

Personaldiagnostik bei Bewerbern aus anderen Kultur- und Sprachräumen

Andreas Frintrup, Brigitte Flubacher

- 2.1 Einführung in die Personaldiagnostik – 6**
- 2.2 Qualitätsanforderungen von Personalauswahlverfahren – 7**
- 2.3 Statistische Fairness von Personalauswahlentscheidungen – 12**
- 2.4 DIN 33430 – 15**
 - 2.4.1 Ablauf einer berufsbezogenen Eignungsbeurteilung – 16
 - 2.4.2 Qualitätsanforderungen an Auftragnehmer und Durchführende – 17
- 2.5 Sprach- und Kulturgebundenheit diagnostischer Methoden – 17**
- 2.6 Integrationsförderliche Personalauswahlprozesse – 18**

Bei der Gestaltung eines gleichermaßen fairen und qualitativ hochwertigen Verfahrens ist die Kenntnis von Methoden und Verfahren der Personalauswahl sowie deren Qualitätsanforderungen unerlässlich. In den folgenden Abschnitten werden grundsätzliche Qualitätsanforderungen an Personalauswahlverfahren erläutert. Aufgrund seiner hohen Relevanz bei der Auswahl von Bewerberinnen und Bewerbern mit Migrationshintergrund wird zudem auf das Kriterium der statistischen Fairness sowie auf die DIN-Norm 33430 eingegangen.

2.1 Einführung in die Personaldiagnostik

Berufsbezogene Auswahlentscheidungen sind sowohl für das Unternehmen als auch für den Bewerber von zentraler Bedeutung. Für den Bewerber hängt in der Regel der individuelle Lebensweg davon ab – für das Unternehmen gehört die Personalauswahl zu den konsequenzenreichsten sowie profitabelsten Investitionen (Görlich und Schuler, 2006). Es ist deshalb unerlässlich, Personalauswahlverfahren hinsichtlich ihrer prognostischen Möglichkeiten und Qualität kritisch zu beurteilen. Nur mit validen und bedarfsorientierten Verfahren können die bestgeeigneten Bewerber für eine Position oder Laufbahn ausgewählt werden.

Die strategische Positionierung einer Organisation ist ein weiterer wesentlicher Aspekt bei der Selektion geeigneter diagnostischer Instrumente: Stehen Veränderungen und Innovationen im Fokus der Tätigkeit, sollte der Schwerpunkt bei der Auswahl künftiger Mitarbeiter auf Merkmale wie Innovation, Lernbereitschaft und Offenheit gesetzt werden. Sollen hingegen Diversität oder die Integration von Minderheiten erklärte Ziele der Organisation sein, müssen Methoden gewählt werden, die entweder speziell auf diesen Personenkreis abgestimmt sind oder die möglichst geringe Unterschiede zwischen den Gruppen hervorbringen. Derartige Methoden, die damit verbundenen Konsequenzen sowie ihre Zusammenhänge mit berufsrelevanten Eignungsmerkmalen sind Gegenstand der vorliegenden Ausarbeitung.

Neben den spezifischen berufserfolgsrelevanten Eignungsmerkmalen einer Position müssen bei der Berufseignungsdiagnostik auch die allgemein anforderungsrelevanten Merkmale einbezogen werden – ganz unabhängig von Zugehörigkeit zu spezifischen Subgruppen. Allgemein anforderungsrelevante Merkmale bezeichnen Anforderungen, die für sämtliche Tätigkeiten von essentieller Bedeutung für den beruflichen Erfolg sind, beispielsweise kognitive Fähigkeiten (Intelligenz), Gewissenhaftigkeit, Leistungsmotivation und Integrität.

Die weitreichende strategische Bedeutung der Personalauswahl spiegelt sich darüber hinaus in weiteren Themen des Personalwesens wider:

Bei der Personalplanung, -gewinnung, -entwicklung und Leistungsbeurteilung. All diese Themen setzen die Anforderungen einer Tätigkeit mit den Kompetenzen der Inhaber in Beziehung und verwenden daher eine Anforderungsanalyse als methodische Basis, um berufserfolgsrelevante Anforderungen einer Position klar zu definieren.

Als besonders folgeschwer stellt sich die Selektion künftiger Führungskräfte dar, da von ihr nicht nur die positionsspezifische Arbeitsleistung, sondern auch die Steuerung, Motivation sowie das Arbeitsklima und damit verbunden die Arbeitsleistung einer gesamten Abteilung abhängt. Viele Studien (z. B. Cascio, 1997) belegen den monetären Nutzen von Personalentscheidungen, die sich in Form von Leistungs- und Qualitätssteigerung, verminderter Fluktuation und Kostensenkung äußern. Wesentlich deutlicher waren die finanziellen Auswirkungen eines Auswahlverfahrens bei der Auswahl von Führungskräften.

Aufgrund der hohen strategischen und operativen Bedeutsamkeit ist es daher ratsam, die Unternehmensleitung von Anfang an in die Erstellung von Anforderungsprofilen und in die Auswahl geeigneter diagnostischer Methoden mit einzubeziehen.

Praxistipp

Definieren Sie vorab gemeinsam mit der Unternehmensleitung die strategischen Ziele Ihrer Organisation (z. B. Innovation oder Diversität). Anschließend sollten die Anforderungen an die zu besetzenden Stellen mit dafür geeigneten Instrumenten der qualitativen und quantitativen Anforderungsanalyse ermittelt werden.

Die nachfolgenden Abschnitte sollen ein grundlegendes Verständnis für Methoden und Verfahren der Personalauswahl sowie deren Qualitätsanforderungen vermitteln.

2.2 Qualitätsanforderungen von Personalauswahlverfahren

Das Grundprinzip personaldiagnostischer Entscheidungen besteht in dem Abgleich individueller (Leistungs-)Voraussetzungen auf Seiten der Bewerber mit den (Leistungs-)Anforderungen der jeweiligen Position. Es gilt daher, einerseits die **Anforderungen** an eine Position/Laufbahn klar und umfassend zu erheben und andererseits, die **Ausprägung** dieser Merkmale auf der Personenseite zu ermitteln. Auf Personenseite gelangen personaldiagnostische Instrumente zum Einsatz, mit denen die individuelle Aus-

prägung gemessen wird. An jede Messung, auch in der Personalauswahl, sind Qualitätsansprüche zu stellen, anhand derer beurteilt werden kann, ob die Messung als solche korrekt ist und ob es damit zulässig ist, auf dieser Basis weitreichende Entscheidungen zu treffen.

Die Qualität eines diagnostischen Instruments wird maßgeblich durch die wissenschaftlichen Gütekriterien moderiert. Die so genannten Hauptgütekriterien eines Verfahrens sind **Objektivität**, **Reliabilität** und **Validität**. Daneben werden die Nebengütekriterien Nützlichkeit, **Fairness**, Transparenz, Ökonomie und Normierung unterschieden.

■ Objektivität

Die Objektivität eines Testverfahrens bezeichnet den Grad, in dem die Ergebnisse eines Tests unabhängig vom Untersucher sind. Je nach Phase der Testdurchführung unterscheidet man zwischen Durchführung-, Auswertungs- und Interpretationsobjektivität.

Unter **Durchführungsobjektivität** versteht man den Grad der Unabhängigkeit der Testergebnisse vom Verhalten des Untersuchungsleiters während der Testdurchführung. Eine Maximierung der Durchführungsobjektivität kann durch eine genaue und schriftliche Festlegung der Testinstruktion erreicht werden sowie durch eine weitgehend standardisierte Untersuchungssituation. Bei einem computergestützten Testverfahren kann man demnach von einer sehr hohen Durchführungsobjektivität ausgehen. Bei einem Verfahren mit Beteiligung und Einwirkung eines Testleiters oder eines Rollenspielers, ist die Durchführungsobjektivität deutlich geringer.

Die **Auswertungsobjektivität** bezeichnet das Ausmaß, in dem die Auswertung der gegebenen Antworten respektive des gezeigten Verhaltens unabhängig von der Person des Auswertenden ist. Entsprechend ist sie am höchsten, wenn kein Mensch daran beteiligt ist, der Lese-, Schreib-, Rechen- oder Übertragungsfehler verursachen könnte. Verfahren, die automatisch ausgewertet werden, haben demnach die höchste Auswertungsobjektivität. Auch bei Auswertungen per Scanner (in der Personaldiagnostik heute kaum noch gebräuchlich), über den Antwortbögen in einen Computer eingelesen werden, ist von einer hohen Auswertungsobjektivität auszugehen. Die Auswertung mittels Schablone oder eine Übertragung von Antworten in einen Computer per Hand bergen hingegen die Gefahr unsystematischer Fehler. Hier ist von einer etwas geringeren Auswertungsobjektivität auszugehen. Die geringste Auswertungsobjektivität weisen jedoch alle interaktiven Verfahren auf – lediglich Fallstudien oder (teil)standardisierte Interviews mit einer numerischen Antwortverankerung sowie interaktive Verfahren mit verhaltensverankerten Skalen, wie sie bei manchen Assessment-Centern zum Einsatz kommen, ermöglichen eine vergleichsweise gute objektive Auswertung. Hierbei verwenden

die Interviewer bzw. Assessoren einen Bewertungsleitfaden, in dem eindeutig festgelegt ist, welche Entscheidung, welche Antwort bzw. welches Verhalten wie bewertet wird.

Die **Interpretationsobjektivität** beschreibt den Grad der Unabhängigkeit der Ergebnisinterpretation vom Anwender. Sie ist dann gegeben, wenn aus gleichen Ergebnissen verschiedener Probanden gleiche Schlussfolgerungen, wie z. B. Berufserfolgsprognosen gezogen werden. Zur leichten und eindeutigen Interpretierbarkeit der Ergebnisse sollte ein Verfahren nach der Auswertung idealerweise einen Gesamtergebniswert in normierten Standardwerten (z. B. auf einer Z-Skala von 70–130 bei einem Mittelwert von 100 und einer Standardabweichung von 10) aufweisen. So wird sichergestellt, dass bei der Interpretation des individuellen Ergebnisses auch die Ergebnisverteilung in der gesamten Stichprobe Berücksichtigung findet. Sind große Bewerberzahlen zu diagnostizieren, kann im Idealfall zudem ein Cut-Off-Wert festgelegt werden, bei dessen Nichterreichen Bewerber automatisch zur Ablehnung vorgesehen werden. Für Spekulationen über die Bedeutung des Ergebnisses besteht bei einer solchen Festlegung kein Spielraum und die Entscheidungsfindung genügt strengen Anforderungen der HR-Compliance, d. h. der zu befolgenden Regeln bzw. Leitlinien eines Unternehmens.

■ Reliabilität

Die Reliabilität oder Zuverlässigkeit eines Testverfahrens betrifft das Ausmaß der Genauigkeit, mit dem ein bestimmtes Merkmal gemessen wird. Reliabilität umschreibt in diesem Sinne die Güte eines Tests hinsichtlich Präzision bzw. Messgenauigkeit. Diese kann auf verschiedene Arten geprüft werden. Zwei gebräuchliche Maße sind die Interne Konsistenz (Cronbachs Alpha) und die Retest-Reliabilität. Die **Interne Konsistenz (Cronbachs Alpha)** gibt an, wie homogen die Einzelskalen eines Verfahrens sind, d. h. inwiefern alle Items einer Skala tatsächlich das gleiche Konstrukt erfassen. Eine alternative Vorgehensweise zur Bestimmung der Internen Konsistenz ist die **Split-Half-Methode**. Dabei wird die Skala in zwei Hälften geteilt, und diese werden miteinander korreliert. Diese Methode ist nicht so zuverlässig wie Cronbachs Alpha. Beispielsweise kann es sein, dass in der zweiten Hälfte der Skala die Probanden schon ermüdet sind, und aufgrund dessen sich die beiden Skalenhälften unterscheiden. Eine Möglichkeit, dies zu umgehen, ist die Anwendung der **Odd-even-Methode**, bei der die Skala nicht in der Mitte geteilt wird, sondern jeweils die ungeraden und die geraden Items einen Skalenteil bilden. Diese ist jedoch gegenüber Cronbachs Alpha in den Hintergrund getreten.

Das andere, aussagekräftige Maß für die Reliabilität ist die **Retest-Reliabilität**. Hierbei wird das gleiche Verfahren von identischen Probanden

nach einem gewissen Zeitraum nochmals bearbeitet. Sollte das Verfahren das erfasste Merkmal exakt messen, müssten die Ergebnisse der beiden Testzeitpunkte hoch miteinander korrelieren. Lerneffekte durch die erste Anwendung werden entweder statistisch kontrolliert oder durch einen größeren zeitlichen Abstand der beiden Testdurchführungen zueinander vermieden. Die Prüfung der Äquivalenz einer Messung gelingt mit der **Paralleltest-Methode**. Hierbei wird nicht das gleiche Verfahren von den Kandidaten noch einmal bearbeitet, sondern ein äquivalentes, d. h. paralleles Verfahren. Wichtig ist bei dieser Art der Reliabilitätsbestimmung, dass Stichprobe und dazwischen liegender Zeitraum groß genug sind und dass das zu messende Merkmal selbst eine hohe Stabilität aufweist.

■ Validität

Die Validität eines Verfahrens bezeichnet den Grad der Genauigkeit, mit dem ein Verfahren das Leistungs- oder Persönlichkeitsmerkmal, das es zu messen beabsichtigt, auch tatsächlich misst. Etwas anders ausgedrückt bedeutet Validität die Angemessenheit oder Sinnhaftigkeit der Schlüsse, die aus diagnostischen Werten gezogen werden können. Validität hat insofern in der Eignungsdiagnostik eine herausragende Bedeutung, da sie Aufschluss über die Ausprägung des interessierenden Merkmals gibt und erlaubt, Vorhersagen über künftiges Verhalten und künftigen (Berufs-) Erfolg zu treffen.

Eine Möglichkeit zur quantitativen Bestimmung dieses Gütekriteriums ist die **Konstruktvalidität**. Hierbei wird untersucht, inwiefern das Testverfahren einen Zusammenhang (Korrelation) mit einem etablierten Testverfahren zum gleichen Konstrukt aufweist. Eine weitere Möglichkeit zur Feststellung der Konstruktvalidität ist die Faktorenanalyse. Hierbei handelt es sich um eine multivariate statistische Methode, mit der eine große Menge empirisch ermittelter Datenpunkte (z. B. Items eines Testverfahrens oder einer Befragung) auf dahinterliegende Ähnlichkeitsmerkmale (latente Variablen = Faktoren) reduziert werden kann. Die Faktorenanalyse kann herangezogen werden, um die Konstruktvalidität eines Verfahrens zu belegen, indem sie zeigt, dass sich die Items tatsächlich zu den jeweils hypothesen- oder konstruktkonformen Dimensionen zusammenschließen.

Zum begründeten Einsatz eines Tests in der Eignungsdiagnostik ist jedoch ein Nachweis seiner **Kriteriumsvalidität** unerlässlich. Diese beantwortet die Frage, ob ein Instrument in der Lage ist, Berufserfolg vorherzusagen. Geprüft wird dies durch die Korrelation der Testwerte mit einem Berufserfolgskriterium. Je nach Berufsbild und Beschäftigungsform gelten verschiedene Kriterien als Indikator für beruflichen Erfolg. Am aussagekräftigsten sind die erreichte hierarchische Position, Beförderungen und

Vorgesetztenurteile und dort, wo keine absolute Definition durch Tarife vorliegt, auch die Höhe der relativen Vergütung und individuelle Bonuszahlungen. Als noch befriedigende Kriterien gelten Kollegeneinschätzungen und Ausbildungs-/Studiumsnoten. Absolutes Gehalt, eigene Berufserfolgseinschätzung, Tätigkeitszufriedenheit und die Übereinstimmung von Fremd- und Selbstbild sind weniger geeignet, da diese durch viele andere Aspekte stärker beeinflusst werden bzw. mit Berufserfolg im engeren Sinne wenig zusammenhängen. So ist z. B. das Gehalt von Branche, Firmengröße etc. abhängig. Optimal sind multiple Kriterien, die sich durch Aggregation mehrerer Einzelkriterien ergeben sowie die Auspartialisierung irrelevanter Einflüsse, wie z. B. des Alters beim Gehalt (letzteres ist wichtig, da z. B. mit steigendem Alter aufgrund von Effekten der Regelbeförderung, des Senioritätsprinzips oder/und aufgrund von Tarifvereinbarungen das Gehalt ebenfalls steigt, ohne dass hieraus Rückschlüsse auf die berufliche Leistung zu ziehen wären).

Praxistipp

Prüfen Sie die Hauptgütekriterien eines Auswahlverfahrens vor dem ersten Einsatz. Testanbieter müssen in der Lage sein, entsprechende Kennwerte zu berichten. Anbieter spezifischer Diagnoseverfahren, wie z. B. von Elementen eines Assessment-Centers, bieten eine Vor-erprobung und – sofern die Entwicklungsbedingungen (Stichprobe, Kriterien, Budgetlage, Zeit) das erlauben – entsprechende Validierungsstrategien an. Ein qualitativ hochwertiges Verfahren sollte zudem geeignete Normen für Ihre Zielgruppe liefern können. Nur ein valides und an einer geeigneten, vergleichbaren Stichprobe geeichtes (normiertes) Verfahren ermöglicht zuverlässige Berufserfolgsprognosen.

■ Fairness

Die Fairness gilt als Nebengütekriterium eines Verfahrens. In der psychologischen Diagnostik wird ein Verfahren dann als fair bezeichnet, wenn es Kandidaten nicht aufgrund ethnischer, soziokultureller oder z. B. geschlechtsspezifischer Gruppenzugehörigkeit systematisch benachteiligt. In diesem Fall müsste sowohl im Prädiktor als auch im Kriterium der Anteil erfolgreicher Kandidaten identisch sein. Sprich: Sowohl in dem Instrument, das den Berufserfolg vorhersagen soll (z. B. Einstellungsinterview) als auch im Beruf selber. Eine Grundvoraussetzung hierfür ist der Einsatz prognostisch valider Verfahren. Aufgrund seiner hohen Relevanz bei der Auswahl von Bewerberinnen und Bewerbern mit Migrationshintergrund wird dieses Nebengütekriterium im Folgenden ausführlicher behandelt.

2.3 Statistische Fairness von Personalauswahlentscheidungen

Statistische Fairness ist insbesondere in den USA ein zentrales Thema der Personalauswahl. Darunter wird die Fairness im Hinblick auf die Auswahlentscheidung verstanden, bei der es keine Benachteiligung von Subgruppen – in der Regel ethnischer Minderheiten – geben sollte. Vielmehr müsse sichergestellt sein, dass der Prozentsatz erfolgreicher Kandidaten im Testverfahren derselbe sein sollte wie der Prozentsatz erfolgreicher Kandidaten im Beruf. In Intelligenztests erreichen dunkelhäutige Personen häufig schlechtere Testwerte als Weiße – die Differenz zeigt sich jedoch nicht in dieser Ausprägung bei der Beurteilung ihrer Arbeitsleistung (Wigdor und Sackett, 1993; Outtz, 2002). Diese Benachteiligung von Minderheiten – die auch bei Bewerbern süd- und mittelamerikanischer Abstammung (Hispanics) in verminderter Ausprägung gefunden wurde, wird »**Adverse Impact**« genannt. Eine derartige Benachteiligung von Minderheiten existiert jedoch weltweit. Besonders gravierend wird diese Benachteiligung, wenn die Auswahlentscheidung ausschließlich aufgrund kognitiver Maße erfolgt (z. B. Intelligenztests, Schulnoten), wenn gleich eine erfolgreiche Ausübung der Tätigkeit weitere Qualifikationen erfordert.

Kapitel 4 stellt einzelne Verfahrenstypen vor und bewertet diese im Hinblick auf ihre diskriminative Wirkung auf Basis empirischer Studien (► Kap. 4).

Adverse Impact kann bei validen Verfahren, die jedoch Subgruppenunterschiede erzeugen, vermieden werden, indem Normgruppen gebildet werden. Hierbei werden die Subgruppen einzeln normiert, so dass aus jeder Gruppe über definierte Quoten die besten Bewerber gewählt werden können. Diese Vorgehensweise ist in vielen Unternehmen und Verwaltungen gängige Praxis, beispielsweise wenn es eine Frauenquote zu erfüllen gibt oder wenn bei der Besetzung von Ausbildungsplätzen sowohl Real- als auch Hauptschüler berücksichtigt werden sollen. Mit diesem Vorgehen lassen sich Diversitätsziele verfolgen, ohne auf den Einsatz eines validen Auswahlverfahrens verzichten zu müssen. Der Nachteil dieses Vorgehens liegt jedoch auf der Hand: Da sich die Testrohwerte der Subgruppen voneinander unterscheiden, erfolgt eine inverse Diskriminierung. Gruppenmitglieder der Majorität werden trotz besserer Testwerte (und bei Vorliegen eines validen Auswahlverfahrens mithin besserer Berufserfolgsprognose) nicht eingestellt. Aus diesem Grund wurde die Subgruppenbildung in den USA 1991 zivilrechtlich verboten. Eine Quotierungslösung bewirkt stets eine schlechtere Auswahlentscheidung als eine rein am Erfolgsprädiktor orientierte Entscheidung, da nicht die geeignetsten Bewerber unabhän-

gig von ihrer Gruppenzugehörigkeit (z. B. Frauen, Männer, Realschüler, Deutsche, Migranten) ausgewählt werden.

Das so genannte **Banding**, bei dem Bänder gebildet werden, umgeht die Subgruppenbildung mit ihren Nachteilen und vermeidet gleichzeitig Adverse Impact. Hierbei werden über definierte Punktwerte Bänder über alle Rohwerte gebildet. Den unterschiedlichen Rohwerten innerhalb eines Bands wird derselbe Leistungswert zugewiesen. Die Bänder werden statistisch mittels der Konfidenzintervalle errechnet, mit denen überprüft werden kann, ob sich zwei Testwerte signifikant voneinander unterscheiden. Entsprechend geht man davon aus, dass sich die Testwerte innerhalb eines Bandes, d. h. zwischen zwei vorab definierten Punktwerten, statistisch nicht signifikant voneinander unterscheiden und folglich alle Personen innerhalb eines solchen Bandes gleich leistungsfähig sind.

Um die Konfidenzintervalle zu berechnen, wird der Standardmessfehler benötigt. Dieser kann über die Standardabweichung und die Retest-Reliabilität eines Verfahrens bestimmt werden. Ausgehend von dem höchsten Wert wird die Grenze des ersten Bandes errechnet, die zweite Grenze wird mit derselben Punktedifferenz darunter festgelegt etc.

Sollten bei der Personalauswahl Selektionsquoten angewandt werden, müssten für die zu berücksichtigenden Subgruppen eine entsprechende Anzahl an Personen aus einem Band ausgewählt werden.

Die Annahme derselben Leistungsfähigkeit innerhalb eines Bandes wirft jedoch zwangsläufig die Frage auf, warum die Personen am oberen und unteren Ende eines Bandes gleich leistungsfähig sein sollen – während sich Personen im mittleren Bereich eines Bandes von Personen im mittleren Bereich des nächsten Bandes unterscheiden. Zudem ist es statistisch inkorrekt, von einer Gleichheit der Testwerte zu sprechen, nur weil kein statistisch signifikanter Unterschied vorliegt – ähnlich als würde im Sport ein Unterschied von einigen hundertstel Sekunden Laufleistung nicht als Leistungsdifferenz sondern als Leistungsgleichheit gewertet werden. Auch das Konfidenzintervall an sich ist für die Bestimmung der Bänder nur bedingt geeignet, da es eigentlich statistisch signifikante Unterschiede zweier Messwerte untersucht. Auf mehrere Probanden übertragen weisen sie nur eine geringe Teststärke auf.

Zuletzt wird durch diese Vorgehensweise eine Einbuße an Aussagekraft des Verfahrens in Kauf genommen, indem intervallskalierte, differenzierte Werte auf eine Skala transformiert werden, die nur grob unterscheidet. Hierdurch wird die Messgenauigkeit von reliablen, messgenauen Instrumenten künstlich reduziert – genauso als würde im Sport die Zeitmessung nicht mehr in 100/sec, sondern nur noch in Intervallen gemessen, die weit unterhalb der Messgenauigkeit liegen – dann kommen zwei sehr unterschiedlich leistungsstarke Läufer z. B. im gleichen Zeitintervall von 4 sec

ins Ziel und die individuellen Leistungsunterschiede werden durch Banding des Zeitintervalls negiert. Gleiches gilt auch, wenn Standardwerte in kategoriale Ergebnisgruppen wie z. B. Schulnoten transformiert werden.

Die am häufigsten verwendete Prozedur um eine etwaige Benachteiligung von Minderheiten zu identifizieren, stellt die so genannte 4/5-Regel bzw. 80%-Regel dar. Diese prüft, ob sich die Einstellungsquote einer Bewerberminderheit (z. B. Bewerber mit Migrationshintergrund) von der Einstellungsquote der Bevölkerungsmehrheit praktisch bedeutsam unterscheidet. Nach der 4/5-Regel ist das per Definition dann der Fall, wenn gilt

$$\frac{\text{Einstellungsquote Minderheit}}{\text{Einstellungsquote Mehrheit}} < 80\%$$

Die 4/5-Regel wurde ausdrücklich für die Unternehmenspraxis entwickelt und ist generell in allen Auswahlprozessen anwendbar, was für alternativ einsetzbare statistische Signifikanztests nur bedingt gilt, da diese bei sehr kleinen oder großen Stichproben teilweise geringe Aussagekraft besitzen. Obwohl die Anwendung der 4/5-Regel grundsätzlich unabhängig von der Größe der beobachteten Fallzahlen bestimmbar ist, zeigt sich das Vorgehen bei kleinen Fallzahlen anfällig für Zufallseffekte. Wenn beispielsweise 15 von 80 Personen aus der Gruppe der Bevölkerungsmehrheit ausgewählt werden, liegt per Definition Adverse Impact vor, wenn gleichzeitig lediglich eine Person aus der Gruppe der Minderheiten eingestellt wird. Allerdings hebt bereits eine (!) weitere eingestellte Person aus der Minderheitengruppe die Indikation für Benachteiligung auf, was die geringe Stabilität der Ergebnisse in diesen Fällen verdeutlicht.

Bei besonders attraktiven Stellen, bei denen Personalabteilungen einen Großteil der Bewerber ablehnen, signalisiert die Anwendung der 4/5-Regel ebenfalls vergleichsweise (vor-)schnell den Verdacht auf Benachteiligung von Teilpopulationen. Wenn z. B. generell 90% aller Bewerber für ein Nachwuchsführungskräfteprogramm abgelehnt werden, dann darf die Einstellungsquote der potenziell benachteiligten Gruppe (z. B. Frauen) lediglich um 2% unter jener der empirisch zu bestimmenden Mehrheit (z. B. Männer) liegen, sonst zeigt die 4/5-Regel den Verdacht auf praktische Benachteiligung an.

Unter anderem aus diesen Gründen ist es in der Unternehmenspraxis von inhaltlicher und mittlerweile in den USA auch von rechtlicher Bedeutung, statistische Signifikanztests bzw. Effektstärkemaße (bei großen Stichproben) als ergänzende Informationen heranzuziehen, um die geschilderten und einige weitere Limitationen der 4/5-Regel zu kompensieren. Sonst laufen Praktiker eventuell Gefahr, vorliegende Personalauswahlprozesse

Analyse protected group: **Alter**

	Gesamt	davon angestellt	SelectionRate	Adverse Impact Ratio
<20	7417	435	5,9%	reference
<30	2394	62	2,6%	-56,0%
<40	62	0	0,0%	-100,0%
<50	13	0	0,0%	-100,0%
<60	3	0	0,0%	-100,0%
<70	0	0	0,0%	-100,0%

Geschlecht

	Gesamt	davon angestellt	SelectionRate	Adverse Impact Ratio
Female	3225	110	3,4%	-41,0%
Male	6667	387	5,8%	reference

■ **Abb. 2.1** Adverse Impact Indicator (System Jobmatcher von HR Diagnostics): Die Einhaltung der 4/5 Regel wird dynamisch im laufenden Prozess überwacht und kann für verschiedene Diskriminierungsmerkmale wie z. B. Alter oder Geschlecht dargestellt werden. Personalverantwortliche können so rasch auf Fehlentwicklungen reagieren und juristische Risiken vermeiden. Dieses Instrument erlaubt es, ohne statistische Expertise die Auswahl fairness für die betreuten Bewerbergruppen sicherzustellen.

vorschnell als diskriminierend einzuschätzen. Eine fallbezogene Auswahl angemessener Prüfstatistiken und die sorgfältige Interpretation der gewonnenen Daten verlangt jedoch ein fortgeschrittenes testmethodisches Wissen bzw. eine technische Lösung innerhalb des Bewerbermanagements (■ Abb. 2.1).

2.4 DIN 33430

Bei der **DIN 33430**, die 2002 in Deutschland verabschiedet wurde, handelt es sich um eine Prozessnorm, die eingeführt wurde, um Qualitätsstandards der berufsbezogenen Eignungsdiagnostik zu erhöhen. Durch die Nutzung wissenschaftlicher Erkenntnisse bei der Personalauswahl wurden Leitsätze festgelegt, mit denen Organisationen die Qualität ihrer Prozesse sowie die Seriosität von Anbietern, Prozessen und Methoden bewerten können. Konkret beinhaltet die Norm Qualitätskriterien für die Verfahren, Richtlinien für den Prozessablauf sowie Qualifikationsanforderungen an die am Prozess beteiligten Personen.

Da es sich bei der DIN 33430 um keine Produkt-, sondern eine Prozessnorm handelt, ist diese nicht zur isolierten Bewertung der Qualität eines Verfahrens außerhalb des Rahmens seiner spezifischen Anwendung geeignet.

2.4.1 Ablauf einer berufsbezogenen Eignungsbeurteilung

Ein Personalauswahlprozess nach der DIN 33430 sollte folgende Schritte umfassen:

1. **Auswahl und Zusammenstellung der diagnostischen Instrumente:** Die eingesetzten Instrumente sollten einen Anforderungsbezug haben – sprich auf Grundlage einer Arbeits- und Anforderungsanalyse ausgewählt worden sein. Sie sollten klare Hinweise zur Handhabung enthalten, Angaben zu den Gütekriterien sowie geeignete Normen beinhalten und theoretische Grundlagen sowie die detaillierte Verfahrenskonzeption dokumentieren.
2. **Planung der Eignungsbeurteilung:** Im Vorfeld müssen alle Mitwirkenden spezifisch geschult, Beurteilungs- und Interpretationsregeln festgelegt und eine Anforderungsanalyse durchgeführt werden. Jedes Verfahren, das zum Einsatz kommt, muss mindestens einer zu operationalisierenden Anforderung zugeordnet werden können. Weiterhin wird festgelegt, wie auf vorhersehbare Nachfragen der Kandidaten reagiert wird und wie den Kandidaten ihre Ergebnisse vermittelt werden.
3. **Durchführung:** Die Objektivität sollte bei der Durchführung gegeben sein. Sämtliche Vorgaben und Instruktionen müssen durch die Anwender beachtet werden. Alle Kandidaten sollten in gleichem Maße Zugang zu eventuellen Vorinformationen haben. Anweisungen an die Kandidaten müssen nach Möglichkeit standardisiert erfolgen und verständlich sein.
4. **Auswertung, Interpretation und Urteilsbildung:** Die Auswertung und Interpretation muss nach den vorab definierten Regeln erfolgen. Die Urteilsbildung muss belegbar sein.
5. **Dokumentation:** Die Ergebnisse der Anforderungsanalyse, die Auswahl und Zusammenstellung der diagnostischen Verfahren, sämtliche Regeln zu Aspekten der Durchführung und der Auswertung sowie das Vorgehen bei der Urteilsbildung müssen nachvollziehbar dokumentiert sein.

2.4.2 Qualitätsanforderungen an Auftragnehmer und Durchführende

Entsprechend der DIN-Norm sollten bei allen Entwicklern und Anwendern fundierte Kenntnisse von Eignungsbeurteilungen und praktische Erfahrungen in der Konzeption und Durchführung von Verfahren zur Eignungsbeurteilung sowie deren Evaluation vorliegen. Sie sollten weiterhin Kenntnis der verfügbaren Verfahren und Prozesse sowie deren Qualität und Einsatzvoraussetzungen besitzen und über die Konstrukte informiert sein, die den Eignungsmerkmalen zugrunde liegen. Die genannten Qualifikationen müssen nachgewiesen werden können und regelmäßige fachliche Fortbildungen sollten besucht werden.

Speziell bei Eignungsinterviews und Verhaltensbeobachtungen sollten weiterhin Kenntnisse über Rahmenbedingungen von Verfahren zur mündlichen Informationsgewinnung und über einschlägige Evaluationen vorliegen.

Praxistipp

Erkundigen Sie sich bei Ihrem Testanbieter, ob der gemeinsam definierte Prozess DIN-konform ist. Verlassen Sie sich jedoch nicht auf eine pauschale Aussage, sondern gehen Sie die einzelnen Punkte gemeinsam mit dem Anbieter durch und prüfen Sie sie kritisch. Lassen Sie sich nicht durch Aussagen, dass ein bestimmtes Verfahren der DIN-Norm entspreche, täuschen. Die Norm bezieht sich nicht auf einzelne Verfahren, sondern den gesamten Prozess berufsbezogener Eignungsuntersuchungen.

2.5 Sprach- und Kulturgebundenheit diagnostischer Methoden

Die Mehrzahl aller diagnostischen Verfahren zeichnet sich durch einen hohen Sprachanteil aus – sei es in Instruktionen, konkreten Aufgabenstellungen oder dargebotenen Antwortalternativen. Diese Sprachlastigkeit von Verfahren benachteiligt Personen mit eingeschränktem Sprachverständnis, unter anderem diejenigen Bewerber, deren Muttersprache eine andere ist. Selbst wenn die Sprache von nichtmuttersprachlichen Kandidaten ausreichend gut beherrscht wird, um alle Wörter zu verstehen, wirkt sich ein hoher Anteil von Sprache negativ beim Verständnis sprachlicher Feinheiten sowie auf die Bearbeitungsdauer aus, was sich insbesondere bei zeit-

begrenzten Verfahren direkt auf den erreichten Testwert auswirkt. Bei interaktiven Verfahren hingegen spielt Sprache eine noch zentralere Rolle. Ein Muttersprachler vermag beispielsweise eine Interviewfrage deutlich schneller und exakter zu erfassen sowie sprachgewandter zu beantworten oder sich in einer Gruppendiskussion oder einem Rollenspiel rascher und sprachlich souveräner einzubringen.

Daneben zeigt eine Vielzahl empirischer Untersuchungen die Kulturgebundenheit vieler diagnostischer Verfahren auf. Neben tatsächlich bestehenden kulturellen Unterschieden in der Ausprägung von Persönlichkeitsmerkmalen eignen sich auch unterschiedliche Methoden mehr oder weniger gut in den einzelnen Kulturbereichen. So eignen sich beispielsweise Gruppendiskussionen weniger für den Einsatz im ostasiatischen Raum, wo aufgrund der eher kollektivistischen Gesellschaft ein Selbstkonzept vorherrscht, das auf sozialer Integration und Eingebundenheit beruht. Während in den westlichen Kulturen eher individuell agiert wird, agieren asiatische Personen eher zurückhaltend und als Teil eines Kollektivs, was die Durchführung von kompetitiven Gruppendiskussionen und Bewertung nach westlichen Standards per se erschwert. Auch bei der Bearbeitung von sozialen Dilemmata (Gefangenen-Dilemma-Spiel) zeigten sich unterschiedliche Verhaltensweisen von westlichen und asiatischen Kandidaten. Letztere wählen häufiger die kooperative Lösungsoption (Bonta, 1997). Auch die Tendenz, sich positiv darzustellen, was viele Persönlichkeitsverfahren provozieren, unterscheidet sich deutlich zwischen westlichen und asiatischen Kulturen, wobei sich letztere im Vergleich mit westlichen Kulturen durch eine größere Bescheidenheit und Zurückhaltung auszeichnen.

In Kapitel 4 wird ausführlich auf die gängigsten diagnostischen Methoden, empirische Daten zu Gruppenunterschieden und darauf basierend ihre Eignung für die Auswahl von Migranten eingegangen (► Kap. 4).

2.6 Integrationsförderliche Personalauswahlprozesse

Empirische Studien belegen die Existenz von Subgruppenunterschieden bei Geschlecht und ethnischer Zugehörigkeit bei den meisten Persönlichkeitsmerkmalen. Manche Unternehmen entscheiden daher, Migranten bei gleichen oder ähnlichen Testwerten zu bevorzugen (Risavy und Hausdorf, 2011). Dies wirft jedoch berechtigterweise die Frage der reversierten Benachteiligung auf. Ein solches Vorgehen – ebenso wie eine Personalentscheidung aufgrund einer Quotenregelung – macht daher die Ungleichbehandlung von Bewerbern dauerhaft evident, die sie eigentlich eliminieren möchte (Blits und Gottfredson, 1990). Denn ein valides Personalauswahl-

verfahren wird nicht konstruiert, um Gruppenunterschiede zu erfassen oder zu produzieren, sondern um die Leistungsstärke der Organisation zu maximieren, indem berufserfolgsrelevante Anforderungsmerkmale der vakanten Position(en) reliabel operationalisiert und Auswahlentscheidungen auf Basis individueller Übereinstimmungen mit dem Anforderungsprofil objektiv und unabhängig von nicht eignungsrelevanten Kriterien getroffen werden.

Bei einer Darbietung übersetzter oder gänzlich sprachfreier Testverfahren stellt sich daher die Frage der Vernachlässigung relevanter Anforderungsmerkmale. So ist beispielsweise das Beherrschen der deutschen Sprache bei vielen Tätigkeiten ein zentraler Aspekt für die erfolgreiche Besetzung der vakanten Stelle. Zudem lassen sich nicht alle relevanten Anforderungen sprachfrei erfassen. Man würde folglich in Kauf nehmen, auf einige berufserfolgsrelevante Anforderungen zu verzichten oder diese weniger stark zu gewichten. Methodisch würden hier Varianzanteile für die Erfolgsprognose systematisch verschenkt.

Sprachkenntnisse von Migranten können auf der anderen Seite bei einigen Positionen von Vorteil für eine erfolgreiche Besetzung der Stelle sein, beispielsweise bei internationalen Aktivitäten, multikulturellen Belegschaften oder bei Tätigkeiten, die regelmäßig die Interaktion mit Angehörigen des gleichen Sprach- und Kulturkreises erfordern, so wie bei international agierenden Unternehmen, Arztpraxen oder Behörden wie der Bundespolizei, dem Zoll oder der Polizei bis hin zum Einwohnermeldeamt.

Fallbeispiel

Eine deutsche Landespolizeibehörde strebte bei der Neuentwicklung der Auswahlverfahren und des Prozesses zur Auswahl von Polizeianwärtern einerseits Diskriminierungsfreiheit unter anderem hinsichtlich des Merkmals Geschlecht an, andererseits danach, bevorzugt Bewerber mit muttersprachlichen Kenntnissen in den für den polizeilichen Vollzugsdienst besonders relevanten Sprachen (= Bedarfssprachen) einzustellen.

Mit statistischen Verfahren konnte identifiziert werden, welche diagnostischen Verfahren besonders anfällig im Sinne geschlechtsspezifischer Diskriminierung sind, d. h. weibliche Bewerber benachteiligen. Durch Umgestaltung der Gewichtung verschiedener Verfahrenselemente zueinander (geringere Gewichtung figuraler und numerischer kognitiver Fähigkeiten, Höhergewichtung interpersonaler Eigenschaftsmerkmale, im weiteren Sinne soziale Kompetenz) konnten prädiktorseitige Benachteiligungen ausgeschlossen werden und eine Besetzung des Polizeidienstes mit Männern und Frauen im gleichen Verhältnis wie der Verteilung des Geschlechts in der Bewerbungspopulation sichergestellt werden. Die sich hieraus ergebende Minderung der Gesamtvalidität



bewegte sich in einer zu vernachlässigenden Größenordnung. Voraussetzung für diese Optimierung war das Vorliegen einer signifikanten Stichprobe an realen Bewerbern mit Datenpunkten zu allen Einstellungsverfahren – nicht alle Effekte lassen sich a priori bei der Konstruktion eines Verfahrens vermeiden. In der Mehrheit der personalpsychologischen Anwendungsfälle ist die Verfügbarkeit realer Felddaten der fraglichen Population unabdingbar, um Fragestellungen hinsichtlich Validität, Adverse Impact, Diskriminierung und Behebung von Diskriminierung sachkundig und nachhaltig beantworten zu können.

Im Hinblick auf die Einstellung von Bewerbern mit Kenntnis in den Bedarfssprachen wurde ein Bonussystem zur Berechnung der individuellen Gesamteignung und des individuellen Rankings entwickelt. Zusätzlich zu den für alle Bewerber identischen Auswahlverfahren werden auf freiwilliger Basis so genannte Bedarfssprachentests angeboten, in denen entsprechende Muttersprachler ihre Sprachkenntnisse dokumentieren können. Das Ergebnis geht in Form eines Bonus in die Gesamtbewertung ein, so dass auch schlechtere Deutschkenntnisse innerhalb definierter Grenzen (Mindestanforderungen an die Beherrschung deutscher Sprache) durch die souveräne Beherrschung einer polizeilichen Bedarfssprache kompensiert werden können. Eine Benachteiligung durch das Merkmal »Migrationshintergrund« konnte durch die bewusste Nutzung eines »Migrationsvorteils« (Beherrschung einer für den Polizeivollzugsdienst relevanten Sprache) geheilt werden.

Diversity Management in der Personalauswahl
Kulturelle Vielfalt in Unternehmen und Behörden
ermöglichen

Frintrup, A.; Flubacher, B.
2014, X, 101 S., Softcover
ISBN: 978-3-642-41433-6