test genutzt werden, um zwischen seriösen und unseriösen Anbietern zu differenzieren. Seriöse Testanbieter stellen die laut Checkliste geforderten Informationen in schriftlicher Form zur Verfügung. Sie weisen darüber hinaus auf die Grenzen der Verfahren hin. Auch wenn einige Tests aussagekräftiger als Interviews und AC sind, erfassen Tests doch nur Verhaltensstichproben und erlauben lediglich Prognosen, sie vermitteln keinesfalls Gewissheiten. Testergebnisse sollten nicht isoliert betrachtet werden, sondern im Zusammenhang mit anderen Informationen über die Person und Situation interpretiert werden.

Folgerungen

Um sich ein solides Urteil über einen Test zu bilden, bedarf es einerseits Informationen über den Test und andererseits der Expertise, um aus den Informationen richtige Schlüsse zu ziehen. Notwendig sind unter anderem Grundkenntnisse in Diagnostik, Evaluation und Statistik. Sofern die Expertise zur Testbeurteilung nicht im Rahmen einer Ausbildung oder eines Studium erworben wurde, sollte man entsprechende Fortbildungen absolvieren (vgl. z. B. Kersting 2010). Alternativ kann es sinnvoll sein, sich in dieser Frage externe Expertise einzukaufen. Es ist leider nicht einfach, einen Test auf seine Qualität hin zu testen. Der Aufwand lohnt sich allerdings, denn gute Tests sind ein wertvolles Instrument moderner Personalarbeit.

Summary

Selection Criteria for Tests and Personality Questionnaires

The more companies must select candidates from so-called marginal groups – that is, job applicants with unusual professional biographies or whose suitability is not apparent at first glance – the more carefully selection procedures must be organized. Tests and personality questionnaires provide valuable methodological support in selecting personnel and in assessing job applicant potential. Some tests have very high validity. For example, the validity of cognitive competence tests in predicting the future professional success of an applicant is significantly higher than the validity of structured suitability interviews. As a result, such tests will probably play a more important role in future than up to now. The author initially explains the fundamental criteria of test assessment and then describes indicators for good as well as inadequate test quality. After that, he presents a practical checklist and a new test assessment system.

Anmerkung

1 Die vollständige Checkliste steht unter www.kersting-internet.de/DIN-Buch/din-buch_downloads.html zum freien Download bereit. Darüber hinaus existiert eine Online-Version der Liste (www.kersting-internet.de/DIN-Screen.html).

Literatur

Holzenkamp, M. / Spinath, F. M. / Höft, S. (2010): Wie valide sind Assessment Center im deutschsprachigen Raum? Eine Überblicksstudie mit Empfehlungen für die AC-Praxis, in: Wirtschaftspsychologie, 12 (2), 17–25

Höft, S. / Muck, P. M. (2009): TBS-TK Rezension: „Golden Profiler of Personality (GPOP). Deutsche Adaptation des Golden Personality Type Profiler von John P. Golden", in: Report Psychologie, 34 (7/8), 322–323

Jäger, R. S. (2004): Test im Test. Insights MDI – wissenschaftlich betrachtet, in: PersonalMagazin, 6 (1), 22

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

Kersting, M. (2008a): Qualität in der Diagnostik und Personalauswahl: Der DIN-Ansatz, Göttingen

Kersting, M. (2008b): Zur Akzeptanz von Intelligenz- und Leistungstests, in: Report Psychologie, 33 (9), 420–433

Kersting, M. (2010): Diagnostische Fortbildung am Beispiel des Trainings zur Eignungsbeurteilung nach DIN 33430, in: U. P. Kanning / L. v. Rosenstiel / H. Schuler (Hg.): Jenseits des Elfenbeinturms. Psychologie als nützliche Wissenschaft, Göttingen, 223–240

Kramer, J. (2009): Allgemeine Intelligenz und beruflicher Erfolg in Deutschland: Vertiefende und weiterführende Metaanalysen, in: Psychologische Rundschau, 60 (2), 82–98

Marcus, B. (2003): Das Wunder sozialer Erwünschtheit in der Personalauswahl, in: Zeitschrift für Personalpsychologie, 2 (3), 129–132

McDaniel, M. A. / Whetzel, D. L. / Schmidt, F. L. / Maurer, S. D. (1994): The validity of employment interviews: A comprehensive review and meta-analysis, in: Journal of Applied Psychology, 79 (4), 599–616

Testkuratorium der Föderation Deutscher Psychologenvereinigungen (2010): TBS-TK. Testbeurteilungssystem des Testkuratoriums der Föderation Deutscher Psychologenvereinigungen, in: Psychologische Rundschau, 61 (1), 52–56

Wottawa, H. (2002): Einige wichtige Entwicklungen der Psychologischen Diagnostik im letzten Jahrzehnt, in: Psychologie in Österreich, 22, 2–5