Kersting, M. (1995). Der Einsatz "westdeutscher" Testszur Personalauswahl in den neuen Bundesländern und die Fairneßfrage. Auswirkungen der Testleistungsdisparität zwischen Ost und West auf die Auswahlentscheidung. *Report Psychologie*, 20, 32-41.

Martin Kersting

DER EINSATZ „WESTDEUTSCHER" TESTS ZUR PERSONALAUSWAHL IN DEN NEUEN BUNDESLÄNDERN UND DIE FAIRNESSFRAGE

Auswirkungen der Testleistungsdisparität zwischen Ost und West auf die Auswahlentscheidung

Der Artikel diskutiert anhand von Erfahrungsberichten den Einsatz „westdeutscher" Tests bei der Personalauswahl in den neuen Bundesländern. Der dabei zutage tretende Befund der im innerdeutschen Vergleich niedrigeren Testleistungen ostdeutscher Bewerber(innen) ist Ausgangspunkt einer Modellrechnung, bei der die Testergebnisse exemplarisch in Auswahlentscheidungen umgesetzt werden. Dieser Ansatz erlaubt es, die praktische Relevanz der Testleistungsdisparität für die Personalauswahl unter verschiedenen Perspektiven der Fairneß zu beurteilen. Eine empirisch-technische Herangehensweise erscheint geboten, da die Diskussion um den Einsatz „westlicher" Tests bei Auswahlentscheidungen in den neuen Ländern im Alltag oft durch implizite Annahmen und Einstellungen zu Themen wie „Ost-West", „Intelligenz", „Tests", „Fairneß" und „Personalselektion" dominiert wird.

Seit dem Fall der Mauer finden „im Westen" entwickelte und an westdeutschen Stichproben geprüfte Tests auch in den neuen Bundesländern Anwendung und werden z. B. bei der Vergabe von Ausbildungs-/Studien- oder Arbeitsplätzen als Auswahlkriterium verwandt. Für diejenigen, die die Entscheidung treffen oder von ihr betroffen sind, stellt sich die Frage, ob mit dem Transfer der Tests über die ehemaligen innerdeutschen Grenzen hinweg nicht auch die Grenzen der Gültigkeit der aufgrund der Testergebnisse getroffenen Aussagen überschritten werden.

Mittlerweile liegen erste Analysen vor, in denen Daten aus „westlichen" Eignungstests auf Ost-West-Unterschiede hin untersucht wurden. Dabei zeigten sich zumeist geringfügig schlechtere Ergebnisse für die Personen aus den neuen Bundesländern.

Im Test für medizinische Studiengänge (TMS) übertrafen die 16761 westdeutschen Teilnehmer(innen) der 1990 durchgeführten Testung die 1688 Teilnehmer(innen) aus den neuen Ländern um durchschnittlich 3.8 Standardpunkte (Standardabweichung 9.75 West und 9.14 Ost). Die Testerhebung 1991 erbrachte folgende Ergebnisse: „Ost" (N = 3305): $\underline{M}$ = 97.98 ($\underline{SD}$ = 9.04); „West" (N = 18812): $\underline{M}$ = 101.16 ($\underline{SD}$ = 9.78). Im Jahre 1992 fiel die Ost-West-Differenz im TMS mit einem Betrag von 4.7 Standardpunkten zwischen den Mittelwerten von 101.27 ($\underline{SD}$ = 9.68) für die Westgruppe (N = 16772) und 96.60 ($\underline{SD}$ = 9.05) für die Ostgruppe (N = 2302) noch deutlicher aus. Die größten Leistungsdifferenzen zeigten sich 1992 bei den Aufgabengruppen „Textverständnis", „Diagramme und Tabellen" und „Medizinisch-naturwissenschaftliches Grundverständnis". (Quellenangaben: Trost[1], persönliche Mitteilung vom 11.8. 1993; Blum & Hensgen, 1993, 1994; Hensgen & Blum 1992.)

Melter[1] (persönliche Mitteilung vom 3.8.1993) vom Personalstammamt der Bundeswehr registrierte beim Ost-West-Vergleich der Testleistungen von 5177 Offiziersanwärtern, die im Zeitraum vom 1. Juli 1991 bis zum 30. Juni 1992 getestet wurden, eine durchschnittliche Überlegenheit der Westgruppe (N = 4125) in der Größenordnung von 3.6 Standardpunkten in einem allgemeinen Intelligenztest (Standardabweichung in beiden Gruppen annähernd 10) und in der Größenordnung von 1.7 Stan-

ardpunkten im Test der Konzentrationsleistung (Standardabweichung Ost: 14; West: 16.4). Im Test für Mathematik zeigten sich hingegen insgesamt keine Leistungsunterschiede zwischen den Teilgruppen.

Über Mittelwertsungleichheiten im WILDE-Intelligenztest zuungunsten von ca. 500 ostdeutschen Bewerber(inne)n für anspruchsvolle Führungsfunktionen im Vertriebsbereich berichtet Stratemann (1992), die diese Daten aus dem Jahre 1990 mit Daten gleich großer westlicher Probandengruppen verglich. Die bei Stratemann (1992) auf S. 71 abgebildete Graphik verzeichnet für alle Testdimensionen im Durchschnitt höhere Werte für die Westgruppe. Angaben über die exakten Leistungswerte, die Standardabweichungen und die Stichprobengröße fehlen. Laut Text konnte nur der Unterschied im rechnerischen Denken zufallskritisch abgesichert werden. Allerdings gingen lediglich die Daten von ca. 200 Personen in die statistische Analyse ein (persönliche Mitteilung Hübbe[1] vom 24.11.1993).

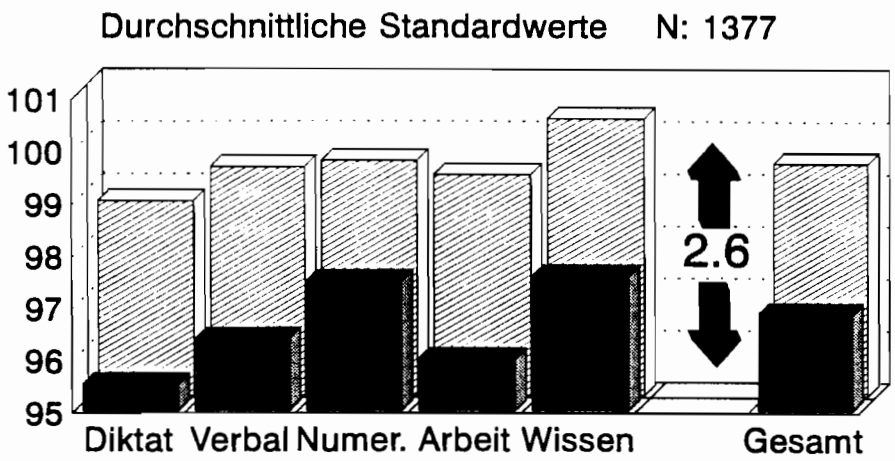
Bei einer 1991 durchgeführten psychologischen Eignungsuntersuchung von 285 Bewerber(inne)n für eine Ausbildung zum gehobenen Dienst der Bundesanstalt für Arbeit in Berlin und Brandenburg fiel der Leistungsvergleich zwischen Ost und West in Abhängigkeit von der jeweiligen Leistungsdimension unterschiedlich aus. Während die 106 Bewerber(innen) aus dem Westteil Berlins in den Tests zum induktiven und deduktiven Denken und beim Rechtschreibtest im Durchschnitt geringfügig bessere Werte erzielten als die 179 Personen aus dem Osten Berlins und aus Brandenburg, schnitt die zuletzt genannte Gruppe im Leistungsvergleich der Ergebnisse im „KLT" und im Test „Kürzen" (Eigenkonstruktion der Bundesanstalt für Arbeit) etwas besser ab. Der Leistungsunterschied betrug dabei jeweils um die 0.3-z-Wert-Punkte. (Persönliche Mitteilung Hustedt[1] vom 25.8.1994.)

Für die Informationen und für die Unterstützung möchte ich Herrn Dipl.-Psych. *E. Hübbe*, Herrn Dr. *Hustedt*, Herrn Dr. *Melter* und Herrn Dr. *Trost* herzlich danken.

Untersuchung

Grundlage der folgenden vergleichenden Analysen sind die Daten von insgesamt 1377 Personen, die sich anläßlich ihrer Bewerbung um eine Ausbildung zum gehobenen nichttechnischen Verwaltungsdienst im Zeitraum Juli 1991 bis Juni 1992 einem Eignungstest unterzogen haben. Die Eignungsuntersuchungen wurden von der Deutschen Gesellschaft für Personalwesen (DGP) durchgeführt. 696 Personen mit einem Durchschnittsalter von 20,6 Jahren hatten sich bei ostdeutschen Verwaltungen beworben. Die 681 Kandidat(inn)en für Verwaltungen der Alt-Bundesländer waren im Durchschnitt zum Testzeitpunkt 21,8 Jahre alt. Frauen waren mit einem Anteil von jeweils 56 % in beiden Gruppen etwas häufiger vertreten als Männer.

Die Bewerber(innen) durchliefen zunächst einen sogenannten „Vortest". Aufgrund der Vortestergebnisse wurde über die Zulassungsemptehlung zum sogenannten „Haupttest" entschieden. Alle zugelassenen Bewerber(innen) mußten beim später gelegenen Haupttesttermin noch einmal eine – diesmal umfangreichere – Reihe von schriftlichen Tests bewältigen. Im Haupttest fand aufgrund von Beobachtungen der Bewerber(innen) bei situativen Aufgaben (Gruppen- und Einzelgesprächen) außerdem eine Verhaltensbeurteilung statt, die im vorliegenden Artikel aber keine Berücksichtigung findet. Die folgenden Ausführungen beschränken sich auf die hinsichtlich Leistungstests unausgelesene Gruppe des Vortests.



	Diktat	Verbal	Numer.	Arbeit	Wissen		Gesamt
West (N: 681)	98.74	99.39	99.51	99.24	100.31		99.46
Ost (N: 696)	95.52	96.39	97.5	96	97.6		96.87
F-Wert	30.5	54.3	25.3	35	28		66.9
Std.Abw. Ost	11.37	7.48	7.42	9.97	9.76		5.94
Std.Abw. West	10.39	7.66	7.59	10.37	9.85		5.84
Anzahl Tests	1	2	4	1	1		9

Abb. 1: Signifikante Ost-West-Leistungsunterschiede im Vortest

Das Balkendiagramm in Abbildung 1 verzeichnet die durchschnittlich in Ost (schwarz) und West (schraffiert) im Vortest erzielten Leistungen. Abgetragen sind die Werte für das Diktat, für die Tests zum gedanklichen Umgang mit sprachlichem („Verbal", zwei Tests) und mit numerischem Aufgabenmaterial („numer.", vier Tests), für einen Test zur Geschwindigkeit und Sorgfalt des Arbeitsverhaltens („Arbeit") und für die Überprüfung gemeinschaftskundlicher Kenntnisse („Wissen") sowie für ein aggregiertes Maß („Gesamt"). Bei dem angewandten Test handelt es sich um eine unveröffentlichte Eigenkonstruktion der DGP. Mit Ausnahme des Lückendiktats, des Kenntnistests und der tendenziell nach Art einer Arbeitsprobe gestalteten Überprüfung des Arbeitsverhaltens entsprechen die übrigen Aufgaben vom Typ her üblichen Intelligenztestaufgaben (z. B. wird die Weiterführung von sprachlichen Analogien oder Zahlenprogressionen verlangt). Beispielitems für alle Tests können beim Autor angefordert werden.

In allen getesteten Dimensionen ergaben sich höhere Werte für die Westgruppe. Die deutlichsten Unterschiede zeigten sich im sprachlichen Denken, beim Diktat und im Arbeitsverhalten, die kleinste Differenz war für den Bereich der Verarbeitung komplexer zahlengebundener Informationen zu verzeichnen. Der in der Graphik gewählte Maßstab ist eine starke Ausschnittsvergrößerung der von 70–130 variierenden Leistungsskala. Im Durchschnitt unterschieden sich die beiden Gruppen um 2.6 Standardpunkte. Alle dargestellten Unterschiede sind statistisch hochsignifikant, allerdings erklärt (bindet) die Variable „Kulturzugehörigkeit" („Ost" oder „West") nur 4.6 % der Varianz. Auch innerhalb der West- und der Ostgruppe konnten für Subgruppen Leistungsdifferenzen festgestellt werden. So erzielten z. B. in der Westgruppe die Teilnehmer(innen) mit Fachhochschulreife im Durchschnitt ein schlechteres Testergebnis (Aggregat der Einzelaufgaben) als die Teilnehmer(innen), die über die allgemeine Hochschulreife verfügten. Innerhalb der Ostgruppe erzielten die Absolvent(inn)en der Erweiterten Oberschule (EOS) die höchsten Testleistungen. Die verfügbaren demographischen und edukativen Gruppierungsdaten konnten aber keinen hinreichenden Beitrag zur Aufklärung der Unterschiede in den Testergebnissen auf der Ost-West-Vergleichsebene leisten. Auf dieser Ebene blieb es z. B. auch dann noch beim Befund der Testleistungsunterschiede, wenn man die Analysen auf die Subgruppen der relativ gesehen testleistungsstarken EOS-Absolvent(inn)en (N=506) im Osten und die relativ testleistungsschwachen West-Bewerber(innen) mit Fachhochschulreife (N=286) begrenzte. ($\underline{M}$ West=98.68 gegenüber $\underline{M}$ Ost=97.28, $\underline{F}(3,788)=5.63$; $\underline{p}<.05$.) Vorbildungseffekte lassen sich gleichwohl nicht ausschließen, da sich aus den vorliegenden Daten von 1991 und 1992 natürlich keine Subgruppen von Testteilnehmer(inne)n bilden lassen, die über formal (Dauer der Schulzeit) und inhaltlich vergleichbare Schulausbildungen verfügen.

Interpretationsmöglichkeiten

Grundsätzlich kann man den Befund der innerdeutschen Leistungsdifferenzen im Eignungstest als Indikator für Unterschiede in der Leistungsfähigkeit oder als Indikator für die Unangemessenheit der verwendeten diagnostischen Verfahren interpretieren. Keine Interpretationsrichtung kann *a priori* – ohne theoretische Begründungen und empirische Bestätigungen – Gültigkeit für sich beanspruchen. Zugunsten der folgenden Ausführungen über die *praktische* Bedeutung der aufgezeigten Befunde für die Personalauswahl wird hier auf eine ausführliche Interpretation/Erklärung der Befunde verzichtet und lediglich ein Interpretationsraum skizziert. Weitergehende Überlegungen zur Interpretation sowie erste empirische Analysen zu spezifischen Hypothesen über mögliche Ursachen der Leistungsunterschiede finden sich bei Blum und Hensgen (1994) sowie bei Kersting (1994).

Zur Argumentation der „unangemessenen diagnostischen Verfahren" zählt der Hinweis auf den möglicherweise unterschiedlich ausgeprägten Umfang der Testerfahrung in den beiden Teilgruppen (siehe z. B. Ettrich & Guthke, 1991, S. 18) und die damit verbundene unterschiedliche Vertrautheit mit der Testsituation (Zeitdruck, Antwortformat usw.). Die oben dargestellten Differenzen in den Testleistungen sind von der Größenordnung her mit den Effekten von Testtrainingsmaßnahmen auf die Testergebnisse vergleichbar (siehe z. B. Kulik, Bangert-Drowns & Kulik, 1984). Im Sinne der kulturellen Testverzerrung kann man auch argumentieren, daß die Tests (insbesondere Kenntnistests, Sprachtests) eventuell kulturspezifische Eigenheiten (z. B. Schulcurricula, Sprachgebrauch) der „westdeutschen Kultur" auf die Population der Ostdeutschen übertragen. Im Rahmen dieser Argumentationsrichtung wird man die Aufmerksamkeit schließlich auch auf nicht-kognitive Variablen (z. B. Motivation, Testängstlichkeit) richten. In Ost und West unterschiedliche Testleistungen könnten u. a. durch möglicherweise unterschiedliche Ausprägungen in solchen nichtkognitiven Variablen bedingt sein.

Will man hingegen die Auffassung begründen, daß die Differenzen im Testergebnis

der beiden Gruppen tatsächliche Leistungsunterschiede widerspiegeln, kann man z.B. argumentieren, daß die beiden Stichproben womöglich jeweils unterschiedliche Grundgesamtheiten repräsentieren. Aus Gründen der (1.) Binnenwanderungsprozesse (siehe Maretzke & Möller, 1992), (2.) der ungleichen Verteilung der formellen Bewerbungsvoraussetzungen und (3.) aus Gründen der unterschiedlich gelungenen Selbstselektion (siehe z.B. Stratemann, 1992) könnten sich hüben und drüben *jeweils anders zusammengesetzte Personengruppen* um Stellen oder Ausbildungen bewerben. Dies würde auch die zum Teil anderslautenden Befunde zu Ost-West-Leistungsunterschieden bei solchen Untersuchungen erklären, die mit weniger vorausgewählten Stichproben arbeiten. So zeigten sich bei Strohschneider (1994) z.B. keine innerdeutschen Unterschiede hinsichtlich der allgemeinen intellektuellen Leistungsfähigkeit von Berufsschüler(innen) aus beiden Teilen Berlins. Baumert (1994) berichtet über eine Untersuchung, in der u.a. mit Hilfe standardisierter Tests die Schulleistungen von ost- und westdeutschen Schüler(inne)n der 7. Jahrgangsstufe des Schuljahres 1991/92 verglichen wurden. Der Vergleich fiel insbesondere in den Fächern Biologie und Physik im Mittel zugunsten der Schüler(innen) aus den neuen Bundesländern aus.

Die Fairneß von Entscheidungen über Angehörige testleistungsdisparater Gruppen

Die Testleistungsdiversität der beiden deutschen Gruppen führt – ungeachtet ihrer Ursachen – dazu, daß der relative Anteil Ausgewählter aus der Gruppe der ostdeutschen Bewerber(innen) vergleichsweise geringer ist als der entsprechende Anteil aus der Westgruppe.

Dieser Befund könnte die auf den Bereich der beruflichen Auswahlentscheidungen bezogene Diskussion der Fairneß-Frage beleben, welche laut Schuler und Funke (1989, S. 318) im deutschen Sprachraum bisher kaum stattgefunden hat. Deutschsprachige Arbeiten zum Thema „Testfairneß" waren vor allem auf konzeptuell-methodischer Ebene angesiedelt (z.B. Gösslbauer, 1977; Simons & Möbus, 1978; Möbus, 1978, 1983; Bartussek, 1982). Einen praktischen Bezug bekam das Thema in Deutschland im Bereich der (Hoch-)Schule (z.B. Simons & Möbus, 1976; Wottawa &

Amelang, 1980; Trost, 1985). Die Gruppen wurden zumeist anhand des Kriteriums „Geschlecht" oder „Schichtzugehörigkeit" definiert.

In den oben angeführten deutschsprachigen konzeptionellen Arbeiten zur Fairneßfrage sowie in zahlreichen englischsprachigen Artikeln sind die existierenden Fairneßmodelle, ihr Verhältnis zueinander und die jeweils abzuleitenden unterschiedlichen Forderungen an ein „faires" Auswahlverfahren übersichtlich und verständlich beschrieben. Im wesentlichen läßt sich die Unterschiedlichkeit der Forderungen auf die Unterschiedlichkeit der *Perspektiven* zurückführen, mit der *ein und dieselbe Sache* betrachtet wird. Soll ein Verfahren „*fair*" sein aus der Sicht der Bewerber(innen) und, wenn ja, (1.) aus Sicht des einzelnen, (2.) der Gesamtgruppe oder (3.) aus der Perspektive einer Untergruppe innerhalb der Bewerber(innen), also z. B. der Ostdeutschen? Soll es „*fair*" *sein* (4.) aus Sicht der Institution oder (5.) aus Sicht des Gemeinwohls? (Perspektiven nach Gösslbauer, 1977, S. 100.) Was sich aus der einen Perspektive als „*fair*" erweist, kann unter einer anderen Perspektive betrachtet „*unfair*" sein. Zum Teil resultieren sogar aus einem Perspektivwechsel *innerhalb* eines Fairneßmodells, also indem man z. B. die für die potentiell Erfolgreichen formulierte Definition aus Sicht der potentiell Nichterfolgreichen formuliert, gänzlich andere, zu den ursprünglichen Forderungen womöglich inkompatible Ansprüche an ein „faires" Verfahren.

Während die Wissensgrundlagen zum Thema Test- und Selektionsfairneß hinreichend ausgearbeitet sind, ermangelt es der Fairneßauseinandersetzung in der Praxis oft einer Quantifizierung des konkreten Problems. Die Fairneßdiskussion läuft Gefahr, „Bodenhaftung" zu verlieren, da sie kaum handlungsrelevante Schlußfolgerungen für den konkreten Fall bereithält. Der vorliegende Artikel versucht deshalb, über die theoretische Diskussion und die Beschreibung der Größenordnung der Testleistungsdifferenzen hinauszugehen, indem mit Hilfe einer Modellrechnung die praktische Bedeutung der Testleistungsdifferenzen für die Fairneß der Auswahlentscheidung veranschaulicht wird. Die Generalisierbarkeit der Modellrechnung wird dabei durch die Setzung einiger Parameter beschränkt. Die Explikation des Vorgehens erlaubt es aber, für eigene Anwendungsfälle andere Parametersetzungen vorzunehmen.

Modellrechnung: Annahmen und Parameter

Die Modellrechnung basiert auf einigen vereinfachenden Annahmen, die bei der Interpretation berücksichtigt werden müssen. Es wird davon ausgegangen, daß

1. Eignungsurteil und Ausleseentscheidung zusammenfallen;
2. das Eignungsurteil allein aufgrund der Testergebnisse getroffen wird;
3. in beiden Gruppen derselbe „cutoff" im Test für das Eignungsurteil gesetzt wird;
4. die Kriteriumsvalidität der Tests in beiden Gruppen mit .54 gleich hoch ist.

Die Annahmen 1–3 lassen sich durch die Fragestellung begründen. Abgeschätzt werden soll der Effekt auf die Auswahlentscheidung, der von den Leistungsdifferenzen der beiden deutschen Gruppen im *Test* ausgeht. Dazu ist es notwendig, andere Einflüsse – wie Verhaltensbeurteilungen, eventuelle unterschiedliche „Cutoff"-Setzungen in den Gruppen oder weitere Aspekte, die in der Realität eine Nichtübereinstimmung von Eignungsurteil und Ausleseentscheidung begründen können –, aus der Modellrechnung auszuschließen.

Die Setzung der Kriteriumsvalidität bedarf der Erläuterung. Zu der grundlegenden Frage, ob die Kriteriumsvalidität der Tests in Ostdeutschland gleich hoch ausfällt wie in Westdeutschland oder ob man von singulären oder differenzierten Werten der Kriteriumsvalidität ausgehen muß, liegen nach Kenntnis des Autors zur Zeit keine empirischen Daten vor. Diesbezüglich lassen sich daher gegenwärtig – faute de mieux – nur Erwartungen formulieren, die sich allerdings an den Ergebnissen empirischer Untersuchungen in anderen Kulturen orientieren können. Dabei konnte für Gruppen, die sowohl kulturell als auch von den Testergebnissen her deutlich disparater waren als „Ost-" und „Westdeutsche", immer wieder festgestellt werden, daß es *keine* nennenswerten gruppenspezifischen Vorhersageunterschiede gab (z. B. Hunter & Hunter, 1984; Hunter, Schmidt & Hunter, 1979; Schmidt, Berner & Hunter, 1973). Diese Literaturbefunde begründen die hier gesetzte Prämisse der in beiden deutschen Gruppen übereinstimmenden Kriteriumsvaliditäten des Eignungstests. Die Annahme ist weitreichend, da sie etwas voraussetzt, was normalerweise der Forschungsgegenstand von Untersuchungen ist. Zur empirischen Untersuchung von Fairneßfragen wird nach Bartussek (1982) am häufigsten das „*(Regressions)Modell der fairen Vorhersage*" nach Cleary angewandt. Ein Auswahlverfahren ist diesem Modell zufolge „*dann fair, wenn das dafür verwendete Vorhersageinstrument (Test) für das Kriterium in keiner der beiden zu vergleichenden Gruppen eine systematische Über- und Unterschätzung ihrer Kriteriumswerte erbringt*" (ebd. S. 3). Die hier getroffene Annahme gleicher Kriteriumsvaliditäten unterstellt, daß die Basis einer fairen Auswahl im Sinne dieser Definition gesichert ist, insofern in beiden Gruppen die *voraussichtlich* besten Bewerber(innen) bevorzugt werden, wenn man in Ost und West der Auswahl denselben „Testcutoff" zugrunde legt. Thematisiert wird im folgenden lediglich die Frage, welche Probleme selbst unter der, aufgrund der Literaturlage zu erwartenden, Gültigkeit dieser Prämisse noch bestehen.

Der eingesetzte Wert von .54 für die Kriteriumsvalidität des Tests in beiden Gruppen geht – in Ermangelung diesbezüglicher empirischer Werte für die neuen Bundesländer – ebenfalls auf die Fachliteratur zurück. Dieser Wert wurde in der Meta-Analyse von Hunter und Hunter (1984, S. 81) für den Durchschnitt über alle Berufe als Validitätswert von kognitiven Fähigkeitstests für die Vorhersage des Ausbildungserfolgs bestimmt.

Empirischer Ausgangspunkt der Modellechnung ist die Quote der aufgrund der Testrgebnisse ausgesprochenen Zulassungsempehlungen in der oben beschriebenen Stichrobe. Gemäß der Annahme 1 einer Identität on Eignungsurteil und Ausleseentscheidung ird diese Quote als Selektionsquote aufgeißt. Legt man für beide Gruppen den identichen Anforderungsmaßstab zugrunde (welher hier aus Platzgründen nicht näher bechrieben werden kann), so erzielten nach Abolvierung des Haupttests 153 der ursprüngch 696 „Ost-"Bewerber(innen), also rund 2 %, und 213 der 681 „West-" Bewerber(innen) (rund 31 %) eine Zulassungsempfehlung ufgrund des Testverfahrens. Den Anteil der weils „valide" oder „falsch" vorhergesagten ugelassenen oder Abgelehnten erhält man, idem man den korrelativen Wert der Kriteiumsvalidität mit der von Rosenthal (1984) orgestellten Formel zur binomialen Effektrößendarstellung in ein 4-Felder-Schema ansformiert.

Modellrechnung: Konsequenzen ür die Auswahlentscheidung

Abbildung 2 stellt die (gerundeten) Erebnisse der Modellrechnung dar. Abgetraen ist zunächst, wie sich je 100 Bewerber(innen) der alten (Tabelle 1) und der neuen Bunesländer (Tabelle 2) auf die vier möglichen rgebniskategorien der Vorhersage verteilen valide [richtig] positiv, falsch positiv, valide egativ und falsch negativ).

Am Ende der Abbildung werden tabellaisch verschiedene Indikatoren für die Fairneß er Auswahlentscheidung miteinander verlichen. Da für beide Gruppen derselbe Korelationskoeffizient zugrundegelegt wurde, reultiert eine gleich hohe Anzahl an korrekten ntscheidungen (Anzahl der korrekt vorheresagten erfolgreichen und der korrekt vorergesagten nicht-erfolgreichen Personen). Mit 77 % ist auch der prozentuale Anteil der rfolgreichen Personen an allen ausgewählten ersonen (valide Positive /Selektionsrate) für eide Gruppen gleich. Diese Gleichheit der Trefferquote" entspricht einem fairen Auswahlverfahren nach dem Fairneßmodell der *gleichen Wahrscheinlichkeiten*" von Linn. Das oben angesprochene Modell von Cleary tellt einen Spezialfall dieses Gleichwahrcheinlichkeitsmodells dar (siehe Möbus, 983, S. 183).

Dabei werden bei der Bewertung der airneß der Auswahlentscheidung die Effekte der Ablehnung ignoriert. Dies entspricht den Interessen der einstellenden Institution. Personen, die nicht ausgewählt wurden, spielen für die Leistungsfähigkeit der Organisation keine Rolle. Der Nutzen für die Institution entsteht durch die erfolgreichen Personen, die eine Zulassung erzielt haben; der Schaden erwächst aus der Zulassung nicht-erfolgreicher Personen. Abgelehnte Personen bringen der Institution keinen Nutzen, und die Ablehnung selbst fügt ihr – außer den Verfahrenskosten – so lange keinen Schaden zu, wie die geringere

Beispiel aufgrund des an einer Stichprobe von 1377 Personen ermittelten Verhältnisses zwischen abgelehnten und angenommenen Bewerber(innen) einerseits und eines (nach Hunter & Hunter, 1984) auf .54 geschätzten Zusammenhangs zwischen Test und Ausbildungserfolg in beiden Gruppen andererseits.

1. West (von je 100)

		Test	
		abgelehnt (69)	angenommen (31)
Kriterium	Erfolg	16	24
	Mißerfolg	53	7

2. Ost (von je 100)

		Test	
		abgelehnt (78)	angenommen (22)
Kriterium	Erfolg	18	17
	Mißerfolg	60	5

3. Vergleich	**West**	**Ost**
Basisrate	40 %	35 %
Selektionsrate	31 %	22 %
Proportion korrekter Entscheidungen	77 %	77 %
Anteil der validen Positiven an allen Selektierten (Modell gleicher Wahrscheinlichkeiten)	77 %	77 %
Anteil der Selektierten an allen potentiell Erfolgreichen (Modell konstanter Verhältnisse)	77 %	63 %
Anteil der validen Positiven an allen potentiell Erfolgreichen (Modell bedingter Wahrscheinlichkeiten)	60 %	49 %

Abb. 2: Modellrechnung: Unterschiedliche Häufigkeiten der Vorhersagefehlertypen bei identischer Kriteriumsvalidität der in Ost und West eingesetzten Testverfahren

Selektionsquote durch eine Erhöhung der Bewerber(innen)zahlen kompensiert werden kann. Unter der Maßgabe einer Maximierung der „Treffer" bei den Akzeptierten wird der im Vergleich der Testleistungen ungünstigeren Basisrate im Osten durch eine strengere oder konservativere Auswahl entgegengesteuert. Dies hat aber nicht nur einen – im Vergleich zum Westen – höheren Bedarf an Kandidat(inn)en zur Konsequenz. Infolge des unterschiedlichen Anteils der Zugelassenen bei Ost und West sind die beiden Fehlertypen der Vorhersage im Gruppenvergleich unterschiedlich häufig besetzt. In der letzten Zeile der Tabelle ist der Wert für den prozentualen Anteil der ausgewählten Positiven an allen potentiell Erfolgreichen berichtet (valide Positive / Basisrate): Während in den alten Bundesländern dieser Modellrechnung zufolge 60 % aller potentiell erfolgreichen Bewerber(innen) im Auswahlverfahren „erkannt" werden, sind es in den neuen Bundesländern nur 49 %. Für die westdeutschen Bewerber(innen) besteht eine – im deutsch-deutschen Vergleich – erhöhte Gefahr, *über*schätzt zu werden (Zusage und späterer Mißerfolg). Demgegenüber droht den ostdeutschen Bewerber(inne)n etwas häufiger das „Schicksal", vergleichsweise *unter*schätzt zu werden (Absage trotz potentiellem Erfolg). Der Modellrechnung zufolge werden mehr potentiell Erfolgreiche Ostbewerber(innen) zurückgewiesen als angenommen. Diese ungleiche Häufigkeit der Fehlertypen ist der Grund, warum nach manchen Fairneßmodellen – beispielsweise nach dem „*bedingten Wahrscheinlichkeitsmodell*" von Cole (1973, zitiert nach Möbus, 1983) – ein Test auch dann als „*unfair*" angesehen werden kann, wenn für beide Gruppen der Nachweis gleich guter Trefferquoten in der Vorhersage erbracht wird. „*Fair*" ist ein Verfahren dieser Auffassung zufolge nur dann, wenn für die *potentiell Erfolgreichen* (Basisrate) in den beiden Gruppen gleich hohe Wahrscheinlichkeiten der Auswahl bestehen. Die Modellrechnung erlaubt es nun, den Effekt zu quantifizieren, den die gefundene Ost-West-Leistungsdisparität in Tests mit – vermutlich – identischer Kriteriumsvalidität auf die Auswahlentscheidung zeitigt: Ausgehend von jeweils 100 Kandidat(inne)n werden im Westen 40 % von 40, das sind *16*, im Osten 51 % von 35 potentiell erfolgreichen Bewerber(inne)n, das sind rund *18* Personen, *nicht* erkannt (grau hinterlegte Felder der Tabellen). Dieses Ergebnis wird neben der empirisch ermittelten Selektionsquote von der geschätzten Höhe der Kriteriumsvalidität determiniert. Setzt man für die Kriteriumsvalidität statt .54 mit .32 den Wert ein, der in der Meta-Analyse von Hunter und Hunter (1984) am unteren Ende der Verteilung der entsprechenden Werte steht, so steigt die Zahl der gegenüber dem Westen im Osten aufgrund der „konservativen" Vorhersage zusätzlich „verkannten" Bewerber(innen) der Modellrechnung zufolge von zwei auf drei.

Ausblick / Konsequenzen

Bei der Beurteilung der in Ost und West unterschiedlichen Häufigkeit der einzelnen Fehlertypen aus Sicht der *kulturellen* Testfairneß gilt es zu bedenken, daß es ohne fehlerfreie Messungen keine unter allen Perspektiven vollständig „fairen" testbasierten Entscheidungen bei leistungsdisparaten Gruppen geben kann. Solange die Messung fehlerbehaftet ist, ist die Frage nach der Fairneß eine Frage danach, *wen* dieser Fehler betrifft. Darüber hinaus ist festzuhalten, daß in der Praxis anstelle der abgewiesenen Personen in der überwiegenden Zahl der Fälle (zumindest im hier gewählten Beispiel der Ausbildung im öffentlichen Dienst) Personen aus der *gleichen kulturellen Gruppe* zum Zuge kommen. Eine direkte Konkurrenz zwischen Ost- und Westbewerber(inne)n dürfte realiter eher die Ausnahme (z.B. bei einer bundeseinheitlichen Vergabepraxis) denn die Regel darstellen. Innerhalb dieser Ausnahmefälle sind Situationen denkbar, die – *außerhalb des diagnostischen Begründungszusammenhangs* – Anlaß dazu geben, eine der Gruppen *leistungsunabhängig* zu bevorzugen. Auf den Bericht von Gruppenleistungsunterschieden wird häufig

mit der Forderung nach „Korrekturmaßnahmen" reagiert (siehe z.B. Wigdor & Sackett, 1993). Denkbar ist z.B. die Einführung von spezifischen Normen und/oder Anforderungen für bestimmte Personenkreise bzw. die Einführung von Bonus/Malus-Systemen. Damit wird die Unfairneß nicht aufgehoben, sondern „verschoben", indem sich die Fairneß aus Sicht des einzelnen – unabhängig von der Gruppenzugehörigkeit – verringert. Außerdem muß in vielen Fällen eine Verschlechterung der Kriteriumsvalidität in Kauf genommen werden. Maxwell und Arwey (1993) konnten mit Bezug auf die Grundgesamtheiten zeigen, daß die validesten Testverfahren methodisch zwingend auch die fairsten Verfahren im Sinne der Cleary Definition sind. Jegliche Form der gruppenspezifischen Test- aus- oder -bewertung wirft schließlich ausufernde Zuordnungsprobleme auf. Zahlreiche Personen könnten mit der gleichen Begründung einen eigenen Gruppenstatus einfordern (z.B. Gruppenbildungen aufgrund geschlechts-, bildungs-, schicht- oder regionalspezifischer Merkmale). Auch der oft propagierte Ersatz der herkömmlichen westdeutschen Tests durch „neue" Testverfahren ist problematisch, da für diese Verfahren die notwendigen Kennwerte und Erfahrungen fehlen.

Die Modellrechnung sollte es den Leser(inne)n erleichtern, sich ein eigenes Urteil über die Angemessenheit des Einsatzes „westdeutscher" Leistungstests zur Unterstützung von Auswahlentscheidungen über Neubundesbürger(innen) zu bilden. Dieses Urteil wird von der jeweiligen Perspektive auf den Sachverhalt beeinflußt. Dem Autor erscheint ein Einsatz „westdeutscher" Leistungstests zur Unterstützung von Auswahlentscheidungen in den neuen Bundesländern unter folgenden Bedingungen gerechtfertigt: Der Testeinsatz muß (1.) durch eine Kontrolle der gruppenspezifischen Testleistungen begleitet werden. Sofern sich gruppenspezifische Leistungsdifferenzen zeigen, ist (2.) der Effekt dieser Nichtübereinstimmung auf die Auswahlentscheidung zu explizieren. Außerdem sollte (3.) eine Erklärung der Leistungsunterschiede geleistet werden, indem (3.1) eine Suche nach Indikatoren einer kulturabhängigen Testverzerrung bzw. (3.2) nach testunabhängigen Ursachen eingeleitet wird. Gegebenenfalls sind (4.) kontrollierte Testmodifikationen notwendig, wobei grundsätzlich (5.) zukünftig die Neubundesbürger(innen) bereits bei der Test*konstruktion* zu berücksichtigen sind. Insgesamt ist (6.) der Nachweis der Berechtigung des Geltungsanspruchs des Tests für die „neue" Population und insbesondere der Nachweis der Kriteriumsvalidität zu erbringen.

Zusammenfassung

Der vorliegende Beitrag berichtet anhand eines innerdeutschen Vergleichs der Testergebnisse von 1377 Bewerber(inne)n über Erfahrungen mit dem Einsatz von „westlichen" Tests zur Eignungsdiagnostik in den neuen Bundesländern. In diesem Vergleich zeigten sich geringfügig höhere Testleistungen der westdeutschen Bewerber(innen). In einer Modellrechnung wird die Bedeutung dieser Ost-West-Testleistungsunterschiede für die Fairneß der Auswahlentscheidung quantifiziert. Die Annahme einer für Testteilnehmer(innen) in beiden Teilen Deutschlands gleich hohen Kriteriumsvalidität sichert zwar eine identische „Trefferquote" der Vorhersage in bezug auf die jeweils *ausgewählten* Personen. Potentiell geeignete Bewerber(innen) aus den neuen Ländern werden im testbasierten Auswahlverfahren aber etwas häufiger „verkannt" als Bewerber(innen) aus den Altbundesländern. Der Modellrechnung zufolge gibt es im Osten 18% und im Westen 16% „falsch Negative". Die Befunde werden unter dem Aspekt der „Fairneß" diskutiert.

Martin Kersting, Dipl.-Psych., geb. 1964. Studium am Institut für Psychologie der Freien Universität Berlin. Zusätzliches Grundstudium der Germanistik. Bis 1992 Mitglied der Forschungsgruppe um Prof. Dr. A. O. Jäger. Seit 1992 Angestellter der Deutschen Gesellschaft für Personalwesen. Bevorzugte Arbeitsbereiche: Diagnostik, Arbeits- und Organisationspsychologie, Differentielle Psychologie. Themenschwerpunkte: Intelligenz, Wissen und Problemlösen.

Anschrift: Deutsche Gesellschaft für Personalwesen, Grassistraße 12, 04107 Leipzig.

BARTUSSEK, D. (1982). *Modelle der Testfairneß und Selektionsfairneß.* (Trierer Psychologische Berichte). Trier: Universität Trier.

BAUMERT, J. (1994). Bildungsvorstellungen, Schulleistungen und selbstbezogene Kognitionen in Ost- und Westdeutschland. In: D. Benner & D. Lenzen (Hrsg.), *Bildung und Erziehung in Europa. Beiträge zum 14. Kongreß der Deutschen Gesellschaft für Erziehungswissenschaft* (S. 272–276). Weinheim: Beltz.

BLUM, F. & HENSGEN, A. (1993). Zahlenmäßige Anteile, Test- und Schulleistungen einzelner Gruppen von Testteilnehmern. In: G. Trost (Hrsg.), *Test für medizinische Studiengänge (TMS): Studien zur Evaluation. 17. Arbeitsbericht* (S. 22–58). Bonn: Institut für Test- und Begabungsforschung.

BLUM, F. & HENSGEN, A. (1994). Vergleichende Analysen der deutschen Teilnehmergruppen aus den alten und den neuen Bundesländern sowie der ausländischen Testbearbeiter. In: G. Trost (Hrsg.), *Test für medizinische Studiengänge (TMS): Studien zur Evaluation. 18. Arbeitsbericht* (S. 54–116). Bonn: Institut für Test- und Begabungsforschung.

ETTRICH, K.U. & GUTHKE, J. (1991). Pädagogisch-psychologische Diagnostik in der DDR – Ein Rück- und Überblick aus gegebenem Anlaß. In: K. Ingenkamp & R.S. Jäger (Hrsg.), *Tests und Trends 9* (S. 13–42). Weinheim: Beltz.

GÖSSLBAUER, J.P. (1977). Tests als Selektionsinstrumente – fair oder unfair? *Psychologie und Praxis, 21*, 97–111.

HENSGEN, A. & BLUM, F. (1992). Vergleich einzelner Teilnehmergruppen beim sechsten Termin des besonderen Auswahlverfahrens: zahlenmäßige Anteile, Test- und Schulleistungen. In: G. Trost (Hrsg.), *Test für medizinische Studiengänge (TMS): Studien zur Evaluation. 16. Arbeitsbericht* (S. 22–95). Bonn: Institut für Test- und Begabungsforschung.

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

HUNTER, J.E., SCHMIDT, F.L. & HUNTER, R. (1979). Differential validity of employment tests by race: a comprehensive review and analysis. *Psychological Bulletin, 86*, 721–735.

KERSTING, M. (1994). *Personalauswahl in den neuen Bundesländern: Tests ohne Grenzen oder Grenzen der Tests? Eine vergleichende Analyse der durchschnittlichen Testergebnisse in Ost und West (Teil II)* (DGP-Informationen). Hannover: Deutsche Gesellschaft für Personalwesen, *53*, 47–98.

KULIK, J.A., BANGERT-DROWNS, R.L. & KULIK, C.L. (1984). Effectiveness of coaching for aptitude tests. *Psychological Bulletin, 95*, 179–188.

MARETZKE, S. & MÖLLER, F.O. (1992). Binnenwanderungsprozesse in Deutschland 1991. *Mitteilungen der Bundesforschungsanstalt für Landeskunde und Raumordnung*, 6–7.

MAXWELL, S.E. & ARVEY, R.D. (1993). The search for predictors with high validity and low Adverse Impact: Compatible or incompatible goals? *Journal of Applied Psychology, 78*, 433–437.

MÖBUS, C. (1978). Zur Fairness psychologischer Intelligenztests: Ein unlösbares Trilemma zwischen den Zielen von Gruppen, Individuen und Institutionen? *Diagnostica, 24*, 191–234.

MÖBUS, C. (1983). Zur praktischen Bedeutung der Testfairneß als zusätzliches Kriterium zu Reliabilität und Validität. In: R. Horn, K. Ingenkamp & R.S. Jäger (Hrsg.), *Tests und Trends 3* (S. 155–203). Weinheim: Beltz.

ROSENTHAL, R. (1984). *Meta-analytic procedures for social research.* Beverly Hills: Sage.

SCHMIDT, F.L., BERNER, J.G. & HUNTER, J.E. (1973). Racial differences in validity of employment tests: reality or illusion? *Journal of Applied Psychology, 58*, 5–9.

SCHULER, H. & FUNKE, U. (1989). Berufseignungsdiagnostik. In: E. Roth (Hrsg.), *Enzyklopädie der Psychologie. Wirtschafts-, Organisations- und Arbeitspsychologie. Band 3* (S. 281–320). Göttingen: Hogrefe.

SIMONS, H. & MÖBUS, C. (1976). Untersuchungen zur Fairneß von Intelligenztests. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 8*, 1–12.

SIMONS, H. & MÖBUS, C. (1978). Testfairneß. In: K.J. Klauer (Hrsg.), *Handbuch der pädagogischen Diagnostik* (Bd. 1, S. 187–197). Düsseldorf: Schwann.

STRATEMANN, I. (1992). *Psychologische Aspekte des wirtschaftlichen Wiederaufbaus in den neuen Bundesländern.* Göttingen: Verlag für angewandte Psychologie.

STROHSCHNEIDER, S. (1994). Strategien beim Umgang mit einem komplexen Problem: Ein deutsch-deutscher Vergleich. *Zeitschrift für Arbeits- und Organisationspsychologie, 38*, 34–40.

TROST, G. (1985). Pädagogische Diagnostik beim Hochschulzugang, dargestellt am Beispiel der Zulassung zu den medizinischen Studiengängen. In: R.S. Jäger, R. Horn & K. Ingenkamp (Hrsg.), *Tests und Trends 4* (S. 41–81). Weinheim: Beltz.

WIGDOR, A.K. & SACKETT, P.R. (1993). Employment Testing and Public Policy: The Case of the General Aptitude Test Battery. In: H. Schuler, J.L. Farr & M. Smith (Hrsg.), *Personnel Selection and Assessment. Individual and Organizational Perspectives* (S. 183–204). Hillsdale, NJ: Erlbaum.

WOTTAWA, H. & AMELANG, M. (1980). Einige Probleme der „Testfairneß" und ihre Implikationen für Hochschulzulassungsverfahren. *Diagnostica, 26*, 199–221.