GERMAN JOURNAL OF
HUMAN RESOURCE
MANAGEMENT
Zeitschrift für Personalforschung

Article

German Journal of
Human Resource Management
1–28

DOI: 10.1177/2397002216637289
gjh.sagepub.com
SAGE

Adverse Impact in der Personalauswahl einer deutschen Behörde: Eine Analyse ethnischer Subgruppendifferenzen[1]

Siegfried Stumpf
Technische Hochschule Köln, Germany

Wolf Rainer Leenen
Technische Hochschule Köln, Germany

Alexander Scheitza
Kölner Institut für Interkulturelle Kompetenz, Germany

Zusammenfassung

Während die englischsprachige Personalforschung in den letzten Jahren zunehmend differenzierte Ergebnisse zum Adverse Impact in Auswahlverfahren vorlegen konnte, fehlen solche Untersuchungen zu Subgruppenunterschieden in der Personalauswahl deutscher Organisationen. Am Beispiel einer großen deutschen Behörde wird ein mehrstufiger Personalauswahlprozess im Hinblick auf das Abschneiden von Bewerbern mit versus ohne Migrationshintergrund analysiert. Die Ergebnisse zeigen, dass die in einer frühen Auswahlphase zum Einsatz kommenden Testverfahren zur Messung der kognitiven Fähigkeiten sowie der Rechtschreibung erhebliche Subgruppendifferenzen aufweisen. Differenzen zwischen den Bewerbern mit versus ohne Migrationshintergrund sind auch trotz der Vorselektion mit den Tests in dem in einer späteren Auswahlphase durchgeführten Assessment Center nachweisbar. Dabei wird der Erfolg in allen Phasen des Auswahlverfahrens von der Art des Migrationshintergrundes beeinflusst: Während bei ausländischen Staatsbürgern die größte Diskrepanz zu den Bewerbern ohne Migrationshintergrund auftritt, liegen die Ergebnisse von Bewerbern, die von Geburt an Deutsche sind, aber einen ausländischen Vater und/oder eine ausländische Mutter haben, nahezu auf dem Niveau der Bewerber ohne Migrationshintergrund. Diese Befunde werden auf dem Hintergrund der internationalen Forschungslage zum Adverse Impact analysiert. Abschließend werden Maßnahmen zur Reduktion von Subgruppendifferenzen bei der Personalauswahl von Organisationen am Beispiel der betrachteten Behörde diskutiert.

Anschrift des Verfassers:
Siegfried Stumpf, Technische Hochschule Köln, Campus Gummersbach, Steinmüllerallee 1, 51643 Gummersbach, Germany.
Email: siegfried.stumpf@th-koeln.de

Schlagwörter
Adverse Impact, Personalauswahl, ethnische Subgruppenunterschiede, Migration

Einleitung

Der Begriff „Adverse Impact" ist im Kontext der politischen und rechtlichen Diskussion über Fragen der Chancengleichheit in den USA geprägt und über entsprechende Verordnungen des US-amerikanischen Arbeits- und Justizministeriums ab Ende der 1970er Jahre verbreitet worden (Equal Employment Opportunity Commission et al., 1978). Von „Adverse Impact" wird gesprochen, wenn Personalentscheidungen im Rahmen von Einstellungen oder Beförderungen erheblich unterschiedliche Auswahlquoten zum Ergebnis haben, die zu Lasten der Mitglieder einer bestimmten Rasse, einer ethnischen Gruppe oder eines Geschlechts gehen. Als Indikator für „erheblich" („substantial") wird in der US-amerikanischen Rechtsprechung bei der Personalauswahl die 80%-Regel verwendet, nach der Adverse Impact dann vorliegt, wenn die Erfolgschancen der Mitglieder einer Minorität weniger als 4/5 der Erfolgschancen der Mitglieder der Majorität betragen (Zedeck, 2010). Aus der Feststellung eines Adverse Impact kann nicht zwangsläufig auf eine Bevorzugung oder Diskriminierung von Personen geschlossen werden. Es kann sich auch um das Phänomen einer nicht intendierten Schlechterstellung einer Gruppe durch ein standardisiertes, auf alle gleich angewendetes Verfahren handeln. Tritt Adverse Impact auf, so ist über die Ursachen des Effekts per se noch nichts bekannt: Er kann durch tatsächliche Unterschiede bei den Vergleichsgruppen hinsichtlich job-relevanter Merkmale, wie z.B. Unterschiede in der Körpergröße oder Körperkraft zwischen Männern und Frauen, bewirkt werden; er kann aber auch auf verzerrende Vorgehensweisen und Standards im Prozess der Personalauswahl zurückgehen.

Die US-amerikanische Arbeitsrechtsprechung nimmt bereits die Feststellung eines Adverse Impact zum Anlass, vom Arbeitgeber eine Rechtfertigung zu verlangen: Die Auswahlinstrumente und -praktiken, die den Effekt auslösen, müssen im Hinblick auf die Arbeitsanforderungen als schlüssig darstellbar sein und im Einklang mit Betriebsnotwendigkeiten stehen. Adverse Impact löst in der US-amerikanischen Rechtsprechung also eine Beweislastumkehr aus, die für betroffene Organisationen mit erheblichen Kosten und Imageeinbußen verbunden sein kann. Eine solche Beweislastumkehr sieht auch das 2006 in Deutschland eingeführte Allgemeine Gleichbehandlungsgesetz (AGG) vor, aus dem Schadensersatz- bzw. Entschädigungsansprüche bei nachgewiesener direkter oder indirekter Benachteiligung ableitbar sind.

Unabhängig von möglichen juristischen Konsequenzen erhalten Analysen zum Adverse Impact für die öffentlichen Arbeitgeber erhebliche Bedeutung, wenn diese beginnen, Diversityziele in ihrer Personalpolitik zu verfolgen (Leenen et al., 2014a). Nach einer Zeit des integrationspolitischen Stillstands in den 1980er und 1990er Jahren hat der Öffentliche Dienst in Deutschland die mangelhafte Repräsentation von Personen mit Migrationshintergrund als Problem erkannt und wirbt inzwischen aktiv um diese „anderen Deutschen", die mit ihrer Mehrsprachigkeit und ihren Kenntnissen anderer kultureller Milieus wichtige Kompetenzen in den bisher eher kulturhomogenen Verwaltungsbereich einbringen könnten. Vor allem im „staatsnahen"

Beschäftigungsbereich, im Bereich der Sicherheitskräfte und der staatlichen Leistungs- und Eingriffsverwaltung, ist diese Unterrepräsentation eklatant (Daten dazu bei Leenen et al. 2014a: 17–19). Die politische Brisanz dieses Tatbestands wird deutlich, wenn man andererseits steigende Anteile dieser Gruppe an der Bevölkerung insgesamt, insbesondere in bestimmten Bevölkerungssegmenten, in Betracht zieht. Im Jahr 2014 hatten insgesamt 16,4 Millionen Menschen in Deutschland – das sind immerhin 20,3% der Gesamtbevölkerung – einen Migrationshintergrund. Rund 2,8 Mio. von ihnen haben ihre Wurzeln in der Türkei, 2,9 Mio. in den Nachfolgestaaten der ehemaligen Sowjetunion, und jeweils 1,5 Mio. in den Nachfolgestaaten des ehemaligen Jugoslawiens und in Polen (Statistisches Bundesamt, 2013) Der Durchschnittswert von 20,3% wird im städtischen Umfeld und bei der jüngeren Bevölkerung allerdings deutlich überschritten. Nordrhein-Westfalen (NRW) hat daher sogar gesetzlich eine Erhöhung des Anteils der Menschen mit Migrationshintergrund im öffentlichen Dienst als Ziel fixiert (§ 6 „Gesetz zur Förderung der gesellschaftlichen Teilhabe … vom 14. Februar 2012). Dies soll allerdings dezidiert „ohne verpflichtende Quote" (Ministerium für Arbeit, Integration und Soziales des Landes Nordrhein-Westfalen, 2013: 6) erreicht werden. Zwangsläufig muss es daher politisch von besonderem Interesse sein, Zugangshindernisse bei der Auswahl und der Einstellung von Menschen mit Migrationshintergrund identifizieren und gegebenenfalls auch abbauen zu können.

Im Folgenden wird am Beispiel des mehrstufigen Personalauswahlprozesses einer deutschen Großbehörde, der Polizei NRW, das Abschneiden von Bewerbern[2] mit vs. ohne Migrationshintergrund analysiert. Der Anteil von Angehörigen ethnischer Minderheiten in deutschen Polizeibehörden wird für die Zeit bis zur Jahrtausendwende auf unter 1% geschätzt, während z.B. die niederländische Polizei diesen Anteil bereits für das Jahr 1999 mit 4,8% angibt. Exakte Zahlen kann man auch für die Zeit nach dem Jahr 2000 nicht benennen, weil „Zugehörigkeit zu einer ethnischen oder kulturellen Gruppierung" im Polizeidienst statistisch grundsätzlich nicht erhoben wird. Die Polizei NRW erhebt allerdings seit 2002 bei Bewerbungen und Einstellungen den Anteil der Personen mit Migrationshintergrund gemäß der Definition des Statistischen Bundesamtes. Über eine Hochrechnung der Einstellungszahlen kann man erschließen, dass der Anteil dieser Personengruppe am Personalbestand inzwischen lediglich bei etwa 3% liegen dürfte (Leenen et al. 2014a: 20). Da das politische Ziel einer weiteren Erhöhung des Anteils von Personen mit Migrationshintergrund an den im öffentlichen Dienst Beschäftigten erklärtermaßen „ohne verpflichtende Quote" und ohne den Gleichheitsgrundsatz verletzende Maßnahmen erreicht werden soll, ist es politisch von besonderem Interesse, „mögliche Hemmnisse bei der Auswahl und der Einstellung von Menschen mit Migrationshintergrund" (Ministerium für Arbeit, Integration und Soziales des Landes Nordrhein-Westfalen, 2013: 6) identifizieren zu können. Von daher sind nicht nur die Einstellungsquoten, sondern auch die „Erfolgsquoten" im Auswahlverfahren interessant. Die Feststellung eines Adverse Impact muss vor diesem Hintergrund als ein ernst zu nehmender Hinweis auf mögliche Zugangshindernisse gesehen werden.

Ergebnisse einer groß angelegten empirischen Untersuchung zum Adverse Impact in Deutschland und insbesondere zur Personalauswahl im Öffentlichen Dienst werden unseres Wissens hier erstmalig veröffentlicht. Von forschungsbasierten Hypothesen ausgehend wird untersucht, welche Subgruppenunterschiede im Auswahlprozess

nachweisbar sind, mit welchen Faktoren solche Unterschiede korrelieren und welche Auswirkungen diese Unterschiede auf die Einstellungschancen haben. Weiterhin wird diskutiert, wie die Wirkung des Adverse Impact abgemildert und ein stärker auf Diversitygerechtigkeit achtendes Verfahren konstruiert werden könnte.

Forschungsstand, Fragestellungen und Hypothesen

Das Konzept des Adverse Impact hat eine Vielzahl organisationspsychologischer Arbeiten zur Personalauswahl stimuliert. Dabei wird insbesondere betrachtet, wie Angehörige unterschiedlicher Gruppen in Auswahlprozessen abschneiden, inwieweit Erfolgsunterschiede von den eingesetzten Verfahren abhängig sind, und mit welchen Maßnahmen man Subgruppenunterschiede in der Personalauswahl reduzieren kann, ohne die Validität des Verfahrens maßgeblich zu reduzieren (siehe dazu den Überblicksartikel von Outtz, 2010). Die Frage nach ethnischen Subgruppenunterschieden in leistungsrelevanten Merkmalen ist aber, zumindest mit Bezug auf kognitive Fähigkeiten, in der Psychologie schon wesentlich älter, bereits Galton (1982) und Thorndike (1921) haben sich damit befasst (Roth et al., 2001).

Die meisten Untersuchungen zum Adverse Impact in der Personalauswahl stammen aus den USA und es dominieren folglich insbesondere Vergleiche zwischen weißen und schwarzen US-Amerikanern. In einem geringeren Ausmaß berücksichtigt die Forschung auch Minderheiten wie Hispanics oder Asian Americans (Ployhart und Holtz, 2008: 169). Untersuchungen zum Adverse Impact bei der Personalauswahl deutscher Organisationen und unter Berücksichtigung der in Deutschland lebenden Minderheiten fehlen allerdings bisher.

Mittlerweile liegen zu diesem Forschungsfeld auch zahlreiche Meta-Analysen (z.B. Roth et al., 2001) und Überblicksartikel (z.B. Ployhart und Holtz, 2008; De Soete et al., 2012; Lindsey et al., 2013) vor. Ethnische Subgruppenunterschiede werden darin üblicherweise anhand des Effektstärkenmaßes Cohen's d beschrieben (Cohen, 1988). Cohen's d dividiert die zwischen zwei Gruppen bestehende Mittelwertdifferenz durch die mit den Stichprobenumfängen gewichtete Standardabweichung der Messwerte in den Gruppen (Roth et al., 2001: 299). Damit werden Mittelwertdifferenzen in Standardabweichungseinheiten ausgedrückt. Ein d von z.B. 1,0 bezeichnet somit eine Mittelwertdifferenz, die genau der Merkmalsstreuung innerhalb der Gruppen, und damit einer Standardabweichungseinheit, entspricht. Das Vorzeichen wird dabei üblicherweise so verwendet, dass ein positives d für ein besseres Abschneiden der Majoritätsgruppe steht, ein negatives dagegen für das bessere Abschneiden der Minoritätsgruppe. Die Verwendung von Cohen's d als Indikator für Subgruppenunterschiede hat gegenüber Unterschieden in Erfolgs- oder Durchfallquoten den Vorteil, dass Cohen's d unabhängig von willkürlich festsetzbaren Selektionsquoten ist (Roth et al., 2001: 299; Sackett und Ellingson, 1997; Kehoe, 2010). Gemäß der üblichen Konvention gilt ein d ab 0,20 als kleine Effektstärke, ab 0,50 als mittlere, und ab 0,80 als große Effektstärke (Bortz und Döring, 1995: 568). Die Forschungslage stellt sich wie folgt dar:

1. *Ethnische Subgruppenunterschiede variieren mit der Auswahlmethodik*: Die höchsten Unterschiede zwischen Majoritäts- und Minoritätsgruppen wurden

bei kognitiven Fähigkeitstests („Intelligenztests“) ermittelt. In ihrer Metaanalyse stellten Roth et al. (2001) für den Vergleich von schwarzen vs. weißen US-Amerikanern in verschiedenen Berufsfeldern (Industrie, Bildung, Militär) ein durchschnittliches unkorrigiertes d von 1,10 fest. In einer der wenigen europäischen Studien analysieren De Meijer, Born, Terlouw und van der Molen (2006) Bewerbungsdaten bei der niederländischen Polizei und kommen für den eingesetzten Intelligenztest beim Vergleich von Majoritätsbewerbern vs. Migranten der ersten Generation auf ein d von 0,97. Für Tests, die den allgemeinen Intelligenzfaktor g erfassen, liegen die d-Werte oftmals höher als für (Sub-)Tests, die spezifische Intelligenzfacetten messen (vgl. z.B. Roth et al., 2001; Outtz und Newman, 2010: 65). Bei Assessment Centern sind die d-Werte für Subgruppenunterschiede lediglich in mittlerer Effekthöhe: In der Metaanalyse von Dean, Bobko und Roth (2008) ergibt sich für den Vergleich von Weißen vs. Schwarzen ein d von 0,52. Für Persönlichkeitstests ergeben sich die niedrigsten d-Werte: In ihrer Metaanalyse berichten Foldes, Duehr und Ones (2008) im Vergleich von Weißen vs. Schwarzen je nach Big-5-Trait d-Werte zwischen –0,07 (Gewissenhaftigkeit) und 0,16 (Extraversion). In der Studie von De Meijer et al. (2006) wird für den Big-5-Trait der Gewissenhaftigkeit, der mit generellem beruflichem Erfolg besonders stark korreliert, sogar ein d von –0,44, d.h. zu Gunsten der Migranten der ersten Generation, berichtet. Ferner ergeben sich laut der Übersicht in De Soete, Lievens und Druart (2012) für weitere eignungsdiagnostische Methoden folgende d-Werte: Für Tests zur sprachlichen Fertigkeit 0,40 bis 0,76, für strukturierte Interviewverfahren 0,36 bis 0,56, für biografische Daten 0,33, für Arbeitsproben 0,52 bis 0,73, und für Situational-Judgment-Tests 0,24 bis 0,38.

2. *Ethnische Subgruppenunterschiede sind abhängig von der jeweils betrachteten Minoritätsgruppe*: Hispanics schneiden im Vergleich zu weißen US-amerikanischen Bewerbern besser ab als schwarze Bewerber; für kognitive Fähigkeitstests wird hierzu in Roth et al. (2001) ein d von 0,72 im Vergleich Weiße vs. Hispanics berichtet im Gegensatz zu dem d von 1,10 im Vergleich Weiße vs. Schwarze. In der Assessment-Center-Metaanalyse von Dean, Bobko und Roth (2008) ergibt sich für den Vergleich Weiße vs. Hispanics ein d von 0,28 im Kontrast zu einem d von 0,52 im Vergleich Weiße vs. Schwarze. De Meijer et al. (2006) stellen in ihrer Studie deutliche Unterschiede im Abschneiden von Migranten der ersten und zweiten Generation fest: Im Vergleich von Majorität vs. Migranten der zweiten Generation ergeben sich d-Werte von lediglich 0,48 für den Intelligenztest und von 0,27 für das Assessment Center, während für den Kontrast von Majorität vs. Migranten der ersten Generation d-Werte von 0,97 für den Intelligenztest und 0,38 für das Assessment Center festgestellt wurden. Melchers und Annen (2010) fanden in ihrer Untersuchung der Offiziersauswahl in der Schweizer Armee nur sehr geringe d-Werte (–0,18 bis 0,09) zwischen deutschsprachiger Majorität vs. italienischsprachigen und französischsprachigen Minoritäten für den Intelligenztest und das Assessment Center. Um sprachliche

Verzerrungen zu reduzieren, konnten hier die Tests und Übungen allerdings in allen drei Muttersprachen absolviert werden.

3. *Vernachlässigte Theorieentwicklung*: Im Fokus bisheriger Forschung steht die Beschreibung des Adverse Impact in Abhängigkeit von ethnischer Subgruppe und Auswahlmethodik. Die Erklärung dieser Befunde hat bislang weniger Forschungsaufmerksamkeit erfahren und die Theorieentwicklung hierzu steckt in den Anfängen. Zur Erklärung ethnischer Subgruppeneffekte haben Outtz und Newman (2010) ein Modell entwickelt, das insbesondere Umweltvariablen (sozioökonomischer Status, familiäre Entwicklungsumgebung, gesellschaftliche Bildungschancen usw.) sowie kulturelle und identitätsbildende Faktoren (Werte, soziale Selbst- und Fremdkategorisierung usw.) berücksichtigt, und kausalorientierte Forschungen anleiten könnte. Fähigkeiten in der für das Auswahlverfahren maßgeblichen Sprache fehlen in diesem Modell, das vor allem auswahlbezogene Subgruppeneffekte beim Vergleich zwischen Schwarzen und Weißen in der US-Amerikanischen Gesellschaft erklären will. In Europa dürfte der Faktor Sprache aber besonders wichtig sein. In ihrer Untersuchung zur Personalauswahl der niederländischen Polizei fanden De Meijer et al. (2006), dass die Beherrschung der holländischen Sprache, gemessen über einen Dutch-Language-Proficiency-Test, mehr Varianz im Abschneiden beim Auswahlverfahren erklärt als Bildungsunterschiede und Ethnizität.
4. *Interventionsorientierung*: Besondere Aufmerksamkeit hat die Frage erfahren, durch welche Strategien und Maßnahmen man in der Personalauswahl Adverse Impact reduzieren kann. Als „Validitäts–Diversitäts-Dilemma" der Personalauswahl (Pyburn et al., 2008) wird das Problem gekennzeichnet, dass gerade der Einsatz besonders valider Prädiktoren für die spätere berufliche Leistung, wie z.B. von Intelligenztests (Schmidt und Hunter, 1998), erhebliche Subgruppendifferenzen erzeugt. Eine besondere Bedeutung kommt somit der Frage zu, wie man Leistungsunterschiede im Auswahlverfahren zwischen ethnischen Subgruppen reduzieren kann, ohne die Validität des Verfahrens maßgeblich zu beeinträchtigen und somit Diversitäts- und Validitätsziele zugleich realisieren kann. Als zielführend bei der Lösung des Dilemmas werden insbesondere folgende Maßnahmen eingestuft (Ployhart und Holtz, 2008; De Soete et al., 2012): Einbezug beruflich relevanter Anforderungsmerkmale, die spezifische Ressourcen ethnischer Minoritäten sein können (z.B. arbeitsplatzrelevante Fremdsprachenkenntnisse); Rückgriff auf möglichst realistische simulationsnahe Verfahren wie Assessment-Center-Übungen, Arbeitsproben oder situative Interviews anstelle oder in Ergänzung zu kognitiven Fähigkeitstests; Reduktion von für die berufliche Leistung irrelevanten Varianzanteilen im Auswahlverfahren, wozu insbesondere die Verringerung sprachlicher Anforderungen („verbal load") gehört, soweit diese für die Erfassung eines Anforderungsmerkmals und/oder die Bewährung am Arbeitsplatz irrelevant sind.

Bisher fehlen systematische Untersuchungen zu ethnischen Subgruppendifferenzen und zum Adverse Impact in der Personalauswahl deutscher Organisationen. Es gibt lediglich Untersuchungen, die zeigen, dass Hinweisreize zur sozialen Kategorisierung

(z.B. ausländisch klingende Namen) die Chancen reduzieren, überhaupt zu einem Vorstellungsgespräch eingeladen zu werden (Kaas und Manger, 2012). Vor diesem Hintergrund werden in der deutschsprachigen Forschung auch die Effekte eines anonymisierten Bewerbungsverfahrens diskutiert (Krause et al., 2012). In dem im Folgenden analysierten Auswahlverfahren der Polizei NRW spielen solche Effekte allerdings keine Rolle, da Bewerber strikt nach objektiven Kriterien wie Alter, Körpergröße oder Schulabschluss zum Verfahren eingeladen werden. Gleichwohl zeigen jährliche Analysen der Bewerbungs- und Einstellungsdaten, dass die Erfolgschancen von Bewerbern mit Migrationshintergrund in diesem für die Personalauswahl einer deutschen Behörde typischen mehrstufigen Verfahren (ein Test der kognitiven Fähigkeiten, ein Test zur Rechtschreibung in deutscher Sprache und einem darauf folgenden Assessment Center) deutlich schlechter sind als für Bewerber ohne Migrationshintergrund. Dieser Sachverhalt soll anhand von Daten aus diesem Auswahlverfahren im Hinblick auf die zentralen Auswahlkomponenten eingehend untersucht werden. Die gewonnenen Erkenntnisse sollen Strategieüberlegungen und Maßnahmen anleiten, wie die Chancen von Bewerbern mit Migrationshintergrund vergrößert werden können, ohne die Validität des Verfahrens zu gefährden. Die Kategorisierung der Bewerber und Einteilung in unterschiedliche Migrantengruppen (ausländische Staatsbürger, eingebürgerte Deutsche, Spätaussiedler, Deutsche mit mindestens einem ausländischen Elternteil) erfolgte auf der Grundlage von Bewerberangaben vorab durch die Behörde. Gestützt auf die oben angegebenen Forschungsbefunde werden folgende Hypothesen überprüft:

- Hypothese I: Bewerber mit Migrationshintergrund schneiden in den Tests und dem Assessment Center schlechter ab als Bewerber ohne Migrationshintergrund.
- Hypothese IIa: Die Erfolgsunterschiede zwischen Bewerbern mit vs. ohne Migrationshintergrund variieren mit dem Auswahlinstrument.
- Hypothese IIb: Beim kognitiven Fähigkeitstests stellen sich größere Subgruppenunterschiede ein als beim Assessment Center. Gestützt auf die allgemeine Forschungslage zu Subgruppendifferenzen bei Tests zu sprachlichen Fertigkeiten sollten die Unterschiede im Rechtschreibtest auf dem Niveau der Subgruppenunterschiede beim Assessment Center liegen.
- Hypothese IIIa: Das schlechtere Abschneiden von Bewerbern mit Migrationshintergrund wird durch die Art des Migrationshintergrundes moderiert.
- Hypothese IIIb: Da Bewerber mit ausländischer Staatsbürgerschaft vermutlich in ihrer Sozialisation weniger Nähe und Vertrautheit zur deutschen Sprache und zu Erfahrungs- und Verhaltensmustern der Mehrheitsgesellschaft aufbauen konnten, wird angenommen, dass diese im Verfahren am schlechtesten abschneiden; bei deutschen Bewerbern mit einem deutsch–ausländischen Elternpaar sollte diese Nähe und Vertrautheit dagegen stark gegeben sein und sie sollten am besten abschneiden. Die Erfolge von eingebürgerten Bewerbern und Spätaussiedlern sollten zwischen diesen beiden Gruppen liegen.

Die Hypothesen IIb bzw. IIIb stellen jeweils Konkretisierungen der umfassender formulierten Hypothesen IIa bzw. IIIa dar und setzen deren Gültigkeit voraus.

Weiterhin werden verfügbare biografische Daten wie insbesondere die Art des Schulabschlusses und Schulnoten genutzt, um diese sowohl in Bezug zum Abschneiden im Auswahlverfahren als auch in Beziehung zu auftretenden Subgruppenunterschieden zu setzen.

Daten und empirische Methodik

Die Behörde und ihr Personalauswahlprozess

Die Untersuchung erfolgt mit Daten aus dem Auswahlprozess für den gehobenen Polizeivollzugsdienst des Landes NRW. Die Polizei NRW beschäftigt im Jahr 2012 insgesamt 41.531 Polizeibeamte und -beamtinnen. Die Werbung und Auswahl für Einstellungen in den Polizeidienst werden vom Landesamt für Ausbildung, Fortbildung und Personalangelegenheiten der Polizei (LAFP) durchgeführt. Die Zahl der jährlichen Neueinstellungen liegt zwischen 1100 (2008) und 1500 (2015) Personen.

Voraussetzung einer Bewerbung für den hier betrachteten gehobenen Polizeivollzugsdienst der Polizei NRW – der mittlere Dienst wurde inzwischen abgeschafft, der höhere Dienst ist Aufsteigern bzw. Quereinsteigern mit abgeschlossenem Hochschulstudium (in der Regel Jura) vorbehalten – ist die (Fach-)Hochschulreife oder eine berufliche Aufstiegsfortbildung gemäß §2 Berufsbildungshochschulzugangsverordnung. Es können sich deutsche Staatsbürger sowie Staatsbürger aus EU-Mitgliedstaaten bewerben. Bei Vorhandensein eines „dienstlichen Interesses“ stellt die Polizei NRW auch Staatsbürger aus Nicht-EU-Staaten mit Aufenthalts- oder Niederlassungserlaubnis ein. Das Verfahren beginnt für alle Bewerber und Bewerberinnen mit einer Online-Bewerbung, in der standardmäßig und obligatorisch Angaben zum Geburtsort, zum Geburtsland, zur eigenen Staatsangehörigkeit sowie zur Staatsangehörigkeit der Mutter und des Vaters abgefragt werden. Bei Staatsangehörigen aus Nicht-EU-Ländern wird auch nach der Aufenthaltserlaubnis und danach gefragt, ob man seine Muttersprache spricht. Deutsche Staatsangehörige müssen angeben, ob sie Deutscher durch Geburt, ob sie Aussiedler sind und wann sie eingebürgert wurden (siehe www.polizeibewerbung.nrw.de). Bei Nicht-Beantwortung dieser Fragen kann die Online-Bewerbung nicht abgeschlossen werden. In der Folge sind der Behörde neben einem Lebenslauf und Nachweisen über den bisherigen Bildungsweg beglaubigte Kopien von Personaldokumenten, die Fahrerlaubnis Klasse B, eine Arztbescheinigung über den aktuellen Gesundheitszustand, eine Kopie des kürzlich bestandenen Deutschen Sportabzeichens sowie ein Schwimmnachweis vorzulegen. Für ausländische Schulzeugnisse ist die Anerkennung durch die zuständige Bezirksregierung erforderlich.

Das Auswahlverfahren der Polizei NRW, dem ein 20 Merkmale umfassendes Anforderungsprofil zugrunde liegt, ist mehrschrittig aufgebaut: Nach der Online Bewerbung und der administrativen Überprüfung der eingereichten Bewerbungsunterlagen auf Vollständigkeit, werden alle Bewerber, die die formalen Einstellungskriterien (z.B. Alter, Bildungsabschluss, Körpergröße usw.) erfüllen, zu dem am Personal Computer (PC) stattfindenden Auswahltest eingeladen. Eine Vorauswahl anhand inhaltlicher Leistungskriterien (z.B. Schulnoten) findet dabei nicht statt. Der zentral an einem LAFP-Standort durchgeführte PC-Test besteht aus einem kognitiven Test, der die analytischen Fähigkeiten sowie die Gedächtnisleistung misst, und einem Rechtschreibtest, der

Orthografie, Grammatik und Zeichensetzung prüfen soll und dem Anforderungsmerkmal der schriftlichen Kommunikationsfähigkeit im Anforderungsprofil zugeordnet wird. Diese beiden PC-gestützten Verfahren setzen sich aus je zehn Subtests zusammen. Bestehen die Kandidaten beide Testverfahren, wobei ein Cutoff-Wert von 85 bei Standardisierung der Testwerte auf einen Mittelwert von 100 und eine Standardabweichung von 10 nicht unterschritten werden darf, so erfolgt anschließend ein Formalgespräch sowie der Wiener Test zur Überprüfung von Reaktionsschnelligkeit, Aufmerksamkeit und Konzentrationsfähigkeit. Kandidaten, die die notwendige Punktezahl erreicht haben, werden dann vom Polizeiarzt auf Polizeidiensttauglichkeit untersucht. Bei bestandener Untersuchung erfolgt eine Einladung zum Assessment Center, dem nach dem PC-Test zweiten zentralen eignungsdiagnostischen Element des Auswahlverfahrens. Das Assessment Center wird dezentral in einer von 10 Einstellungsbehörden des Landes durchgeführt. Es setzt sich zusammen aus (1) einem Rollenspiel zu einer Konfliktsituation im Polizeidienst, (2) einem Kurzvortrag vor der Prüfungskommission, (3) einem Rollenspiel zu Multitasking-Fähigkeiten und (4) einem strukturierten Interview, bei dem neben Fragen zu Lebenslauf und Berufsmotivation auch Fragen zum Verhalten in kritischen Situationen gestellt werden. Die Auswahlkommissionen der Assessment Center bestehen aus zwei Polizeibeamten. Sowohl diese Assessoren als auch die ebenfalls aus dem Polizeidienst stammenden Rollenspieler durchlaufen eine eigens dafür entwickelte Schulung. Im Assessment Center werden die Teilnehmer in allen vier Teilverfahren im Hinblick auf zu beobachtende Anforderungsmerkmale bewertet; daraus wird ein Gesamtindikator für die Leistung im Assessment Center berechnet, für den wiederum ein Cutoff-Wert von 85 bei Standardisierung der Verteilung des Gesamtindikators auf einen Mittelwert von 100 und eine Standardabweichung von 10 gilt. Wird dieser Cutoff-Wert unterschritten, so ist man im Assessment Center gescheitert und scheidet aus dem Verfahren aus. Für die Kandidaten, die das Assessment Center bestanden haben und außerdem nicht durch nachträgliche Veränderungen ihrer Bewerbungsvoraussetzungen (z.B. durch ein laufendes Strafverfahren) aus dem Verfahren ausscheiden, wird ein sogenannter Rangordnungswert gebildet. In den Rangordnungswert gehen die Leistungen aller im Verfahren gemessenen Anforderungsmerkmale mit jeweils spezifischen Gewichten ein, wobei 17 dieser 20 Merkmale ausschließlich im Assessment Center erfasst werden. Anhand dieses Rangordnungswertes wird entschieden, ob ein Kandidat ein Angebot erhält oder nicht. Besteht z.B. die Vorgabe, dass in der gegenwärtigen Bewertungskampagne 1500 Personen einzustellen sind, so erhalten die Bewerber, die in der Reihenfolge des Rangordnungswertes die Plätze 1 bis 1500 inne haben, ein Einstellungsangebot.

Die eingestellten Personen absolvieren eine Ausbildung zum gehobenen Polizeivollzugsdienst, der als Einheitslaufbahn konzipiert ist. Diese beginnt mit einem dreijährigen Bachelorstudium an der Fachhochschule für öffentliche Verwaltung NRW, das neben theoretischen Anteilen auch Praxisanteile beinhaltet. In der darauf folgenden vierjährigen Erstverwendungszeit erfolgt ein Einsatz im Streifen- bzw. Wach- und Wechseldienst (mit Aufgabenschwerpunkt: Verkehrsüberwachung, Kriminalitätsprävention, Ordnungsstörungen) oder in einer Einsatzhundertschaft mit Aufgaben wie der Begleitung von Großveranstaltungen (Fußballspiele, Demonstrationen usw.). Bei Praxisbewährung besteht danach die Möglichkeit, sich

für spezielle polizeiliche Aufgabenfelder (z.B. Wirtschaftskriminalität) zu bewerben und dort Fach- und Führungsaufgaben zu übernehmen.

Stichprobe

Die von der Behörde zur Verfügung gestellten Daten umfassten insgesamt 3771 Bewerber, die in der Bewerbungskampagne 2011/2012 den PC-Test absolvierten. Damit handelt es sich um Bewerber, die sich im Online-Verfahren korrekt beworben und danach ihre schriftlichen Unterlagen bei der Behörde eingereicht haben, die die Voraussetzungen der formalen administrativen Vorauswahl (Alter, Körpergröße, Bildungsabschluss usw.) erfüllten und zum PC-Test antraten. Die Datensätze von 22 Bewerbern mussten ausgeschlossen werden, da diese unvollständig waren oder für diese Bewerber keine Zuordnung von soziodemographischen Daten und Testergebnissen möglich war. Insgesamt gingen somit die Daten von 3749 Bewerbern in die Analyse ein. Jeder Datensatz dieser 3749 Bewerber enthielt sowohl die erforderlichen soziodemografischen Angaben (z.B. Geburtsort, Staatsangehörigkeit, Migrationshintergrund, Noten im Abschlusszeugnis) als auch alle Ergebnisse des Auswahlverfahrens differenziert bis in Subtests und die einzelnen Teilaufgaben.

Von den insgesamt 3749 Bewerbern, die zum PC-Test angetreten sind, haben 635 Personen (16,9%) einen Migrationshintergrund. Auf der Grundlage von Selbstauskünften der Bewerber wurde die Art des Migrationshintergrundes von der Behörde vier Kategorien zugeordnet: (1) Eingebürgerte Deutsche (N=258; 40,7%); (2) Deutsche von Geburt an mit einem oder, im Ausnahmefall, zwei ausländischen Eltern (N=151; 23,8%); und (3) Spätaussiedler (N=85; 13,4%); Gruppe (4) sind ausländische Staatsbürger (N=140; 22,1%)[3].

404 Bewerber mit Migrationshintergrund (64%) sind in Deutschland, die übrigen 231 sind in 38 verschiedenen Ländern geboren, die Schwerpunkte liegen hier auf Kasachstan, Russland, Polen und der Türkei. Bei den Bewerbern mit ausländischer Staatsbürgerschaft dominiert die türkische Staatsbürgerschaft (47%) vor der griechischen (9%) und der italienischen (8%); die restlichen Staatsbürgerschaften verteilen sich auf 25 weitere Länder.

1226 der Bewerber sind weiblich (32,7%), wobei der Anteil der weiblichen Bewerber bei den Bewerbern mit Migrationshintergrund mit 23,6% geringer ist als bei den Bewerbern ohne Migrationshintergrund mit 34,6%. Das Durchschnittalter der Bewerber zum Stichtag 01.01.2012 beträgt 20,6 Jahre. Bewerber mit Migrationshintergrund liegen mit einem Mittelwert von 21,9 Jahren über diesem Altersdurchschnitt. Die allgemeine Hochschulreife (Abitur) besitzen 75,6% der Bewerber ohne Migrationshintergrund; bei den Bewerbern mit Migrationshintergrund sind dies mit 62,0% etwas weniger. Die übrigen Bewerber haben die Fachhochschulreife oder einen äquivalenten Bildungsabschluss.

2393 der 3749 Bewerber nehmen am späteren Assessment Center teil. Der Anteil der Teilnehmer mit Migrationshintergrund beträgt hier 13,5% und der Frauenanteil 34,2%.

Datenauswertung

Die Behörde stellte für die Analyse die Rohdaten des Auswahlverfahrens zur Verfügung, d.h. die Ergebnisse für alle Bewerber auf Itemebene der PC-Tests sowie die

Assessment-Center-Ratings. Diese Daten wurden gemäß der in der Behörde für Auswahlzwecke praktizierten Vorgehensweise verrechnet. Für jeden der 10 Subtests der beiden Testverfahren wurde als Rohpunktwert die Anzahl der gelösten Aufgaben berechnet; zum Ausgleich von Unterschieden der in mehreren Versionen vorliegenden Subtests wurde der Rohpunktwert jeweils z-transformiert und anschließend in eine Verteilung mit Mittelwert 100 und Standardabweichung 10 umgerechnet. Der Leistungsindikator für jeden der beiden Tests ergibt sich aus einer Addition der transformierten Subtestergebnisse mit wiederum anschließender Umrechnung in eine Verteilung mit Mittelwert 100 und Standardabweichung 10. Alle z-Transformationen erfolgten bezüglich der sich in der vorliegenden Stichprobe ergebenen Mittelwerte und Standardabweichungen in den jeweiligen Testversionen. Analog hierzu wird bei der Berechnung der Leistungen im Assessment Center vorgegangen: Die zu jeder Kombination von Anforderungsdimension und Übung vorliegenden Ratings wurden gemittelt und dann sowohl über die Übungen als auch die Anforderungsdimensionen aggregiert. Die sich so für jede Dimension ergebenden Leistungswerte werden zunächst bezüglich Mittelwert und Standardabweichung der Leistungen aller Bewerber in der Anforderungsdimension z-transformiert, in eine Verteilung mit Mittelwert 100 und Standardabweichung umgerechnet und anschließend addiert. Der sich so ergebende Summenwert wird wiederum in eine Verteilung mit Mittelwert 100 und Standardabweichung 10 umgerechnet und stellt in dieser Form den Leistungsindikator im Assessment Center dar.

Die Untersuchungshypothesen werden mittels Mittelwertevergleiche (t-test, Varianzanalysen und linearer Kontraste) per SPSS™ überprüft. Zur Kennzeichnung der Höhe der Mittelwertunterschiede zwischen den verschiedenen ethnischen Subgruppen wird das Effektstärkenmaß Cohen's d wie folgt berechnet (Roth et al., 2001: 299; Ellis, 2009):

$$\text{Cohen's d} = (M_1 - M_2) / SD_{pooled}, \text{ wobei } SD_{\text{pooled}} = \sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{n_1 + n_2 - 2}}$$

und M=Arithmetisches Mittel; SD= Standardabweichung, n=Umfang der Stichproben 1 und 2.

Für die Cohen's d-Werte werden Konfidenzintervalle gemäß Sedlmeier und Renkewitz (2008: 654) berechnet. Ausscheidequoten für die Subgruppen werden mit dem von der Behörde verwandten Cutoff-Wert von 85 errechnet und die sich ergebenden Anteilsunterschiede mittels Chi-Quadrat-Test auf Signifikanz geprüft. Um zu überprüfen, inwieweit der Migrationshintergrund über weitere Variablen wie Geschlecht, Alter, Schulabschluss oder Schulnoten hinaus Varianz im Abschneiden bei Testverfahren und Assessment Center erklärt, werden hierarchische Regressionsanalysen (vgl. z.B. Urban und Mayerl, 2011) durchgeführt.

Ergebnisse

Tabelle 1 stellt für die zentralen Untersuchungsvariablen die Mittelwerte, Standardabweichungen und Korrelationen dar.

Tabelle 1. Mittelwerte, Standardabweichungen und Korrelation der Untersuchungsvariablen.

Variablen	M	SD	1	2	3	4	5	6	7	8	9
1. Migrationshintergrund	0,17	0,38									
2. Geschlecht	0,67	0,47	.09**								
3. Alter	20,59	3,33	.18**	14**							
4. Schulabschluss	0,73	0,44	−.12**	−.09**	−.27**						
5. Deutschnote	7,89	2,30	−.11**	−.16**	.01	−.04*					
6. Mathematiknote	7,49	2,80	−.02	−.08**	−.07**	.02	.29**				
7. Test kognitive Fähigkeiten	100,00	10,00	−.23**	.10**	−.02	.26**	.11**	.14**			
8. Test Rechtschreibung	100,00	10,00	−.21**	−.11**	−.02	.31**	.14**	.08**	.57**		
9. Assessment Center	100,00	10,00	−.11**	−.13**	.09**	.08**	.20**	.06**	.26**	.17**	
10. Rangordnungswert	0,00	1,00	−.09**	−.09**	.11**	.09**	.18**	.06**	.36**	.27*	.98**

Anmerkung: Fallzahl je nach Korrelation zwischen 2.237 (für Korrelationen mit Rangordnungswert) bis zu 3749. In Zellen für Zusammenhänge mit den dichotomen Variablen 1, 2 und 4 punktbiserialer Korrelationskoeffizient, ansonsten Produkt-Moment-Korrelationskoeffizienten.
*p <.05, **p <.01, zweiseitige Fragestellung.
Codierung für die dichotomen Variablen: Migrationshintergrund (ja = 1; nein = 0), Geschlecht (männlich = 1; weiblich = 0), Schulabschluss (allgemeine Hochschulreife= 1; Fachhochschulreife oder Äquivalenz = 0). Schulnoten sind Punktwerte von 0 bis 15 für die beste Note. M: Arithmetisches Mittel; SD: Standardabweichung.

Tabelle 1 zeigt, dass der Migrationshintergrund der Bewerber mit einem schlechteren Abschneiden bei den beiden Auswahltests, dem Assessment Center sowie dem Rangordnungswert einhergeht. Außerdem fällt die mit r = .57 sehr hohe Korrelation zwischen dem kognitiven Fähigkeitstest und dem Rechtschreibtest auf. Die Schulnoten hängen positiv mit den Leistungen im Auswahlverfahren zusammen. Weiterhin ist der Schulabschluss ein guter Prädiktor für das Abschneiden insbesondere in den Testverfahren: Bewerber mit der allgemeinen Hochschulreife erzielen hier bessere Ergebnisse als Bewerber mit Fachhochschulreife oder einem hiermit äquivalenten Bildungsgrad. Männliche Bewerber schneiden im kognitiven Fähigkeitstest etwas besser ab, weibliche Bewerber dagegen in Rechtschreibtest und Assessment Center. Die sehr hohe Korrelation zwischen Rangordnungswert und Assessment Center ergibt sich daraus, dass 17 der 20 Anforderungsmerkmale, die in den Rangordnungswert eingehen und damit die Platzierung eines Kandidaten im Auswahlverfahren bestimmen, ausschließlich im Assessment Center gemessen werden.

Tabelle 2 stellt die Leistungen der Bewerber mit vs. ohne Migrationshintergrund bei den drei Auswahlverfahren, die Ergebnisse des t-Tests zur Prüfung der statistischen Signifikanz der Mittelwertunterschiede sowie die resultierenden Cohen's d-Werte für die Mittelwertdiskrepanzen dar.

Tabelle 2 zeigt, dass Bewerber mit Migrationshintergrund in allen drei Auswahlverfahren signifikant schlechter als Bewerber ohne Migrationshintergrund abschneiden. Hypothese I ist damit bestätigt.

Hypothese IIa postuliert, dass die Höhe der Mittelwertdiskrepanz gemessen am Indikator Cohen's d abhängig ist vom Auswahlinstrument. Da es in Tabelle 2 nicht überlappende Konfidenzintervalle für Cohen's d gibt, ist dies der Fall, und Hypothese IIa damit bestätigt. Die Relationen der Werte für Cohen's d fallen aber nur teilweise aus wie in Hypothese IIb postuliert: Das Konfidenzintervall für Cohen's d des kognitiven Fähigkeitstests liegt wie

Tabelle 2. Mittelwertdiskrepanzen in Abhängigkeit vom Auswahlverfahren.

	Bewerber ohne Migrationshintergrund			Bewerber mit Migrationshintergrund			t-Werte	Cohen's d [CI]
	Anzahl	MW	SD	Anzahl	MW	SD		
Test kognitive Fähigkeiten	3114	101,02	9,43	635	95,01	11,16	12,68***	0,62 [0,53;0,70]
Test Rechtschreibung	3114	100,93	9,56	635	95,45	10,84	11,83***	0,56 [0,47;0,65]
Assessment Center	2069	100,45	9,88	324	97,11	10,34	5,61***	0,34 [0,22;0,44]

Anmerkung: Bei Test kognitive Fähigkeiten und Test Rechtschreibung t-Werte unter Annahme nicht bestehender Varianzhomogenität. CI = 95% - Konfidenzintervall.
MW: Mittelwert; SD: Standardabweichung.
***p <.001, einseitig.

Tabelle 3. Mittelwertdiskrepanzen in Abhängigkeit von Auswahlverfahren und der Art des Migrationshintergrundes.

	Bewerber ohne MH	Bewerber mit MH							
	MW (SD)	Ausländische Staatsbürger		Eingebürgerte		Spätaussiedler		Deutsche mit ausländischem Elternteil	
		MW (SD)	d	MW (SD)	d	MW (SD)	d	MW (SD)	d
Test kognitive Fähigkeiten	101,02 (9,43)	91,06 (11,96)	1,04	94,04 (10,77)	0,73	97,90 (9,67)	0,33	98,74 (10,37)	0,24
Test Rechtschreibung	100,93 (9,56)	91,23 (11,82)	1,00	95,77 (10,85)	0,53	96,41 (9,54)	0,47	98,33 (9,40)	0,27
Assessment Center	100,45 (9,88)	95,31 (10,54)	0,52	96,59 (10,44)	0,39	95,96 (9,54)	0,45	99,67 (10,15)	0,08

Anmerkung: Derer d-Wert bezieht sich jeweils auf den Mittelwertvergleich zwischen den Bewerbern ohne Migrationshintergrund und der Bewerbern der jeweiligen ethnischen Subgruppe.
Anzahl N (Testverfahren / Assessment Center): Bewerber ohne MH (3114/2069); ausländische Staatsbürger (140/58); Eingebürgerte (258/125); Spätaussiedler (85/47); Deutsche mit mindestens einem ausländischen Elternteil (151/93).
MH: Migrationshintergrund; MW: Mittelwert; SD: Standardabweichung; d: Cohen's d.

erwartet ohne Überschneidungen oberhalb des Konfidenzintervalles für Cohen's d des Assessment Centers. Das Konfidenzintervall für das Cohen's d des Rechtschreibtests liegt anders als erwartet weitgehend überlappend mit dem Konfidenzintervall für den kognitiven Fähigkeitstest und zudem außerhalb des Konfidenzintervalles für das Assessment Center. Das Cohen's d des Rechtschreibtests ist somit signifikant größer als beim Assessment Center und unterscheidet sich nicht signifikant vom Cohen's d des kognitiven Fähigkeitstests. Hypothese IIb hat sich damit nur teilweise bestätigt.

Hypothese IIIa postuliert, dass die Höhe der Mittelwertunterschiede in den Auswahlverfahren von der Art des Migrationshintergrundes abhängig ist. Tabelle 3 zeigt hierzu die Ergebnisse.

Aus Tabelle 3 geht hervor, dass das Abschneiden der Bewerber mit Migrationshintergrund deutlich mit der Art des Migrationshintergrundes variiert. Einfaktorielle Varianzanalysen innerhalb der Bewerber mit Migrationshintergrund ergeben jeweils signifikante Effekte für die unabhängige Variable „Art des

Migrationshintergrundes“ auf die abhängige Variable der Leistung im kognitiven Fähigkeitstest ($F = 14{,}96$; $p < .001$), im Rechtschreibtest ($F = 11{,}42$; $p < .001$) und im Assessment Center ($F = 2{,}83$; $p < .04$). Damit ist Hypothese IIIa bestätigt.

Zur Überprüfung der Hypothese IIIb, die eine spezifische Leistungsreihenfolge der Subgruppen bei den Auswahlverfahren postuliert, wurden lineare Kontrastanalysen durchgeführt. Pro Auswahlverfahren werden hier drei Kontraste überprüft: (a) Ausländische Staatsbürger vs. Eingebürgerte/Spätaussiedler; (b) Eingebürgerte/Spätaussiedler vs. Deutsche mit mindestens einem ausländischen Elternteil; (c) Ausländische Staatsbürger vs. Deutsche mit mindestens einem ausländischen Elternteil. Alle drei Kontraste sind bei beiden Testfahren statistisch signifikant (bei Kontrast b im Rechtschreibtest $p < .05$ und im kognitiven Fähigkeitstest $p < .01$, alle weiteren Kontraste $p < .001$, einseitige Fragestellung). Beim Assessment Center sind lediglich die Kontraste b und c signifikant ($p < .01$ einseitige Fragestellung). Bei den beiden Testverfahren schneiden somit ausländische Staatsbürger signifikant schlechter ab als Eingebürgerte und Spätaussiedler, und diese wiederum schlechter als Deutsche mit mindestens einem ausländischen Elternteil. Beim Vergleich der Leistungen der Bewerber aus den unterschiedlichen Subgruppen mit den Bewerbern ohne Migrationshintergrund drückt sich dies in stark unterschiedlichen Cohen’s d-Werten aus, die z.B. beim kognitiven Fähigkeitstest von 1,04 bis 0,24 reichen. Für die Testverfahren bestätigt sich somit Hypothese IIIb vollständig, für das Assessment Center nur teilweise.

Weitere Ergebnisse, die über die in den Tabellen 1 bis 3 dargestellten Zusammenhänge sowie die unmittelbare Hypothesenüberprüfung hinausgehen, sind:

1. *Ergebnisse der Assessment-Center-Teilnehmer in den PC-Testverfahren*: Bei der oben angegebenen Überprüfung von Hypothese IIa hat sich gezeigt, dass die Höhe der Leistungsunterschiede zwischen Bewerbern mit vs. ohne Migrationshintergrund vom Auswahlverfahren abhängig ist. Für die beiden PC-Testverfahren sind diese höher als für das Assessment-Center-Verfahren (vgl. Tabelle 2). Hier ist aber zu beachten, dass es sich bei den Teilnehmern des Assessment Centers um eine vorselektierte Gruppe handelt, in der die in den PC-Testverfahren ausgeschiedenen Teilnehmer nicht mehr enthalten sind. Betrachtet man nur die Gruppe der Assessment-Center-Teilnehmer und vergleicht deren vorherige PC-Testleistungen für Kandidaten mit vs. ohne Migrationshintergrund, so unterscheiden sich die Werte für Cohen’s d nicht mehr signifikant: Für den kognitiven Fähigkeitstest ergibt sich ein d von 0,39 und ein 95% - Konfidenzintervall von [0,27;0,51], für den Rechtschreibtest ein d von 0,31 und ein Intervall von [0,19;0,43]; beide d-Werte unterscheiden sich nicht signifikant von den Werten des Assessment Centers mit einem Cohen’s d von 0,34 und dem Konfidenzintervall von [0,22;0,45] .
2. *Ausscheidequoten*: Die Behörde verwendet in den beiden Testverfahren sowie dem Assessment Center jeweils einen Cutoff-Wert von 85. Bewerber, die diesen Wert unterschreiten, scheiden somit an dieser Stelle aus dem Verfahren aus. Legt man diesen Cutoff-Wert an die Stichprobendaten an, so scheiden beim kognitiven Fähigkeitstest 5,3% der Bewerber ohne Migrationshintergrund

aus im Gegensatz zu 18,1% der Bewerber mit Migrationshintergrund (Chi-Quadrat-Wert: 126,3; $p < .001$); beim Rechtschreibtest liegen die entsprechenden Ausscheidequoten bei 5,5% zu 16,5% (Chi-Quadrat-Wert: 94,5; $p < .001$) und beim Assessment Center bei 5,6% zu 12,7% (Chi-Quadrat-Wert: 23,1; $p < .001$). Bei den beiden Testverfahren scheiden insgesamt 9,1% der Bewerber ohne Migrationshintergrund aus im Gegensatz zu 25,4% bei den Bewerbern mit Migrationshintergrund. Besonders sind bei den Testverfahren die Bewerber mit ausländischer Staatsangehörigkeit betroffen: 40,7% scheitern an einem der beiden oder an beiden Tests.

3. *Einstellungsquoten*: Nimmt man an, dass von den 2237 Kandidaten, die das Assessment Center bestanden haben, 1500 eingestellt werden sollen, so erhalten die hinsichtlich des Rangordnungskoeffizienten 1500 besten Bewerber ein Stellenangebot. Unter diesen 1500 Einzustellenden sind 162 Bewerber mit Migrationshintergrund und 1338 Bewerber ohne Migrationshintergrund. Da 635 Bewerber mit Migrationshintergrund in die Testverfahren gingen, liegt deren Erfolgsquote bei 25,5%. Bei den Bewerbern ohne Migrationshintergrund reüssieren letztlich 1338 von 3114, also 43,0%. Das ergibt ein Adverse-Impact-Ratio von 0,59 (25,5% zu 43,0%), was deutlich unter der in der US-amerikanischen Rechtsprechung favorisierten 80%-Regel liegt.
4. *Analysen auf Subtestniveau*: Bei allen 10 Subtests zu den kognitiven Fähigkeiten, bei den 10 Subtests zur Rechtschreibung sowie in allen vier Assessment-Center-Aufgaben schneiden Bewerber mit Migrationshintergrund im Mittel signifikant schlechter ab als Bewerber ohne Migrationshintergrund (jeweils $p<.01$). Die Cohen's d-Werte variieren dabei auf Subtestebene beträchtlich. Beim kognitiven Fähigkeitstest reichen die d-Werte von 0,13 bis 0,55, wobei besonders niedrige d-Werte bei Subtests mit keinen oder geringen sprachlichen Anforderungen auftreten (Figurenmatrizen, Zahlenreihen, Erinnerungsaufgaben usw.). Auch die Varianz in den Subtests zur Rechtschreibung ist beträchtlich (d-Werte von 0,13 bis 0,57). Auch hier treten höhere d-Werte bei sprachlich komplexeren Aufgaben auf. Beim Assessment Center betragen die d-Werte für die beiden Rollenspiele 0,17 und 0,41, für das Auswahlinterview 0,23 und für den Vortrag 0,32.
5. *Migrationshintergrund im Zusammenspiel mit weiteren Prädiktoren des Abschneidens im Verfahren*: Um zu überprüfen, inwieweit der Migrationshintergrund über weitere Variablen hinaus Varianz erklärt, wurden für jedes der drei Kriterien „Leistung im kognitiven Fähigkeitstest", „Leistung im Rechtschreibtest" und „Leistung im Assessment Center" hierarchische Regressionsanalysen (vgl. z.B. Urban und Mayerl, 2011) durchgeführt. In einem ersten Schritt wurden Geschlecht und Alter als Prädiktoren eingesetzt, in einem zweiten Schritt wurde der Bildungsabschluss hinzugefügt, dann in einem dritten Schritt die schulische Leistung wie die Mathematiknote (bei Kriterium „kognitiver Fähigkeitstest") oder Deutschnote (bei Kriterien „Rechtschreibtest" und „Assessment Center"). In einem vierten und letzten Schritt wurde schließlich der Migrationshintergrund als Prädiktor hinzugefügt.

Tabelle 4. Hierarchische Regressionsanalyse für das Kriterium des Abschneidens im kognitiven Fähigkeitstest.

Prädiktor	Kriterium: Leistung im kognitiven Fähigkeitstest			
	Modell 1 BETA	Model 2 BETA	Modell 3 BETA	Modell 4 BETA
Schritt 1: Kontrolle				
Alter	−0,029	0,046**	0,056**	0,088***
Geschlecht (männlich)	0,099***	0,115***	0,125***	0,139***
Schritt 2: Bildungsgrad				
Allgemeine Hochschulreife (ja)		0,286***	0,287***	0,271***
Schritt 3: Schulleistung				
Mathematiknote			0,152***	0,151***
Schritt 4: Kultureller Hintergrund				
Migrationshintergrund (ja)				−0,217***
R^2	0,010	0,085	0,107	0,152
Änderung in R^2	0,010***	0,075***	0,023***	0,045***

Anmerkung: BETA: Standardisierter Regressionskoeffizient. Geschlecht, Allgemeine Hochschulreife und Migration sind als Dummyvariablen kodiert, wobei die in Klammern genannte Variablenausprägung mit '1' kodiert ist, die gegenteilige Ausprägung mit '0'. Kollinearitätsstatistik: Toleranzen > 0,89.
N =3731. *= p<.05. **= p<.01. ***= p<.001.

Tabelle 4 zeigt die Ergebnisse für das Kriterium der Leistung im kognitiven Fähigkeitstest.

Aus Tabelle 4 wird deutlich, dass das Hinzufügen des Migrationshintergrundes selbst bei vorheriger Berücksichtigung dieser anderen Prädiktoren zu einer signifikanten Zunahme (p <.001) der Varianzaufklärung um 4,5 Prozentpunkte auf 15,2% führt. Ähnlich verhält es sich bei den beiden anderen Kriterien: Das Hinzufügen des Migrationshintergrundes im letzten Schritt führt beim Kriterium der Rechtschreibtestleistung zu einer signifikanten Steigerung (p <.001) der Varianzaufklärung um 2,6 Prozentpunkte auf 15,8%; für das Kriterium der Assessment-Center-Leistung ergibt sich eine signifikante Steigerung (p <.001) der Varianzaufklärung um 0,9 Prozentpunkte auf 8,5%. Der Migrationshintergrund leistet damit einen eigenen statistischen Erklärungsbeitrag zum Abschneiden in allen drei Verfahren, der nicht auf die Wirkung der übrigen Untersuchungsvariablen reduziert werden kann. Grafik 1 verdeutlicht ergänzend dazu das Zusammenwirken von Schulabschluss und Migrationshintergrund am Beispiel der Leistung im kognitiven Fähigkeitstest. Eine zweifaktorielle Varianzanalyse ergibt signifikante Haupteffekte für die Variablen „Schulabschluss" (F = 151,7; p <.001) und „Migrationshintergrund" (F = 144,4; p <.001), ohne dass der Interaktionseffekt zwischen beiden Variablen signifikant wird (F = 0,22; p < .641). Dieses Ergebnis findet sich in gleicher Form auch für den Rechtschreibtest und das Assessment Center. Schulabschluss und Migrationshintergrund sind also unabhängig voneinander wirkende Einflussfaktoren.

6. *Geburtsland der Migranten als Einflussfaktor*: Weiterhin wurde überprüft, ob Bewerber mit Migrationshintergrund, die im Ausland geboren sind (N = 231), in den Verfahren schlechter abschneiden als Bewerber mit Migrationshintergrund,

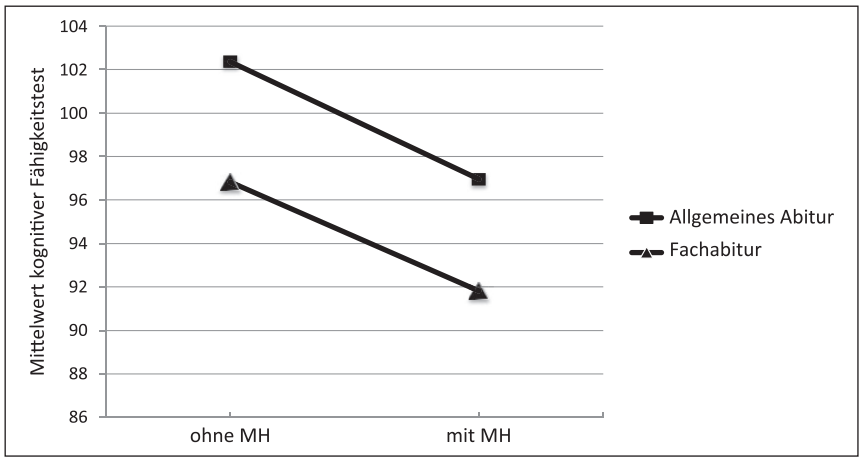

Abbildung 1. Abschneiden im kognitiven Fähigkeitstest in Abhängigkeit von Schulabschluss und Migrationshintergrund (MH) (*N* = 3.739).

die bereits in Deutschland geboren wurden (*N* = 404). Dies ist lediglich beim Rechtschreibtest der Fall (Mittelwert: 94,44 vs. 96,03; t = 1,79; $p < .037$, einseitige Fragestellung, d = 0,15), für den kognitiven Fähigkeitstest und das Assessment Center finden sich dagegen keine signifikanten Unterschiede.

7. *Ausländische Staatsbürgerschaft der Migranten als Einflussfaktor*: Da lediglich die türkische Staatsbürgerschaft in der Bewerberstichprobe in hinreichender Anzahl (N=66) vorhanden war, wurde das Abschneiden dieser spezifischen Gruppe im Vergleich zu den Bewerbern mit allen übrigen ausländischen Staatsbürgerschaften (N=75) untersucht. Es ergeben sich hier keine statistisch signifikanten Mittelwertunterschiede für das Abschneiden im kognitiven Fähigkeitstest und im Rechtschreibtest. Auch für die Leistung im Assessment Center ergab sich kein signifikanter Mittelwertunterschied, was hier aber angesichts des noch kleineren Stichprobenumfangs (N=27 und N=32) wenig aussagekräftig ist.
8. *Leistung im Rechtschreibtest und Deutschnote im Schulabschlusszeugnis*: Eine biografische Variable, die als Alternative zum Rechtschreibtest herangezogen werden könnte, ist die Deutschnote. Während der Rechtschreibtest für den Vergleich von Bewerbern mit vs. ohne Migrationshintergrund ein Cohen's d von 0,56 mit einem 95%-Konfidenzintervall [0,47;0,62] aufweist, ergibt sich für die Deutschnote in diesem Vergleich ein signifikant niedrigeres Cohen's d von 0,29 mit einem Konfidenzintervall von [0,21;0,38].

Diskussion

Ergebnisinterpretation

Die Hypothese I, dass Bewerber mit Migrationshintergrund im kognitiven Fähigkeitstest, im Rechtschreibtest und im Assessment Center schlechter abschneiden, hat sich

bestätigt. Untersuchungen aus den USA und Ländern wie z.B. den Niederlanden (De Meijer et al., 2006) zeigen, dass untersuchte Minoritätsgruppen gegenüber Majoritätsgruppen sowohl in kognitiven Fähigkeitstests (Roth et al., 2001), in Tests zu sprachlichen Fertigkeiten (Hough et al., 2001), in Auswahlinterviews (Roth et al., 2002) als auch in Assessment Centern (Dean et al., 2008) in der Regel schlechter abschneiden. Dieses Muster findet sich auch in dem Auswahlprozess der untersuchten deutschen Behörde. Die auftretenden Subgruppendifferenzen bereiten den Boden für einen erheblichen Adverse Impact in der Personalselektion, was sich bei einer Modellrechnung mit 1500 einzustellenden Personen in dem resultierenden Adverse-Impact-Ratio von 0,59 zeigt. Dieser Wert liegt deutlich unter der in der US-amerikanischen Rechtsprechung favorisierten 80%-Regel und weist auch auf eine Beeinträchtigung möglicher Diversitätsziele der Organisation hin.

Hypothese IIa postuliert, dass Subgruppendifferenzen mit der Auswahlmethodik variieren, was in der Untersuchung bestätigt wird. Für den kognitiven Fähigkeitstest ergibt sich im Vergleich der Bewerber ohne vs. mit Migrationshintergrund wie in Hypothese IIb erwartet ein signifikant höheres Cohen's d als beim Assessment Center (0,62 vs. 0,34). Abweichend von Hypothese IIb liegt der d-Wert für den Rechtschreibtest nicht signifikant unterhalb des d-Werts für den kognitiven Fähigkeitstest (0,62 vs. 0,56). Hierfür kommen zwei Erklärungen in Betracht: (1) Der Rechtschreibtest prüft auf sehr anspruchsvollem kognitiven Niveau Regelkenntnisse in deutscher Grammatik, Interpunktion und Orthographie, und korreliert von daher hoch mit dem kognitiven Fähigkeitstest ($r = .57$); (2) Während in den USA das Englische für weiße und schwarze Untersuchungsgruppen den Charakter einer Muttersprache haben dürfte, wird bei den in Deutschland lebenden Bevölkerungsgruppen mit Migrationshintergrund in den Familien häufig die Sprache des Herkunftslandes gesprochen. Haug (2008: 42) beziffert in einem Integrationsreport des Bundesamtes für Migration und Flüchtlinge noch für die Zeit nach der Jahrtausendwende den Anteil der Migranten, in deren Familie Deutsch gesprochen wird, mit 17,9%, den Anteil derjenigen, die die Herkunftslandsprache sprechen, mit 44,8%, und die einen Mix aus Muttersprache und Deutsch sprechen mit 31,7%.

Die Diskrepanz hinsichtlich der d-Werte beim Vergleich von kognitivem Fähigkeitstest vs. Assessment Center muss allerdings nicht zwangsläufig der Auswahlmethodik geschuldet sein, sondern kann ein Effekt der zunehmenden Selektion der Bewerber sein. Geht man von den Bewerbern aus, die nicht in den PC-Tests ausgeschieden sind und die das Assessment Center erreicht haben, so ergibt sich für diese selektierte Bewerbergruppe hinsichtlich ihrer Leistungen im kognitiven Fähigkeitstest beim Vergleich der Bewerber mit vs. ohne Migrationshintergrund lediglich ein Cohen's d von 0,39, was sich nicht signifikant von dem Cohen's d von 0,34 für das Assessment Center unterscheidet. Die Reduktion des d-Wertes im kognitiven Fähigkeitstest von 0,62 für die Teilnehmer an den PC-Tests zu 0,39 für die Assessment-Center-Teilnehmer ist nachvollziehbar: 18,1% der Bewerber mit Migrationshintergrund scheiden aufgrund schwacher Ergebnisse im kognitiven Fähigkeitstest aus; bei den Bewerbern ohne Migrationshintergrund sind dies lediglich 5,3%. Der Mittelwert der verbliebenen Bewerber für den kognitiven Fähigkeitstest steigt dadurch bei Bewerbern mit Migrationshintergrund stärker als bei den Bewerbern ohne Migrationshintergrund; trotz der ebenfalls abnehmenden Standardabweichungen reduziert sich der d-Wert bereits auf

0,42. Weitere Selektionsprozesse, unter anderem aufgrund des Rechtschreibtests, führen dann zur weiteren Angleichung des d-Wertes bis auf 0,39 bei den Teilnehmern des Assessment Centers.

Für eine bereits stark selektierte Bewerbergruppe ergeben sich in dieser Untersuchung für Assessment Center und kognitiven Fähigkeitstest also vergleichbare Leistungsunterschiede von Bewerbern mit und ohne Migrationshintergrund. Es stellt sich die Frage, ob die in der Literatur aufgeführten niedrigeren d-Werte für Assessment Center im Kontrast zu kognitiven Fähigkeitstests nicht zumindest teilweise auch auf diesen Selektionseffekt zurückgehen. Assessment Center sind kostspielige Methoden und kommen üblicherweise erst in späteren Selektionsschritten zum Einsatz als die kostengünstigen Testverfahren. Ergebnis einer Reihe von Selektionsschritten ist oftmals, dass der verbleibende Personenkreis niedrigere ethnische Subgruppenunterschiede aufweist (Roth et al., 2001: 302): So fallen ethnische Subgruppendifferenzen in kognitiven Fähigkeiten bei Stelleninhabern niedriger aus als bei Bewerbern, bei Studenten niedriger aus als bei Schülern und Hochschulbewerbern, und auch mit steigender Komplexität der Anforderungen von Arbeitsplätzen liegen, unter anderem aufgrund von Selbstselektionsprozessen, geringere ethnische Subgruppendifferenzen vor (Roth et al., 2001).

Hypothese IIIa postuliert, dass die Art des Migrationshintergrundes das Abschneiden in den Auswahlverfahren moderiert. Wie die Varianzanalysen hierzu zeigen, ist dies tatsächlich der Fall. Die in Hypothese IIIb spezifizierte Reihenfolge wird durch die Kontrastanalysen weitgehend bestätigt. Am schlechtesten schneiden in den beiden Testverfahren die ausländischen Staatsbürger ab. Für den kognitiven Fähigkeitstest resultiert beim Vergleich dieser Gruppe mit den Bewerbern ohne Migrationshintergrund ein Cohen's d von 1,04, was ungefähr dem typischen Ergebnis beim Vergleich kognitiver Testleistungen von weißen vs. schwarzen US-Amerikanern entspricht (Roth et al., 2001). Die in allen Verfahren beste Migrantengruppe sind dagegen Bewerber, die von Geburt an die deutsche Staatsbürgerschaft hatten, und bei denen in der Regel entweder der Vater oder die Mutter (und nur in Ausnahmefällen beide Elternteile) eine ausländische Staatsbürgerschaft haben. Diese Familienkonstellation dürfte für die Sozialisation in die deutsche Mehrheitsgesellschaft vergleichsweise günstig sein. Spätaussiedler und eingebürgerte Deutsche liegen bei den Testverfahren zwischen den beiden anderen Gruppen, beim Assessment Center unterscheiden sie sich allerdings nicht signifikant von den ausländischen Staatsbürgern. Diese Ergebnisse zeigen, dass es wichtig ist, zwischen den Migrantengruppen zu differenzieren: Die Leistungsunterschiede zwischen den Subgruppen sind beträchtlich.

In der Untersuchung von De Meijer et al. (2006) zur Personalauswahl der Polizei in den Niederlanden wurden beträchtliche Mittelwertunterschiede zwischen Migranten der ersten und zweiten Generation ermittelt. Da in der vorliegenden Untersuchung keine direkte Information zu dieser Generationsvariablen vorlag, wurde als Annäherung hierfür die Angabe zum Geburtsland verwendet. Lediglich im Rechtschreibtest schneiden im Ausland geborene Migranten signifikant schlechter ab als in Deutschland geborene, das Cohen's d für diesen Unterschied ist mit 0,15 aber gering.

Die Analysen auf Subtestniveau im kognitiven Fähigkeitstest zeigen beträchtliche Unterschiede hinsichtlich der d-Werte zwischen Bewerbern mit vs. ohne

Migrationshintergrund. Dies entspricht dem Forschungsstand: Die höchsten Diskrepanzen zwischen ethnischen Subgruppen treten in der Regel bei Messungen des allgemeinen Intelligenzfaktors g auf, kleinere Diskrepanzen sind dagegen bei spezifischen Intelligenzaspekten zu finden wie z.B. der mathematischen oder verbalen Intelligenz (Roth et al., 2001; Outtz und Newmann, 2010). Da die niedrigsten d-Werte in dieser Untersuchung für sprachfreie oder sprachreduzierte Subtests vorliegen, liegt die Vermutung nahe, dass Unterschiede in den sprachlichen Fähigkeiten maßgeblich zu den unterschiedlichen d-Werten der Subtests beitragen. Man darf annehmen, dass gerade Intelligenzmaße, die starke subethnische Differenzen hervorbringen, stärker durch Faktoren wie kulturelle oder sprachliche Aufladung verzerrt sind, und deswegen möglicherweise weniger konstruktvalide sind (Goldstein et al., 2010).

Bei dem Vergleich der zwischen Bewerbern mit vs. ohne Migrationshintergrund liegenden Mittelwertunterschiede von Rechtschreibtest und Deutschnote im Schulabschlusszeugnis ergaben sich deutliche niedrigere d-Werte für die Deutschnote (0,56 vs. 0,29). Dass Bildungsleistungen niedrigere ethnische Subgruppendifferenzen als kognitive Fähigkeitstests aufweisen, entspricht dem Forschungsstand (Ployhart and Holtz, 2008). Bildungsleistungen sind nicht nur von kognitiven Fähigkeiten, sondern auch von motivationalen Faktoren abhängig, und schlechtere Bildungsausgangslagen lassen sich zum Teil ausgleichen durch eine höhere Motivation und verstärkten Einsatz. De Meijer et al. (2006) fanden z.B. höhere Gewissenhaftigkeitswerte bei Migranten als bei Bewerbern ohne Migrationshintergrund, was darauf hindeutet, dass Migranten, die beträchtliche formale Einstiegshürden wie z.B. die Hochschulreife gemeistert haben, dies unter anderem mit vermehrtem Einsatz und Ehrgeiz schafften und so Handicaps wie z.B. sprachbedingte Schwierigkeiten oder weniger Unterstützung durch das Elternhaus zum Teil kompensieren konnten.

Die regressionsanalytischen Ergebnisse zeigen, dass das Merkmal Migrationshintergrund auch bei Berücksichtigung anderer Untersuchungsvariablen wie Geschlecht, Alter, Schulnoten oder Art des Schulabschlusses ein signifikanter Erklärungsfaktor für das Abschneiden in allen drei Auswahlverfahren bleibt. Leistungsunterschiede im Auswahlverfahren zwischen Bewerbern mit vs. ohne Migrationshintergrund sind somit insbesondere nicht einfach durch den Verweis auf einen niedrigeren Bildungsabschluss (allgemeine vs. fachgebundene Hochschulreife) zu erklären.

Praktische Implikationen

Was kann die Wissenschaft einer Behörde praktisch anraten, bei der ein massiver Adverse Impact festgestellt wurde? Einfach durch Forcierung des Personalmarketings den Pool an Bewerbungen aus dem Migrationsmilieu zu erhöhen, wie das Lindsey et al. (2013) neben anderen Maßnahmen vorschlagen, ist sicherlich keine zielführende Maßnahme: Die Wahrscheinlichkeit ist vielmehr hoch, dass der Adverse Impact weiter ansteigt (Arthur and Woehr, 2013). Aussichtsreicher könnte es sein, gezielte Werbemaßnahmen für die anvisierten Milieus zu entwerfen und schon im Marketingprozess auf eine bessere Passung der Bewerbungen zu den Arbeits- und Organisationsanforderungen Wert zu legen (Newman and Lyon, 2009). Generell wird man jedenfalls nicht versprechen können, den Adverse Impact einfach „auszumerzen“ („eradicating“), wie das der Titel der Veröffentlichung von Lindsey et al. (2013) nahelegt.

Der Versuch einer Vermeidung oder auch nur einer Verringerung des Adverse Impact in einem bestehenden Verfahren führt in grundsätzliche Probleme: Die vorliegende Untersuchung konnte zwar beträchtliche ethnische Subgruppenunterschiede feststellen, die Frage nach den Ursachen aber nicht beantworten. Auch die Frage, ob die ermittelten Subtestdifferenzen auf reale Differenzen in den Merkmalsausprägungen verweisen, das Testverfahren also nur der Indikator hierfür ist, oder aber durch das Verfahren als methodisches Artefakt selbst erzeugt werden (Arthur et al., 2013: 479), lässt sich nicht abschließend aufklären. Verzerrende Effekte des Verfahrens können auf ein breites Spektrum von Einflussfaktoren zurückgehen, von einer starken Sprachgebundenheit einzelner Arbeitsanweisungen bis hin zu einer Kulturlastigkeit ganzer Subtests. Wichtiger als die Frage nach „den" Ursachen von Subgruppenunterschieden ist unter praktischen Gesichtspunkten aber die Frage, wie diese Differenzen reduziert werden können, ohne die Validität des Verfahrens signifikant abzusenken. Auf der Basis von Forschungsergebnissen sind hierzu aussichtsreiche Strategien und Maßnahmen vorgeschlagen worden (z.B. Ployhard and Holtz, 2008; De Soete et al., 2012; Lindsey et al., 2013), die zwar eine Richtung für praktische Lösungen angeben, aber für eine konkrete Umsetzungssituation nicht unbedingt eine Erfolgsgarantie darstellen.

In der Auswahlpraxis großer Verwaltungsbehörden ist man also mit erheblichen Schwierigkeiten und Hindernissen konfrontiert: (a) Die Ursache für die in bestimmten Verfahren auftretende Spreizung zwischen Subgruppen bleibt mehrdeutig; (b) die Installation eines reformierten Verfahrens ist hochkomplex und im Grunde nur iterativ generierbar; (c) zahlreiche Bausteine, die in Aussicht stellen, dem Ziel einer Diversity-gerechten Personalauswahl besser zu entsprechen, so z.B. Simulationsmethoden, sind aufwändiger und damit teurer als standardisierte Testverfahren. Alle diese Probleme lassen sich sehr schlecht im Rahmen der Arbeits- und Entscheidungsstruktur großer Behörden bearbeiten und einer Lösung zuführen.

Da der Versuch der Überarbeitung und Reformierung eines Auswahlverfahrens mit dem Ziel, den Adverse Impact zu reduzieren, von solchen Schwierigkeiten überschattet ist, und ein „Großer Wurf" in der Regel nicht kurzfristig zu erwarten ist, sollte die Überarbeitung des Auswahlverfahrens unter Diversity-Gesichtspunkten als längerfristiger Prozess angelegt und von kompetenten Fachexperten begleitet werden. Grundlagenkonzepte wie die von der Bundesakademie für öffentliche Verwaltung herausgegebene Handreichung zur Interkulturellen Personalauswahl (2012) können dabei zusätzliche Orientierung geben. Diese Handreichung definiert klassische Bereiche von Schwierigkeiten, die einer interkulturellen Öffnung der Personalauswahl von Behörden entgegenstehen, und an denen Verbesserungsmaßnahmen ansetzen können.

Zur Reduktion der ethnischen Subgruppendifferenzen und des damit einhergehenden Adverse Impact in der hier betrachteten Behörde kommen folgende Ansatzpunkte in Betracht:

1. *Verzicht auf den bisherigen Rechtschreibtest*: Dieser Test ist dazu gedacht, das Anforderungsmerkmal „schriftliche Kommunikationsfähigkeit" zu überprüfen. Nach unseren Untersuchungen produziert er beträchtliche Subgruppenunterschiede und, da er auch zur Vorselektion eingesetzt wird, dreimal höhere Ausscheidequoten bei Bewerbern mit Migrationshintergrund als bei Bewerbern ohne Migrationshintergrund (5,5% vs. 16,5%). Es ist aber fraglich, ob der Test überhaupt das richtige Instrument ist, um „schriftliche Kommunikationsfähigkeit"

zu erfassen. Der Test erfasst Grammatik- und Orthografiekenntnisse auf einem sehr anspruchsvollen Niveau. Die Bewerber sollen aber später keine Deutschlehrer werden, sondern, wie die Anforderungsinformationen nahelegen, in ihrer beruflichen Tätigkeit Sachverhalte (z.B. Unfallgeschehnisse, Zeugenaussagen usw.) auf Deutsch sachlich richtig, verständlich und korrekt schriftlich festhalten können. Alternativ könnte man diese Fähigkeit auch über eine geeignete Arbeitsprobe im Assessment Center realitätsnah erfassen (z.B. Protokoll zu einem auf Video gezeigten Unfallgeschehen schreiben); dies würde der Empfehlung entsprechen, Adverse Impact durch verstärkten Einsatz von Simulationen zu reduzieren (De Soete et al., 2012). Will man den Rechtschreibtest dennoch im Auswahlverfahren belassen, so sollte dessen Rolle als Vorauswahlinstrument hinterfragt werden. Bestehende Defizite in der aufgabenbezogenen schriftlichen Kommunikation ließen sich durch Trainingsmaßnahmen (z.B. Schulungen zur Protokoll- und Berichtserstellung) verbessern. Will man generell Kenntnisse in Deutsch als Schriftsprache als Vorauswahlmerkmal nutzen, so bietet sich als Alternative die Deutschnote an, die geringere Subgruppendifferenzen als der Rechtschreibtest erzeugt.

2. *Reduktion sprachlicher Anforderungen im kognitiven Fähigkeitstest*: Simulationsuntersuchungen zum Ausbalancieren von Validitäts-Diversitäts-Zielen (Finch et al., 2009) legen nahe, auf kognitive Fähigkeitstests im Zuge der Vorauswahl ganz zu verzichten. Dazu würden wir nicht raten, da kognitive Fähigkeitstests nach wie vor gute Prädiktoren für beruflichen Erfolg darstellen und Instrumente wie Persönlichkeitstests, die auf Selbstbeschreibungen basieren, als Vorauswahlinstrument ohne Berücksichtigung von z.B. Auswahlinterviewdaten in diagnostischer Hinsicht problematisch sind (Hossiep und Mühlhaus, 2015: 150) und zu Akzeptanzproblemen bei den Bewerbern führen dürften. Allerdings halten wir es für sinnvoll, die sprachlichen Anforderungen im kognitiven Fähigkeitstest einer Prüfung zu unterziehen und sie zu reduzieren, wenn sie für die Erfassung der analytischen Fähigkeiten nicht erforderlich sind. Dies entspräche der grundlegenden Interventionsstrategie, konstruktirrelevante Varianz in Prädiktoren zu reduzieren (vgl. z.B. De Soete et al., 2012).
3. *Erweiterung des Anforderungsprofils um Fähigkeiten in für den Arbeitsalltag relevanten Fremdsprachen*: Wie eine Anforderungsanalyse für interkulturelle Arbeitsfelder in der Polizeiarbeit zeigen konnte (Leenen et al., 2014b), hat der zuwanderungsbedingte gesellschaftliche Wandel beispielsweise das Beherrschen einer oder mehrerer Fremdsprachen zu einer beruflich hochrelevanten Kompetenz im Behördenbereich oder in der Praxis der Polizei werden lassen. Bisher wird für die Einstellung in den Polizeidienst als Fremdsprache lediglich Englisch vorausgesetzt. Weitere Kenntnisse in den im Kontakt zwischen Polizei und unterschiedlichen Migrantengruppen wünschenswerten Sprachen werden nicht betrachtet. Sprachkenntnisse in sogenannten Bedarfssprachen positiv in das Kompetenzprofil aufzunehmen, verschafft die Möglichkeit, eine gewisse Tendenz zum schlechteren Abschneiden von Personen mit Migrationshintergrund, die Deutsch nicht als Muttersprache gelernt haben, durch Nutzung des Vorteils der Mehrsprachigkeit zu kompensieren. Diese Intervention entspricht dem Vorschlag von Ployhart und Holtz (2008), die für eine Arbeitstätigkeit erforderlichen Kompetenzen vollständig zu erfassen

4. *Interkulturelles Wahrnehmungs- und Bewertungstraining*: Jede interaktive Diagnostik, also jede Situation, in der die Leistungen von Bewerbern direkt von Beobachtern bzw. Ratern eingeschätzt und bewertet werden müssen, ist hochgradig anfällig für subjektive Einschätzungen und für den Einfluss nicht weiter reflektierter kultureller Selbstverständlichkeiten. Denk- und Verhaltensweisen von Bewerbern mit einem anderen Kulturhintergrund werden schnell nicht nur als überraschend und unerwartet, sondern negativ gewertet. Die Möglichkeit solch kultureller Fehlattribuierungen ist vor allem ein Problem in Auswahlinterviews sowie in simulationsorientierten Übungen im Assessment-Center-Verfahren. Es ist eine Aufgabe speziell auf die kulturelle Attributionsproblematik zugeschnittener Trainings, die Beurteiler für mögliche Wahrnehmungsverzerrungen und Fehlattributionen im Bewertungsprozess zu sensibilisieren. Entsprechende Schulungen sind also eine weitere Möglichkeit, einen substanziellen Beitrag zu einer diversity-gerechten Personalauswahl zu leisten. Diese Interventionsmöglichkeit wird in der Forschungsliteratur weitgehend übersehen und findet sich als Strategieempfehlung lediglich in Lindsey et al. (2013).

Im Zuge der Überarbeitung des Verfahrens sollten kontinuierlich Daten dazu erhoben werden, wie sich Veränderungen im Auswahlverfahren auf die beobachteten Subgruppendifferenzen auswirken. Weiterhin sind Untersuchungen zur Kriteriumsvalidität erforderlich, um so die Beziehungen zwischen den mit den Auswahlmethoden gemessenen Merkmalen und Validitätskriterien wie Leistung im Polizeidienst oder Studienerfolg an der Polizeihochschule zu erfassen. Nur so kann man überprüfen, ob die ergriffenen Maßnahmen zur Senkung von Subgruppendifferenzen und Adverse Impact im Auswahlverfahren noch eine hinreichende Validität für Auswahlzwecke garantieren. Validitätsdaten sind ferner erforderlich, um zu überprüfen, ob Unterschiede zwischen ethnischen Subgruppen im Auswahlverfahren tatsächlich mit realen Leistungsunterschieden im Organisationskontext einhergehen. Hat das schlechtere Abschneiden von Subgruppen im Auswahlverfahren keine Entsprechung zu Unterschieden in der beruflichen Produktivität, ist das Auswahlverfahren als schlicht unfair und gegen Gerechtigkeitsprinzipien verstoßend abzulehnen (Cleary, 1968; Cook, 2009: 277).

Fazit und Ausblick

Die vorliegende Studie untersucht ethnische Subgruppendifferenzen im Kontext der Personalauswahl deutscher Organisationen und greift somit das in Deutschland noch vernachlässigte Thema der Adverse-Impact-Forschung auf. Am Beispiel einer großen deutschen Behörde wird analysiert, inwieweit Bewerber mit vs. ohne Migrationshintergrund im Personalauswahlverfahren unterschiedlich gut abschneiden und inwiefern dies von der Auswahlmethodik und der Art des Migrationshintergrundes abhängig ist. Angesichts eines zunehmenden Anteils von Personen mit Migrationshintergrund an der Gesamtbevölkerung sieht sich speziell der Öffentliche Dienst mit der Herausforderung konfrontiert, nicht nur den Anteil solcher Personen bei den Bewerbungen, sondern letztlich auch die auf diesen Personenkreis entfallende Quote an Einstellungen weiter zu erhöhen. Scheitern solche Bewerbungen überdurchschnittlich häufig, kann das als ein Zeichen für versteckte Zugangshindernisse interpretiert werden.

In der vorliegenden Untersuchung zeigte sich im Personalauswahlverfahren dieser Behörde, der man eine Verletzung des Gleichbehandlungsgrundsatzes gerade nicht vorwerfen kann, ein Adverse Impact von ganz erheblichem Ausmaß. Dabei sind die Leistungsunterschiede der ethnischen Subgruppen in den in einer frühen Selektionsphase eingesetzten Testverfahren zur Messung der kognitiven Fähigkeiten und der Rechtschreibung am größten. Subgruppendifferenzen finden sich in geringerem Ausmaß auch im Assessment Center, das in einer späteren Auswahlphase eingesetzt wird. Geringere Differenzen im Assessment Center entsprechen durchweg der Forschungslage; dies muss nach der vorliegenden Untersuchung jedoch nicht an der Methodik liegen, sondern kann auch darauf zurückzuführen sein, dass die Teilnehmer am Assessment Center bereits eine stark vorselektierte Gruppe sind.

Die Höhe der Subgruppenunterschiede ist in erheblichem Maße davon abhängig, welche Migrantengruppe betrachtet wird. Die höchsten Abweichungen zur Majorität liegen für ausländische Staatsbürger vor, während die niedrigsten Abweichungen zur Majorität bei Bewerbern bestehen, die von Geburt an die deutsche Staatsbürgerschaft besitzen, und lediglich einen ausländischen Vater und/oder eine ausländische Mutter haben. Migranten sind hinsichtlich der Gefahr des Adverse Impact auf keinen Fall als homogene Bewerbergruppe anzusehen.

Aus den Ergebnissen der Untersuchung werden Handlungsempfehlungen für die Modifikation des Auswahlverfahrens der betrachteten Behörde abgeleitet. Die Umsetzung dieser Handlungsempfehlungen sollte als längerfristiger Prozess angelegt werden, in dessen Zuge kontinuierlich die Auswirkung von Reformmaßnahmen auf die Höhe der Subgruppenunterschiede erfasst und zugleich die Validität des Auswahlverfahrens ermittelt wird.

Das Auswahlverfahren der Polizei NRW untersuchen zu können, hat forschungspolitisch die Chance eröffnet, einen in der Öffentlichen Verwaltung typischen Auswahlprozess mit den klassischen Instrumenten eines PC-gestützten Rechtschreibtests, eines PC-Tests zur Überprüfung analytischer Fähigkeiten sowie eines nachgeschalteten Assessment Centers durchleuchten zu können. Da in die Untersuchung ein ganzer Jahrgang von Bewerbungen Eingang finden konnte, und das hier untersuchte Auswahlverfahren in seiner Mehrstufigkeit und dem darin eingesetzten Instrumentarium (vgl. z.B. Schuler et al., 2007) typisch erscheint, sind die Befunde für die untersuchten Subgruppen hochsignifikant und prinzipiell auch auf ähnlich gelagerte Auswahlverfahren anderer Behörden oder Unternehmen übertragbar.

Allerdings sind weitere Untersuchungen in der Personalauswahl deutscher Organisationen sinnvoll, um ethnische Subgruppenunterschiede zu identifizieren, und um ein besseres Verständnis für die soziokulturellen Ursachen der beobachteten Unterschiede zu gewinnen. Deswegen sollten in Folgeuntersuchungen deutlich mehr Variablen erfasst werden als dies in der vorliegenden Studie möglich war. So wäre es wünschenswert, nicht nur die Zugehörigkeit zu einer bestimmten Zuwanderergeneration, sondern auch Daten zum frühkindlichen Spracherwerb, zur Mehrsprachigkeit und zum Erlernen der deutschen Sprache sowie zum Besuch von Horten und Kindergärten der Mehrheitsgesellschaft zu erfassen, um Anhaltspunkte für die Intensität des „Eintauchens“ in die deutsche Sprache und Kultur zu gewinnen, und diese ins Verhältnis zu Erfolgschancen bei unterschiedlichen Auswahlinstrumenten setzen zu können. Die

Subgruppendifferenzen in den Auswahlverfahren spiegeln zudem offensichtlich nicht nur Kulturunterschiede, sondern auch spezifische sozio-ökonomische Lebensbedingungen der Bewerber wider, die ebenfalls zu erfassen wären (z.B. Einkommen und sozialer Status der Eltern, Qualität der besuchten Schulen, ländliche Lebensregion vs. städtisches Ballungszentrum usw.) Wenn sich Leistungsunterschiede in Auswahlverfahren letztendlich auf Unterschiede in der Qualität der sozio-ökonomischen Lebensbedingungen zurückführen ließen, so wäre dies ein weiterer Hinweis darauf, wie sehr es wichtig ist, auf gesellschaftlicher Ebene allen ethnischen Gruppierungen Chancengleichheit auf hohem Niveau zu ermöglichen. Behörden und Unternehmen sollten aber solche gesellschaftlichen Veränderungsprozesse nicht untätig abwarten, sondern ihre Personalauswahl dahingehend überprüfen, ob diese nicht einzelnen Bewerbergruppen schlechtere Erfolgschancen gibt, so dass darunter Diversitätsziele der Organisation leiden und/oder Fairnessgesichtspunkte verletzt werden. In diesem Zusammenhang werden dringend weitere Untersuchungen dazu benötigt, welche Maßnahmen zur Reduktion eines Adverse Impact bei der Personalauswahl in deutschen Organisationen besonders wirksam sind, ohne dabei die Validität des Gesamtverfahrens bedeutsam zu reduzieren. Insgesamt sollte sich die hohe gesellschaftliche Relevanz dieses Themas verstärkt in weiteren Forschungsanstrengungen widerspiegeln.

Fußnoten

1. Die hier berichteten Ergebnisse entstanden im Rahmen des Projekts „Interkulturelle Kompetenz und Inklusion in der Personalauswahl der Polizei (IKIP)“, das durch das XENOS-Programm „Integration und Vielfalt“ der EU von 2012 bis 2014 gefördert wurde. Die Autoren danken dem Landesamt für Ausbildung, Fortbildung und Personalangelegenheiten der Polizei in Nordrhein-Westfalen (LAFP) für die produktive Zusammenarbeit. Die Autoren danken außerdem zwei unbekannten Gutachtern oder Gutachterinnen für die wertvollen Anregungen zur Optimierung des Manuskriptes.
2. Aus Gründen der Lesbarkeit wird im Folgenden auf *political correctness* in der Schreibweise verzichtet: prinzipiell sind immer beide Geschlechter gemeint. Wenn speziell männliche oder weibliche Personen angesprochen sind, wird das kenntlich gemacht.
3. Für eine Person mit Migrationshintergrund lag keine eindeutige Zuordnung vor, weswegen die Daten dieser Person bei den Subgruppenanalysen, in denen die Art des Migrationshintergrundes bedeutsam war, ausgeschlossen wurde.

Literaturverzeichnis

Arthur W Jr and Woehr D (2013) No steps forward, two steps back: The fallacy of trying to ‘eradicate’ adverse impact? *Industrial and Organizational Psychology* 6(4): 438–442.

Arthur W Jr, Doverspike D, Barrett GV, et al. (2013) Chasing the title VII Holy Grail: The pitfalls of guaranteeing adverse impact elimination. *Journal of Business Psychology* 28(4): 473–485.

Bortz J and Döring N (1995) *Forschungsmethoden und Evaluation* (2., vollständig überarbeitete und aktualisierte Auflage). Berlin: Springer.

Bundesakademie für öffentliche Verwaltung im Bundesministerium des Innern (2012) Interkulturelle Öffnung der Personalauswahl im öffentlichen Dienst. Handreichung. Available at: http://www.bamf.de/SharedDocs/Anlagen/WSB/DE/Downloads/broschuere-interkulturelle-oeffnung.pdf;jsessionid=463BBF4A73C255B04EB4D55543271F5C.1_cid286?__blob=publicationFile (accessed 16 March 2015).

Cleary TA (1968) Test bias: Prediction of grades of negro and white students in integrated colleges. *Journal of Educational Measurement* 5: 115–124.

Cohen J (1988) *Statistical Power Analysis for the Behavioral Sciences*. New York: Erlbaum.

Cook M (2009) *Personnel selection. Adding value through people* (5th edn). Chichester: Wiley–Blackwell.

De Meijer LAL, Born MP, Terlouw G, et al. (2006) Applicant and method factors related to ethnic score differences in personnel selection: A study at the Dutch police. *Human Performance* 19(3) 219–251.

De Soete B, Lievens F and Druart C (2012) An update on the diversity–validity dilemma in personnel selection: A review. *Psychological Topics* 21(3): 399–424.

Dean MA, Bobko P and Roth PL (2008) Ethnic and gender subgroup differences in assessment center ratings: A meta-analysis. *Journal of Applied Psychology* 93(3): 685–691.

Ellis PD (2009) Effect size equations. Available at: http://www.polyu.edu.hk/mm/effectsizefaqs/effect_size_equations2 (accessed 27 February 2015).

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice (1978) Uniform guidelines on employee selection procedures. 29 C.F.R. 1607. Available at: https://www.gpo.gov/fdsys/pkg/CFR-2009-title29-vol4/pdf/CFR-2009-title29-vol4-part1607.pdf (accessed 2 March 2016).

Finch DM, Wallace JC and Edwards BD (2009) Multistage selection strategies: Simulating the effects on adverse impact and expected performance for various predictor combinations. *Journal of Applied Psychology* 94(2): 318–340.

Foldes HJ, Duehr EE and Ones DS (2008) Group differences in personality: Meta-analyses comparing five US racial groups. *Personnel Psychology* 61: 579–616.

Galton F (1892) *Hereditary genius*. London: Macmillan.

Goldstein HW, Scherbaum CA and Yusko KP (2010) Revisiting g: Intelligence, adverse impact, and personnel selection. In: Outtz JL (ed.) *Adverse Impact. Implications for Organizational Staffing and High Stakes Selection*. New York: Routledge, 95–134.

Haug S (2008) *Sprachliche Integration von Migranten in Deutschland*. Forschungsgruppe des Bundesamtes für Migration und Flüchtlinge. Working Paper no.14. Reihe „Integrationsreport" Teil 2. Nürnberg: BAMF.

Hough LM, Oswald FL and Ployhart RE (2001) Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment* 9: 152–194.

Hossiep R and Mühlhaus O (2015) *Personalauswahl und -entwicklung mit Persönlichkeitstests* (2. Auflage). Göttingen: Hogrefe.

Kaas L and Manger C (2012) Ethnic discrimination in Germany's labour market: A field experiment. *German Economic Review* 13(1) 1–20.

Kehoe JF (2010) Cut scores and adverse impact. In: Outtz JL (ed.) *Adverse Impact. Implications for Organizational Staffing and High Stakes Selection*. New York: Routledge, 289–322.

Krause A, Rinne U, Zimmermann KF, et al. (2012) Pilotprojekt „Anonymisierte Bewerbungsverfahren". Abschlussbericht. IZA Research Report no. 44. Available at: http://www.iza.org/en/webcontent/publications/reports/report_pdfs/iza_report_44.pdf (accessed 06 March 2015).

Leenen WR, Groß A, Grosch H, et al. (2014a) *Kulturelle Diversität in der Öffentlichen Verwaltung. Konzeptionelle Grundsatzfragen, Strategien und praktische Lösungen am Beispiel der Polizei*. Münster: Waxmann.

Leenen WR, Stumpf S and Scheitza A (2014b) „Interkulturelle Kompetenz" in der Personalauswahl – Konzeptionalisierung und Integration in bestehende Auswahlsysteme. In: Barié-Wimmer F, von Helmolt K and Zimmermann B (eds) *Interkulturelle Arbeitskontexte. Beiträge zur empirischen Forschung*. Stuttgart: ibidem Verlag, 227–257.

Lindsey A, King E, McCausland T, et al. (2013) What we know and don't: Eradicating employment discrimination 50 years after the Civil Rights Act. *Industrial and Organizational Psychology: Perspectives on Science and Practice* 6(4): 391–413.

Melchers KG and Annen H (2010) Officer selection for the Swiss armed forces. An evaluation of validity and fairness issues. *Swiss Journal of Psychology* 69(2): 105–115.

Ministerium für Arbeit, Integration und Soziales des Landes Nordrhein-Westfalen (2013) *Landesinitiative Nordrhein-Westfalen. Mehr Migrantinnen und Migranten in den Öffentlichen Dienst – Interkulturelle Öffnung der Landesverwaltung*. Zweiter Umsetzungsbericht für den Zeitraum 31. Mai 2012 bis 30. Mai 2013. Available at: https://www.mais.nrw/sites/default/files/asset/document/130000_endfassung-zweiter-umsetzungsbericht_ikoe.pdf (accessed 2 March 2016).

Newman DA and Lyon JS (2009) Recruitment efforts to reduce adverse impact: Targeted recruiting for personality, cognitive ability, and diversity. *Journal of Applied Psychology* 94: 298–317.

Outtz JL (ed.) (2010) *Adverse Impact. Implications for Organizational Staffing and High Stakes Selection*. New York: Routledge.

Outtz JL and Newman DA (2010) A theory of adverse impact. In: Outtz JL (ed.) *Adverse Impact. Implications for Organizational Staffing and High Stakes Selection*. New York: Routledge, 53–94.

Ployhart RE and Holtz BC (2008) The diversity–validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology* 61: 153–172.

Pyburn KM Jr, Ployhart RE and Kravitz DA (2008) The diversity–validity dilemma: Overview and legal context. *Personnel Psychology* 61: 143–151.

Roth PL, Bevier CA, Bobko P, et al. (2001) Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology* 54: 297–330.

Roth PL, Van Iddekinge CH, Huffcutt AI, et al. (2002) Corrections for range restriction in structured interview ethnic group differences: The values may be larger than researchers thought. *Journal of Applied Psychology* 87: 369–376.

Sackett PR and Ellingson J (1997) The effects of forming multi-predictor composites on group differences and adverse impact. *Personnel Psychology* 50: 707–721.

Schmidt F and Hunter J (1998) The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin* 124(2): 262–274.

Schuler H, Hell B, Trapmann S, et al. (2007) Die Nutzung psychologischer Verfahren der externen Personalauswahl in deutschen Unternehmen. Ein Vergleich über 20 Jahre. *Zeitschrift für Personalpsychologie* 6(2): 60–70.

Sedlmeier P and Renkewitz F (2008) *Forschungsmethoden und Statistik in der Psychologie*. München: Pearson.

Statistisches Bundesamt (2013) Bevölkerung und Erwerbstätigkeit. Bevölkerung mit Migrationshintergrund, Ergebnisse des Mikrozensus 2013. Fachserie 1, Reihe 2.2. Available at: https://www.destatis.de/DE/Publikationen/Thematisch/Bevoelkerung/MigrationIntegration/Migrationshintergrund2010220137004.pdf?__blob=publicationFile (accessed 2 March 2016).

Thorndike EL (1921) *Educational psychology*. New York: Teacher's College of Columbia.

Urban D and Mayerl J (2011) *Regressionsanalyse: Theorie, Technik und Anwendung* (4. überarbeitete und erweiterte Auflage). Wiesbaden: Verlag für Sozialwissenschaften.

Zedeck S (2010) Adverse impact: History and evolution. In: Outtz JL (ed.) *Adverse Impact. Implications for Organizational Staffing and High Stakes Selection*. New York: Routledge, 3–27.

Adverse Impact in the personnel selection process of a German government authority: An analysis of ethnical subgroup differences

Siegfried Stumpf
Technische Hochschule Köln, Germany

Wolf Rainer Leenen
Technische Hochschule Köln, Germany

Alexander Scheitza
Kölner Institut für Interkulturelle Kompetenz, Germany

Abstract

While Anglo-American Human Resources (HR) research has been able to deliver highly differentiated results for adverse impact in personnel selection processes in recent years, there is an almost total lack of such research into differences in the performance outcomes of subgroups in the personnel selection procedures of German companies and organisations. In order to address this situation, the multi-stage personnel selection process of a large German government authority was analysed in terms of the success or failure of applicants with and without a migrant background. The results show that the cognitive ability test administered in the early selection stage and the orthographic test completed at the same time demonstrate substantial differences between subgroups; despite the preselection by means of the cognitive ability test and the orthographic test, differences between applicants with and without a migrant background are also found in the assessment centre conducted during a subsequent selection stage. Success in all phases of the selection process is influenced by the nature of a candidate's migration background: foreign nationals show the greatest divergence from applicants who do not have a migrant background. On the other hand, applicants who have been German citizens from birth, but who have a migrant mother and/or father, perform at almost the same level as those whose background does not involve migration. These findings are analysed against the backdrop of international research into adverse impact. In conclusion, possible measures are discussed for reducing the subgroup differences in personnel selection of organisations using the example of the government authority studied.

Keywords

Adverse impact, personnel selection, ethnic subgroup differences, migration