Martin Kersting

Ost-West-Leistungsunterschiede in Berufseignungstests in Abhängigkeit von der kulturspezifischen Wirkung einiger Aufgabenmerkmale

Dokumentation: Kersting, M. (1996). Ost-West-Leistungsunterschiede in Berufseignungstests in Abhängigkeit von der kulturspezifischen Wirkung einiger Aufgabenmerkmale. *Zeitschrift für Arbeits- und Organisationspsychologie. 40 Jg.* (N.F. 14) 3, 106–117.
Schlagwörter: Eignungsdiagnostik, Testfairneß, Neue Bundesländer, Personalauswahl, Intelligenz, Testgüte.

Zusammenfassung

Ostdeutsche Bewerber[1] erzielen in „westdeutschen" Berufseignungstests im innerdeutschen Vergleich schlechtere Ergebnisse. In dem vorliegenden Artikel wird die Frage gestellt, inwieweit die Ost-West-Leistungsdifferenzen von kulturspezifischen Wirkungen bestimmter Testmerkmale abhängen. In der Studie mit 853 Personen aus zwei unabhängigen Untersuchungen wurde der Einfluß (1.) der Testdarbietungszeit, (2.) des „westdeutschen" Sprachmaterials und (3.) der in den Tests repräsentierten „westdeutschen" Wissensdomänen auf die Testleistungen in Ost und West mit unterschiedlichen Methoden analysiert. Als Hauptergebnis kann festgehalten werden: (1.) Bei einigen Aufgaben zeigte sich eine kulturspezifisch unterschiedliche Bearbeitungshäufigkeit, die sich auf die Größe der innerdeutschen Testleistungsunterschiede auswirkte. (2.) Einige wenige Items sprachgebundener Aufgaben zeigten in der ostdeutschen Personengruppe eine – im Kontext der übrigen Items – unerwartete Schwierigkeit. (3.) Die mit einem Kenntnistest gemessenen Ost-West-Differenzen im Umfang des gemeinschaftskundlichen Wissens variierten in Abhängigkeit vom verwandten Aufgabeninhalt. Die untersuchten Testmerkmale wirkten sich zuungunsten der ostdeutschen Teilpopulation auf die *Größe* der innerdeutschen Testleistungsdifferenz aus. Sie konnten diese aber nicht vollständig erklären. Das Ergebnis läßt somit Raum für die Wirksamkeit testexogener Faktoren beim Zustandekommen des Ost-West-Leistungsunterschiedes in den Ergebnissen von Berufeignungstests. Als potentielle testexogene Einflußfaktoren werden insbesondere Stichproben- und Personmerkmale diskutiert.

East-West Performance Differences on West-German Personnel Selection Tests in Relation to Culture-specific Effects of some Test Characteristics

Abstract

East German job applicants perform less well on West German personnel selection tests. The present article examines the extent to which these East-West performance differences are dependent on culture-specific effects of certain test characteristics. In a study on 853 persons from two different aptitude testing sessions the influence of the following factors was investigated: a) the presence of time limits, b) West-German linguistic code, c) West-German knowledge-domains as represented in the test. Main findings were: 1. Processing frequency differed for some tasks in a culturally specific manner which affected the magnitude of inner-German performance differences. 2. As compared to overall scores, specific items pertaining to verbal ability turned out to have an unexpectedly high degree of difficulty within the East German sample. 3. East-West differences as measured with a politics-knowledge-test varied dependent on respective task contents. All these examined test characteristics affected the *magnitude* of performance differences, favoring West-German as compared to East-German performance results, though they are not completely accounting for the existing differences. It may be assumed that test-exogeneous factors also play a part in the occurence of East-West performance differences on aptitude tests. As an example for such exogeneous factors, the role of sample and person characteristics is discussed.

1 Einleitung

Seit der Wiedervereinigung Deutschlands werden bei der psychologischen Diagnose der Berufseignung ostdeutscher Bundesbürger u. a. „westliche" Testverfahren angewandt. Damit stellt sich die Frage, ob die relevanten personellen Leistungsvoraussetzungen der Kandidaten aus den neuen Bundesländern mit diesen Testverfahren adäquat erfaßt werden und ob die Tests in den neuen Bundesländern genauso funktionieren wie in den Altbundesländern.

Ersten Befunden zufolge müssen beim Einsatz „westdeutscher" Eignungstests für ostdeutsche Testanden geringfügig niedrigere Leistungen erwartet werden. Ein westdeutscher Testleisungsvorteil zeigte sich im Rahmen der Auswahl von Bewerbern für medizinische Studienplätze (siehe Hensgen & Blum, 1992, 1995; Blum & Hensgen, 1993, 1994), für anspruchsvolle Führungspositionen (Stratemann, 1992), für die Offizierslaufbahn und für eine Ausbildung zum gehobenen Dienst bei verschiedenen öffentlichen Arbeitgebern. (Nähere Angaben finden sich in der Übersicht bei Kersting, 1995, S. 32 f.)

Gruppenspezifische Leistungsunterschiede werden oft zum Anlaß genommen, umstandslos auf die vermeintliche Unangemessenheit oder unzulängliche Güte des angewandten Testverfahrens zu schließen. Wie es ohne weiteres keine Setzungen darüber geben kann, daß eine Gruppe in ihren Testleistungen einer anderen Gruppe im Durchschnitt überlegen ist, kann es auch kein Diktat des Leistungsgleichstandes geben (siehe z. B. Reynolds & Brown, 1984, S. 24). Solche Setzungen spiegeln egalitäre oder meritokratische Ideologiepositionen, taugen aber nicht als Testgütekriterien. Jensen (1980, S. 370) hat diese Art des Mißverständnisses in der Diskussion um Gruppenunterschiede in Tests als *„egalitarian fallacy"* bezeichnet. Innerdeutsche Leistungsdifferenzen in Eignungstests allein

1 Hier wie im folgenden ist die weibliche Form immer mit gemeint.

rechtfertigen weder die Aussage, daß beim Einsatz in den neuen Ländern *„im Westen erfolgreiche Selektionsinstrumente nicht funktionieren"* (Stratemann, 1994, S. 41) noch die Forderung nach *„ostspezifischen Normen"* (ebd, S. 43). Mit solchen Interpretationen und Forderungen wird das Terrain durcheilt, bevor es vermessen ist. Die Erweiterung des Geltungsanspruchs eines Eignungstests begründet allerdings die Notwendigkeit der Kontrolle der Validitätsvoraussetzungen und der Validität für die neue Population. *„Test users should verify periodically that changes in populations of test takers (...) have not made their current procedures inappropriate."*, heißt es auf Seite 42 der Standards for Educational and Psychological Testing der American Psychological Association (1985).

Dabei ist zwischen (1.) der Analyse der *Fairneß* einer Auswahlentscheidung, bei der u.a. von Testresultaten *Gebrauch* gemacht wird, und (2.) der Analyse des Tests, z. B. in Form der Suche nach testinhärenten Bedingungsfaktoren der Gruppendifferenzen, zu unterscheiden. Interessiert man sich dafür, ob die aus den Testergebnissen gewonnenen *Schlußfolgerungen* valide und fair sind, verlagert man die Betrachtung von Testkennwerten auf die Analyse von gruppenspezifischen Kriterium-Prädiktor-Beziehungen. Einen praktischen Bezug bekam dieses – im Vergleich mit angloamerikanischen Ländern im deutschsprachigen Raum eher selten aufgegriffene – Thema in Deutschland vor allem 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 bezug auf den Einsatz westdeutscher Berufseignungstests in den neuen Bundesländern stellt sich in diesem Zusammenhang die Frage, ob die Kriteriumsvalidität der Tests für beide Gruppen (Ost/West) gleich hoch ausfällt. Hierzu lagen nach Kenntnis des Autors zum Zeitpunkt der Artikelerstellung keine Daten vor. Selbst wenn sich die auf westdeutschen Eignungstests basierenden Auswahlverfahren für ost- und westdeutsche Bewerber zumindest im Sinne des *„(Regressions-)Modells der fairen Vorhersage"* nach Cleary als *„fair"* erweisen sollten[2] – und die Erfahrungen mit kulturell unterschiedlicheren Gruppen (z. B. Schmitt & Noe, 1986; Hunter, Schmidt & Hunter, 1979) begründen diese Annahme – bleibt die Klassifikationsentscheidung für die im Prädiktor unterschiedlich starken Gruppen unter spezifischen Fairneßgesichtspunkten[3] suspekt. Aufgrund fehlerbehafteter Messungen auf seiten des Kriteriums und des Prädiktors gibt es keine perfekten Eignungsprognosen. Selbst bei identischen Kriteriums-Prädiktor-Verhältnissen können sich testleistungsheterogene Gruppen unterschiedlich auf die beiden Fehler*typen* der Vorhersage verteilen (siehe Wigdor & Sackett, 1993). Mit Hilfe einer Modellrechnung (Kersting, 1995, S. 37 f.) auf Basis der empirisch nachgewiesenen innerdeutschen Unterschiede in einem „westlichen" Eignungstest konnte gezeigt werden, daß geeigneten Bewerbern aus den neuen Bundesländern aufgrund ihrer relativen Testdefizite vergleichsweise häufiger zu Unrecht die Zielposition verwehrt wird (falsch Negative), während ihren westlichen Landsleuten relativ häufiger der „angenehmere" Fehler einer Übetschätzung zufällt (Zulassung trotz mangelnder Eignung). Gruppenunterschiede in Eignungstests konstituieren also selbst bei *identischen* Steigungen der Regressionslinien zwischen Test und Kriterium für beide Populationen ein Problem bei der Entscheidungsfindung. Zusätzlich zu der Frage nach der (singulären oder differentiellen) prognostischen Validität sollte daher untersucht werden, *wodurch* gruppenspezifische Leistungsunterschiede bedingt sind und ob sich bestimmte Bedingungsfaktoren ausschalten oder in ihrem Einfluß mindern lassen. In diesem Kontext stellt sich u. a. die Frage, ob bestimmte Merkmale der „westdeutschen" Eignungstests in den neuen Bundesländern ein anderes Antwortverhalten provozieren als in den alten Bundesländern. Diese Fragestellung verfolgt der vorliegende Artikel. Die Arbeitshypothese lautet, daß die innerdeutschen Gruppenunterschiede in Berufseignungstests z. T. durch kulturspezifische Wirkungen einiger Aufgabenmerkmale der verwendeten „westdeutschen" Tests bedingt sind. Dabei geht es *nicht* um die Revision des eingesetzten spezifischen Testverfahrens, sondern die Analysen wollen am Beispiel einiger Tests allgemein für das Thema sensibilisieren und auf einige mögliche Prüfungsstrategien sowie auf textexogene Alternativhypothesen hinweisen.

2 Theoretischer Hintergrund

Die Frage, ob bei der Leistungsmessung spezifische Gruppen durch testinhärente Merkmale privilegiert oder diskriminiert werden, wurde in Deutschland – mit wenigen Ausnahmen, siehe etwa die Arbeiten zum Test für medizinische Studiengänge (TMS, z. B. Klieme, 1991; Hensgen & Blum, 1995) – nur selten aufgegriffen. Dabei dürfte es unumstritten sein, daß Testmerkmale wie z. B. Zeitdruck, Antwortformat, sprachliche Einkleidung und wissensspezifische Elemente das Testbearbeitungsverhalten beeinflussen. Bei der Anwendung eines Tests in unterschiedlichen Gruppen wird meist implizit davon ausgegangen, daß die Testmerkmale sich in gleicher Weise auf diese Gruppen auswirken. In bezug auf den Einsatz „westdeutscher" Eignungstests in den neuen und in den alten Bundesländern ergeben sich bezüglich dieser Gleichwirksamkeitsannahme aber die im folgenden dargestellten Zweifel.

2.1 Formale Aspekte: Zeitbegrenzung und Antwortformat

Testleistungen hängen nicht nur von der Leistungsfähigkeit der Person, sondern auch von der Vertrautheit der

2 Ein Auswahlverfahren ist dem *„(Regressions-)Modell der fairen Vorhersage"* 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"* (Bartussek, 1982, S. 3).

3 Z. B. nach der Betrachtungsweise des Fairneßmodells der *„bedingten Wahrscheinlichkeiten"* nach Cole (1973; zitiert nach Möbus, 1983).

Person mit der Untersuchungssituation ab. Durch Antwortstrategien und durch einen geschickten Umgang mit der knappen Ressource „Bearbeitungszeit" können sich Bewerber unter Ausnutzung der Ratewahrscheinlichkeit in Vorteil setzen. Testtrainings, in denen u.a. Testbearbeitungstechniken vermittelt werden, zeitigen nachweislich einen positiven Effekt auf die Testleistungen (siehe z. B. die Zusammenstellung der entsprechenden Effekt-Größen aus sieben Metaanalysen bei Lipsey und Wilson, 1993).

Auch wenn nicht davon ausgegangen werden soll, daß ein Großteil der westdeutschen Bewerber vor der Testung regelrechte Trainingsprogramme absolviert hat, kann dem Aspekt der mittelbaren oder unmittelbaren Testerfahrung dennoch eine Bedeutung für den innerdeutschen Vergleich von Testleistungen zukommen. Denn nicht nur ein Training, sondern z. B. auch die in Eigenregie, etwa mit Hilfe eines Buches, durchgeführte Vorbereitung (siehe etwa van der Molen, Te Nijenhuis & Keen, 1995) oder die reine Vertrautheit mit der Untersuchungssituation und/oder mit Antwortprozeduren (etwa durch die Erfahrung mit ähnlichen Erhebungsverfahren in der Schule) kann sich positiv auf die Testleistung auswirken.

Die Bewerber aus den neuen Bundesländern dürften im Durchschnitt weniger test- und testlektüreerfahren sein als ihr Gegenüber aus den Altbundesländern. Zwar gab es auch in der DDR Leistungstests, sie fanden aber als *„Überbleibsel der bürgerlichen Psychologie"* und infolge des *„Testverbots in der UdSSR von 1936, dem sog. Pädologie-Dekret des ZK der KPdSU"* eine ungleich geringere Anwendung (Ettrich & Guthke, 1991, S. 18).

Die Geschwindigkeit der Testbearbeitung steht bei einigen Tests aufgrund der Zeitbegrenzung und der Ratewahrscheinlichkeit im Zusammenhang mit der Testleistung und stellt eine wesentliche Komponente der Anfälligkeit von Tests gegenüber Trainings dar (siehe Hartigan & Wigdor, 1989, S. 282). Testtrainings oder Testerfahrung sind aber nicht die einzigen Bedingungsfaktoren der Geschwindigkeit der Testbearbeitung. Neben Persönlichkeitsmerkmalen wie z. B. Motivation oder Ängstlichkeit wird gleichfalls die *Kultur* explizit als eine potentielle Determinante des Bearbeitungstempos diskutiert (Iseler, 1970, S. 234). Die Vorstellung und Bedeutung von „Zeit" und „Geschwindigkeit" dürfte über die Kulturen differieren (Thomas & Helfrich, 1993), und gerade hinsichtlich der „Zeit" dürfte die ehemalige DDR ein wohlhabenderes Land gewesen sein als ihr hektisches westliches Pendant.

Die zu prüfende These lautet, daß die zeitbegrenzte Vorgabe der Aufgaben die möglicherweise langsam aber genau arbeitenden ostdeutschen Bewerber gegenüber den eventuell schnelleren westdeutschen Bewerbern, die sich auch mit Ungefährlösungen oder geratenen Antworten zufrieden geben, benachteiligt.

Für diese These sprechen neben den oben ausgeführten Überlegungen auch erste empirische Befunde. Bei den 1992 und 1993 im Rahmen des besonderen Auswahlverfahrens für medizinische Studiengänge durchgeführten Tests schnitten Teilnehmer aus den neuen Bundesländern in den Aufgabengruppen, bei denen es u.a. auf die Bearbeitungsgeschwindigkeit ankommt, schlechter ab als Deutsche aus den alten Bundesländern (Blum & Hensgen, 1994, S. 72; Hensgen & Blum, 1995, S. 73). Obwohl die Lösung von computersimulierten komplexen Problemen nur bedingt mit der Lösung von Intelligenztestaufgaben vergleichbar ist, kann dennoch der von Strohschneider (1994, S. 36 f.) berichtete Befund, daß ostdeutsche Probanden zu Beginn der Bearbeitung eines komplexen Problems deutlich langsamer (bedächtiger oder zögerlicher) arbeiteten als die westdeutschen Probanden, in diesem Zusammenhang angeführt werden.

2.2 Inhaltliche Aspekte und ihre Auswirkungen auf die Itemschwierigkeit und die Schwierigkeitsabfolge

Sowohl formale Aspekte der Testdarbietung als auch inhaltliche Aspekte der Aufgaben können dazu führen, daß die Lösung von Testitems den beiden deutschen Gruppen unterschiedlich schwer fällt. Erste Ergebnisse hierzu wurden von Hensgen und Blum (1995) vorgestellt. Während die Analysen für die Items von fünf der acht geprüften TMS-Untertests keine bedeutsamen Schwierigkeitsunterschiede zwischen den beiden deutschen Gruppen indizierten, wiesen 12,5 % der 24 Items des Untertests „Quantitative und formale Probleme" in den beiden deutschen Gruppen statistisch bedeutsam unterschiedliche Schwierigkeiten auf. Gruppenspezifisch divergierende Itemschwierigkeiten konstituieren sowohl unmittelbar (z. B. in ihrer Auswirkung auf die Aufgabentrennschärfe) als auch mittelbar – z. B. in Kombination mit der begrenzten Testdarbietungszeit und der Itemdarbietungsfolge – ein psychometrisches Problem für den Einsatz des Test in den unterschiedlichen Gruppen. Die weitverbreitete Reihung der Items nach ansteigender Schwierigkeit hat bei zeitbegrenzter Testvorgabe einen Effekt auf die Leistung (siehe z. B. Sax & Cromack, 1966). Sollte die an der westlichen Referenzgruppe adjustierte Darbietungsfolge der Items für die ostdeutschen Bewerber nicht „stimmen", so besteht die Gefahr, daß sich Personen aus den neuen Bundesländern an „zu früh" dargebotenen schwierigen Items „festbeißen" und nicht mehr rechtzeitig zu den für sie leichteren – weiter hinten dargebotenen – Items vordringen. Sofern Tests bei unterschiedlichen Gruppen angeboten werden, sollte man wachsam gegenüber gruppenspezifisch unterschiedlichen Itemschwierigkeiten sein.

Als potentielle Ursachen für gruppenspezifisch unterschiedliche Itemschwierigkeiten – und somit ggf. auch für Probleme der Itemreihung – nennen Reynolds und Brown (1984, S. 25): (1) Fragen nach Inhaltsgebieten, die in der Minderheitengruppe weniger lernzugänglich sind oder waren, (2) eine willkürliche, den kulturellen Erfahrungen der Minderheitengruppe widersprechende Einstufung der Antworten als „falsch" oder „richtig" und (3) eine Wortwahl der Fragestellung, die einer der zu testenden Gruppen unvertraut ist. Einige dieser potentiellen Ursachen sollen im folgenden noch genauer erläutert werden.

2.2.1 Linguistische Äquivalenz

Als eine mögliche Quelle für kulturspezifische Effekte von Aufgabenmerkmalen wird immer wieder die Sprache genannt. Ergänzend zu den Testmerkmalen – wie z. B. der Schwierigkeit der sprachlichen Formulierung der Items – wird dabei auch die verbale Kommunikation mit den testdurchführenden Personen thematisiert (z. B. bei Reynolds & Brown, 1984, S. 17; bei Helms, 1992, S. 1092 f.; bei van

de Vijver & Poortinga, 1992, S. 19). Die Frage der linguistischen Äquivalenz betrifft natürlich insbesondere Tests, die von einer Sprache in die andere übersetzt und/oder bei Nichtmuttersprachlern angewandt werden. Gleichwohl diskutiert z. B. Helms (1992, S. 1093) in diesem Zusammenhang explizit auch verschiedene linguistische „Versionen" *einer* Sprache. Die Sprache und die Lebenswelt der Sprecher beeinflussen sich wechselseitig. Dies gilt insbesondere für Wortbedeutungen. Ein Wort ruft etwas hervor, das seine Bedeutung zum Teil aus der Erfahrung hat. Auch wenn Wortbedeutungen relativ konstant und intersubjektiv sind, können 40 Jahre der deutsch-deutschen Trennung zu einem – zumindest in Nuancen – unterschiedlichen Sprachgebrauch geführt haben. Die Wörter *„Broiler"* oder *„Sättigungsbeilage"* waren hüben ebenso ungebräuchlich wie drüben die Verwendung der Anrede *„sehr geehrte"* statt des vertrauten *„werte"*[4]. Es läßt sich die These ableiten, daß die sprachlichen Merkmale der Tests sich in beiden Teilpopulationen Deutschlands u. a. in Form unterschiedlicher Itemschwierigkeiten und -schwierigkeitsrangfolgen auswirken.

2.2.2 Repräsentation kulturspezifischer Kenntnisse

Besonders bei Kenntnistests besteht die Gefahr, daß bildungsspezifische Eigenheiten der Kultur von Testkonstrukteuren in unzulässiger Weise auf andere Kulturen übertragen werden. In der Literatur werden solche Effekte u. a. unter den Stichworten *„inappropriate content"* (Reynolds & Brown, 1984, S. 17) oder *„content bias"* (Walsh & Betz, 1985, S. 380) thematisiert. Die Probleme einer unreflektierten Anwendung westdeutscher Tests in der DDR verdeutlichen Ettrich und Guthke (1991, S. 17) *„Zum Beispiel erwies sich die Frage im HAWIK, Warum ist es besser, einer Wohltätigkeitsorganisation Geld zu geben als einem Bettler?' als unverständlich, da weder Wohltätigkeitsorganisationen noch Bettler zum Erfahrungsschatz der Kinder gehörten."* Die Autoren überlegen dann, inwieweit nach der Wiedervereinigung die mit diesen Items angesprochenen Kenntnisse wieder in das Alltagswissen eingegangen sind bzw. eingehen werden. Allgemein läßt sich die These formulieren, daß die in den Items von Kenntnistests repräsentierten „westlichen" Wissensdomänen sich negativ auf die Testleistungen ostdeutscher Testanden auswirken.

3 Probandengruppen und Meßinstrumente

Grundlage der folgenden vergleichenden Analysen sind die Testergebnisse von insgesamt 853 Personen, die sich anläßlich ihrer Bewerbung um eine Ausbildung zum gehobenen nicht-technischen Verwaltungsdienst bei ost- und westdeutschen Behörden einem Eignungstest unterzogen haben. Die Eignungsuntersuchungen wurden von der Deutschen Gesellschaft für Personalwesen (*DGP*) durchgeführt. Die Untersuchungen waren überwiegend als mehrstufiges Auswahlverfahren mit dem sogenannten „Vortest" als erster Selektionsstufe konzipiert. Die folgenden Ausführungen beschränken sich auf diejenigen Testaufgaben, die zum „Vortest" zählen, und somit auf eine hinsichtlich Leistungstests unausgelesene Gruppe.

Zur Absicherung der Ergebnisse gegenüber zeitlichen Schwankungen und gegenüber stichprobenspezifischen Effekten wurden jeweils zwei unabhängige Untersuchungsgruppen analysiert. Die erste Untersuchungsgruppe (U1) setzte sich aus 255 ostdeutschen und 245 westdeutschen Bewerbern zusammen, die im Zeitraum Juli 1991 bis Juni 1992 getestet wurden.[5] Die 110 Bewerber aus den neuen und 243 Bewerber aus den alten Bundesländern der zweiten Untersuchungsgruppe (U2) haben sich im Zeitraum von Oktober 1993 bis September 1994 der Testung unterzogen. Demographische Angaben zu den Personengruppen können der Tabelle 1 entnommen werden.

Bei den insgesamt neun Aufgabentypen, die zum „Vortest" gehören, handelt es sich um unveröffentlichte Eigenkonstruktionen der DGP. Die Rechtschreibkenntnisse wurden mit einen Lückendiktat, das gemeinschaftskundliche Wissen mit einem Kenntnistest überprüft. Dieser Kenntnistest wurde innerhalb des Zeitraums der ersten Untersuchung modifiziert, über Art, Anlaß und Auswirkung dieser Modifikation wird weiter unten berichtet. Neben einer Überprüfung der Geschwindigkeit und Sorgfalt des Arbeitsverhaltens mit Hilfe der sogenannten „Postaufgabe" wurden zwei Tests zum sprachgebundenen und vier Tests zum zahlengebundenen Denken eingesetzt. Zwei dieser vier zahlengebundenen Aufgaben zielten eher auf eine Diagnose der entsprechenden Grundfertigkeiten (Textrechnen und Grundrechnen), während die zwei anderen Aufgaben eher das formallogische Denken erfassen sollen.

Da es hier nicht um eine Testrevision, sondern um eine exemplarische Demonstration der kulturspezifischen Effekte von Testmerkmalen und um Beispiele für entsprechende Analysetechniken geht, wurden für die weiter unten (Abschnitt 4.2–4.4) dargestellten Itemanalysen nur vier dieser neun Subtests auf Itemebene für die elektronische Datenverarbeitung aufbereitet und berücksichtigt. Eine Itemanalyse der Aufgabe zum Arbeitsverhalten ist für die vorliegende Fragestellung wenig sinnvoll, da die Schwierigkeit eines Items in reinen Speedtests lediglich eine Funktion der Itemposition ist. Der Einfluß des verwandten Aufgabenmaterials auf die Rechtschreibleistungen im Diktat läßt sich durch eine experimentelle Variation der Diktattexte besser untersuchen als durch eine Itemanalyse. Die Analysen beschränken sich daher auf die Aufgaben zum sprach- und zahlengebundenen Denken sowie auf den Kenntnistest. Bei den zahlengebundenen Aufgaben wurden von den vier Aufgaben je eine zum formallogischen Denken („Zahlenmatrizen") und eine zur Prüfung der Rechenfertigkeiten („Textrechenaufgaben") für die Analyse ausgewählt, bei den sprachgebundenen Aufgaben fiel die Wahl auf die Aufgabe „Analogien". Alle Aufgaben entsprechen üblichen Intelligenztestaufgaben.[6]

4 Auch innerhalb der alten Bundesländer finden sich sprachliche Varietäten, z. B. zwischen Nord- und Süddeutschland (Apfelsine-Orange; Laken-Leintuch usw.).

5 Es handelt sich dabei um eine in bezug auf wesentliche demographische Merkmale (Geschlecht / Alter) und in bezug auf die zentralen Leistungsvariablen parallelisierte Teilstichprobe aus einer bereits an anderer Stelle (Kersting, 1995) erwähnten Untersuchung.

6 Beispielitems für alle Tests können beim Autor angefordert werden.

Tabelle 1
Demographische Angaben zu den Personengruppen

Untersuchung 1	OST (51 %)			WEST (49 %)			GESAMT		
	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂
Personenzahl	142 55,7 %	113 44,3 %	255	133 54,3 %	112 45,7 %	245	275 55 %	225 45 %	500
Alter (Mittelwert)	19,5	22,1	20,6	20,3	23,5	21,8	19,9	22,8	21,2
Untersuchung 2	OST (31 %)			WEST (69 %)			GESAMT		
	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂
Personenzahl	70 63,6 %	40 36,4 %	110	133 54,7 %	110 45,3 %	243	203 57,5 %	150 42,5 %	353
Alter (Mittelwert)	20,0	20,6	20,2	20,5	23,4	21,8	20,3	22,6	21,3

4 Ergebnisse

4.1 Ost-West Testleistungsdifferenzen

In Tabelle 2 finden sich die Ergebnisse für den Ost-West-Vergleich der Testleistungen. Gerechnet wurden Varianzanalysen mit dem Faktor Kultur (Ost/West) als unabhängige Variable und der Testleistung in Standardwertpunkten als abhängige Variable. Da die erste Untersuchungsgruppe mit unterschiedlichen Versionen des Kenntnistests konfrontiert war, wird auf einen allgemeinen Ost-West-Vergleich der Wissensleistungen dieser Gruppe zugunsten der weiter unten (Punkt 4.4) dargestellten differenzierten Analyse verzichtet.

Im Durchschnitt über die acht Tests (ohne den Kenntnistest) übertraf die Westgruppe die Ostgruppe um 2.5 Standardwertpunkte in der ersten und um 4.5 Standardwertpunkte in der zweiten Untersuchung. Mit Ausnahme eines Tests zum formallogischen Denken (VB) waren alle dargestellten Unterschiede statistisch signifikant. Der Faktor „Kulturzugehörigkeit" („Ost" oder „West") erklärte 5.0 % (U 1) bzw. 10.6 % (U 2) der Varianz. Auch innerhalb der West- und der Ostgruppe konnten für Subgruppen Leistungsdifferenzen festgestellt werden. So erzielten z. B. innerhalb der Ostgruppe der U 1 die Absolventen der Erweiterten Oberschule (EOS) die besten 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.

Tabelle 2
Ost-West Unterschiede im Berufseignungstest

Test	U 1 (N Ost=255, N West =245)				U 2 (N Ost=110, N West =243)			
	West[1]	Ost[1]	Diff.	F/Sig.	West[1]	Ost[1]	Diff.	F/Sig.
Diktat 41 Items	99.3 (9.7)	95.2 (10.8)	4.1	19.8 **	98.6 (9.9)	92.1 (11.8)	6.5	28.9 **
ÄW (verbal) 20 Items	99.2 (8.1)	95.9 (8.9)	3.3	18.0 **	98.7 (9.2)	94.9 (8.0)	3.8	13.6 **
AG (verbal) 23 Items	98.5 (10.0)	96.3 (9.6)	2.2	6.0 **	99.1 (11.2)	94.7 (11.2)	4.4	11.9 **
ZZ (numerisch) 15 Items	99.8 (9.8)	96.5 (9.8)	3.3	13.7 **	99.5 (11.7)	94.7 (10.6)	4.8	13.7 **
VB (numerisch) 15 Items	100.7 (9.7)	100.9 (10.8)	-0.2	0.04 n.s.	100.1 (10.5)	99.0 (9.6)	1.1	0.9 n.s.
TX (numerisch) 17 Items	99.8 (9.0)	97.4 (9.1)	2.4	8.9 **	99.5 (8.9)	96.0 (8.0)	3.5	12.8 **
GR (numerisch) 16 Items	98.2 (10.1)	96.2 (9.3)	2.0	5.5 *	97.4 (10.7)	90.8 (10.2)	6.6	30.2 **
PA (Arbeitsverh.) 26 Items	99.3 (10.0)	96.2 (9.3)	3.1	13.1 **	101.7 (10.3)	97.0 (9.8)	4.7	16.8 **
Gem. (Wissen) 40 Items	In der U 1 verschiedene Testversionen, siehe Text				99.9 (9.7)	96.7 (9.3)	3.2	8.8 **
Gesamt ohne "Gem."	99.3 (5.4)	96.8 (5.6)	2.5	26.3 **	99.4 (6.0)	94.9 (6.0)	4.5	41.8 **

Anmerkungen: [1]) Durchschnitt der Standardwertpunkte und (in Klammern darunter) Standardabweichung; * p<.05; ** p<.01;
Diktat – Lückendiktat,
ÄW – Ähnliche Wortbedeutungen,
AG – Analogien,
ZZ – Zahlenmatrizen,
VB – Verschiedene Beziehungen,
TX – Textrechenaufgaben,
GR – Grundrechnen,
PA – Postaufgabe,
Gem. – Gemeinschaftskunde,
Gesamt– Aggregat (ohne Gem.); 8 Tests.

4.2 Zum gruppenspezifischen Einfluß der Zeitbegrenzung auf die Testleistungen

Als Maß für die Testbearbeitungshäufigkeit wurde für jede der vier Aufgaben die durchschnittliche Bearbeitungshäufigkeit pro Gruppe (Ost / West) als einfaches Aggregat über alle einzelnen Items der jeweiligen Aufgabe bestimmt. Ein Item galt dabei als „bearbeitet", wenn bei diesem Item auf dem Antwortbogen von den Bewerbern irgendeine Lösung

(egal ob falsch oder richtig) vermerkt war[7]. Aufgrund der extrem schiefen Verteilungen (die zuerst dargebotenen Items werden immer, die reihungsletzten Items hingegen relativ seltener in Angriff genommen) sind die Verteilungskennwerte „Mittelwert" oder „Standardabweichung" zur Beschreibung der durchschnittlichen Bearbeitungshäufigkeit ungeeignet. Zur Prüfung der Unterschiede zwischen den Gruppen wurde mit dem U-Test von Mann-Whitney ein parameterfreies Verfahren gewählt. Die Ergebnisse sind in der Tabelle 3 dargestellt.

Bei der Interpretation der Ergebnisse ist zu berücksichtigen, daß die Voraussetzungen zum Auffinden von Gruppenunterschieden in der Bearbeitungshäufigkeit bei den analysierten Aufgaben recht unterschiedlich waren. Während bei der sprachgebundenen Aufgabe beide Gruppen insgesamt gut mit der Darbietungszeit zurechtkamen und fast alle Testanden hier alle Items bearbeiteten, stellte die Bearbeitungszeit bei den zahlengebunden Aufgaben für beide Gruppen ein Problem dar, so daß zahlreiche Testbearbeiter unter dem für einen „power Test" als „magische Grenze" definierten Wert von 90% (Hartigan & Wigdor, 1989, S. 101) bearbeiteter Items blieben. Den in Tabelle 3 berichteten Ergebnisse zufolge zeigte sich ein kulturspezifischer Unterschied in der Bearbeitungshäufigkeit nur bei solchen Aufgaben, bei denen grundsätzlich (in beiden Gruppen) Items unbearbeitet blieben. Im vorliegenden Fall waren das vor allem die zahlengebundenen Aufgaben, bei denen sich Ost-West-Unterschiede in der Bearbeitungshäufigkeit zeigten. Unter diesen Testbedingungen (die Mehrzahl der Testanden läßt Items unbearbeitet) bearbeiteten die westdeutschen Bewerber in der verfügbaren Zeit mehr Items als die ostdeutschen Testanden. Leichte Unterschiede ergaben sich für den Kenntnistest. Bei der nach Art des „power Tests" dargebotenen Analogieaufgabe ließ sich die kulturell unterschiedliche Bearbeitungshäufigkeit nicht zufallskritisch absichern.

Das Vorkommen nicht-bearbeiteter Items ist mit großer Wahrscheinlichkeit u.a. auf das Testmerkmal der zeitbegrenzten Darbietung zurückzuführen. Der Befund der in Ost und West unterschiedlichen Bearbeitungshäufigkeit bei drei der vier analysierten Aufgaben spricht dafür, daß sich dieses Testmerkmal in unterschiedlicher Art und Weise in den beiden Gruppen auswirkt. Die im innerdeutschen Vergleich unterschiedlichen Effekte der Bearbeitungszeitgrenzen der Tests stellt eine kulturgebundene Störvariable des diagnostischen Prozesses dar. Als Folge dieser unterschiedlichen Bearbeitungshäufigkeiten können sich neben Reliabilitätseinschränkungen auch gruppenspezifische Verzerrungen in der Leistungsmessung ergeben. Um eine Abschätzung dieses Effekts zu erreichen, wurde durch eine nachträgliche Auswertung der Tests versucht, die Störvariable der Bearbeitungshäufigkeit konstant zu halten. Dazu wurde die Anzahl der Rohwertpunkte auf die Anzahl der überhaupt bearbeiteten Items relativiert. Diese Auswertung berücksichtigte nun nur noch die „Trefferquote" – unabhängig von der Anzahl der bearbeiteten Items. Eine Person, die von 12 bearbeiteten Items sechs richtig beantwortet hat, bekam dieser nachträglichen Auswertung zufolge ebenso den Wert einer fünfzigprozentigen Trefferquote wie eine Person, die zwar insgesamt neun richtige Antworten gegeben hat, aber auch 18 Items bearbeitet (probiert) hatte.

In der Tabelle 4 sind für jede Aufgabe in der oberen Zeile die gruppenspezifischen Ergebnisse der einfaktoriellen Varianzanalysen bei der konventionellen Auswertungsmethode (Anzahl der richtigen Antworten, zur besseren Vergleichbarkeit ebenfalls als Prozentwert – in Prozent des jeweiligen maximalen Rohwerts – ausgedrückt) und bei der Auswertungsmethode der „Trefferquote" (untere Zeile) abgetragen. Die Wirkung der unterschiedlichen Auswertungsmodi soll am Beispiel der Aufgabe „Zahlenmatrizen" veranschaulicht werden. Betrug der Leistungsvorsprung der Westgruppe bei dieser Aufgabe nach der herkömmlichen Auswertung noch 6% (U 1) bzw. 7,6% (U 2), so verringerte sich dieser Leistungsvorteil unter Elimination der Effekte der unterschiedlichen Bearbeitungshäufigkeit

Tabelle 3
Bearbeitungshäufigkeit; U-Test von Mann-Whitney; Durchschnittliche Differenzen (West-Ost) im mittleren Rang

	U 1 (N Ost = 255, West = 245)		U 2 (N Ost = 110, West = 243)	
	Diff.	**p**	**Diff.**	**p**
AG (Analogien)	17.34	.10	12.50	.25
ZZ (Zahlenmatrizen)	38.67	< .01	35.39	< .01
TX (Textrechenaufgaben)	43.17	< .001	56.54	< .001
Gem. (Gemeinschaftskunde)	21.39	< .01	12.12	< .05

Tabelle 4
Ost-West Unterschiede bei verschiedenen Auswertungsmodi

		U 1 (N Ost = 255, N West = 245)				U 2 (N Ost = 110, N West = 243)			
		West	**Ost**	**Diff.**	**F/Sig.**	**West**	**Ost**	**Diff.**	**F/Sig.**
AG	korrekt	72.0	69.1	2.9	6.8**	73.1	68.0	5.1	11.8**
	Treffer	74.8	72.0	2.8	6.8**	78.4	74.3	4.1	11.0**
ZZ	korrekt	58.9	52.9	6.0	16.3**	57.8	50.2	7.6	12.6**
	Treffer	76.3	72.6	3.7	6.0**	75.7	71.2	4.5	5.7*
TX	korrekt	46.3	40	6.3	18.6**	45.8	37.4	8.4	22.1**
	Treffer	70.4	65.3	5.1	10.4**	69.1	64.6	4.5	5.0*
Gem.	korrekt	In der U 1 verschiedene Testversionen, siehe Text				76.0	69.2	6.8	20.0**
	Treffer					76.4	69.6	6.8	20.8**

Anmerkungen: „korrekt" = Anzahl korrekter Lösungen in Prozent: Prozentanteil der gelösten Items an der Zahl der maximal möglichen Lösungen; „Treffer" = Trefferquote: Prozentanteil der gelösten Items an der Zahl aller bearbeiteten Items; * p<.05; ** p<.01; Abkürzungen siehe Tabelle 2.

7 Bei den folgenden Analysen wird von der vereinfachenden Annahme ausgegangen, daß ein Item, zu dem sich kein Lösungsversuch auf dem Antwortbogen findet, von den Probanden nicht bearbeitet wurde. Tatsächlich ist es auch möglich, daß eine Person ein Item mental ausführlich bearbeitet hat, sich dann aber gegen die Fixierung einer Lösung entschied. Dieser Spezialfall wird in der Folge ignoriert.

auf 3,7% (U1) bzw. 4,5% (U2). Allerdings blieb der innerdeutsche Testleistungsunterschied in beiden Untersuchungen bei allen Aufgaben auch bei dieser modifizierten Auswertungsstrategie statistisch signifikant. Bei Aufgaben mit einer geringeren Speedkomponente – wie z.B. der Analogieaufgabe oder dem Kenntnistest – ging mit der veränderten Auswertung erwartungsgemäß kaum eine Veränderung im Ausmaß des Gruppenunterschieds einher.

Man kann konstatieren, daß das Testmerkmal einer – im Verhältnis zu Schwierigkeit und Menge der Items – engen Begrenzung der Darbietungszeit einen nachweisbaren kulturspezifischen Effekt auf die Quantität der Testbearbeitung, aber nur einen geringen kulturspezifischen Effekt auf die Qualität der Testleistungen zeitigte. Die Speedkomponente der als Powertest intendierten Aufgaben vergrößerte – vermittelt über die in Ost und West unterschiedliche Testbearbeitungshäufigkeit – zwar den innerdeutschen *Abstand* der durchschnittlichen Testleistungen. Die Zeitbegrenzung der Tests allein klärte den Mittelwertsunterschied zwischen den beiden deutschen Gruppen aber nicht vollständig auf. Selbst bei der Anwendung der modifizierten Auswertungsstrategie blieb der Leistungsvorteil der Westgruppe bestehen, obwohl die modifizierte Auswertung nach der Methode der „Trefferquote" tendenziell eine Benachteiligung der bearbeitungsfreudigeren Westgruppe darstellte. Denn mehr Items zu bearbeiten bedeutet bei einem Test mit einer ansteigenden Schwierigkeit der Items auch – sofern die Testanden sich an die Darbietungsfolge halten – schwierigere Items zu bearbeiten. Hätte es den ostdeutschen Bewerbern tatsächlich nur an der Testbearbeitungsgeschwindigkeit gemangelt, hätten sie bei der Auswertung nach der Trefferquote mit ihrer innerdeutschen Vergleichsgruppe gleichauf ziehen müssen. Gleichwohl ist das Problem der kulturspezifisch unterschiedlichen Bearbeitungshäufigkeit nicht zu bagatellisieren. Die vorliegende Analyse kann u.U. eine Unterschätzung des Effekts darstellen. Die beiden Tests, bei denen sich die größten innerdeutschen Unterschiede in der Bearbeitungshäufigkeit fanden (ZZ und TX) wurden nämlich mit einem *freien Antwortformat* vorgegeben; d.h., es gab keine vorgegebenen Antwortalternativen und somit eine geringere Ratewahrscheinlichkeit. Der Effekt einer innerdeutsch uneinheitlichen Testbearbeitungshäufigkeit auf die Leistung könnte bei einem multiple-choice-Antwortformat größer ausfallen.

4.3 Itemschwierigkeit, Schwierigkeitsabfolge und linguistische Äquivalenz

Der Einfluß sprachlicher Testmerkmale auf die kulturspezifische Schwierigkeit (und somit auf die Schwierigkeitsabfolge) der Items sollte sich vor allem in sprachgebundenen Tests manifestieren – obwohl natürlich auch hier andere, z.B. formale Aspekte – die Schwierigkeit eines Items stets mitbestimmen. Am Beispiel des Subtests „Analogien" (AG) wurde zunächst geprüft, ob die einzelnen Items in den gruppenspezifischen Schwierigkeitsabfolgen die gleiche Position einnehmen. Abbildung 1 zeigt die Schwierigkeitsabfolge der Items der Aufgabe „Analogien" für die beiden Gruppen der ersten Untersuchung. Auf der Ordinate ist die Itemschwierigkeit abgetragen. Die Items wurden entsprechend ihrer Schwierigkeit *in der Westgruppe* entlang der Abszisse geordnet. Bei identischen Schwierigkeitsabfolgen müßten die beiden Linien deckungsgleich oder (bei leistungsdisparaten Gruppen) parallel zueinander verlaufen. Der Graphik läßt sich entnehmen, daß die Items in bezug auf ihre Schwierigkeit in den beiden Gruppen eine unterschiedliche Rangordnung aufwiesen.

Entscheidend in der Abbildung 1 ist, daß sich die Linien für die Schwierigkeitsabfolge bei bestimmten Items *kreuzen*. Jensen (1980, S.432) benutzt für das Phänomen solcher Wechselwirkungen den Ausdruck *„groups x items interaction"* und schlägt als Prüfungsmethode u.a. die Varianzanalyse vor.

Die naheliegende Verwendung der Varianzanalyse in diesem Zusammenhang ruft allerdings Einwände methodenkritischer Forscher auf den Plan (z.B. Angoff, 1982, S.107f.; Jensen, 1984, S.535f.; Thissen, Steinberg & Gerrard, 1986, S.119). Grundlage der Kritik ist die Tatsache, daß die Anwendung der Varianzanalyse als Prüfungsmethode der vorliegenden Fragestellung unter bestimmten Bedingungen zu Artefakten führen kann. Eine adäquate, teststarke und häufig (z.B. als Standardprüfverfahren beim „Educational Testing Service") angewandte Methode zur Identifizierung gruppenspezifischer Itemfunktionen stellt der sogenannte „Mantel-Haenszel-Ansatz" dar (siehe Dorans & Holland, 1993). Ein u.a. auf dieser Prüfstatistik

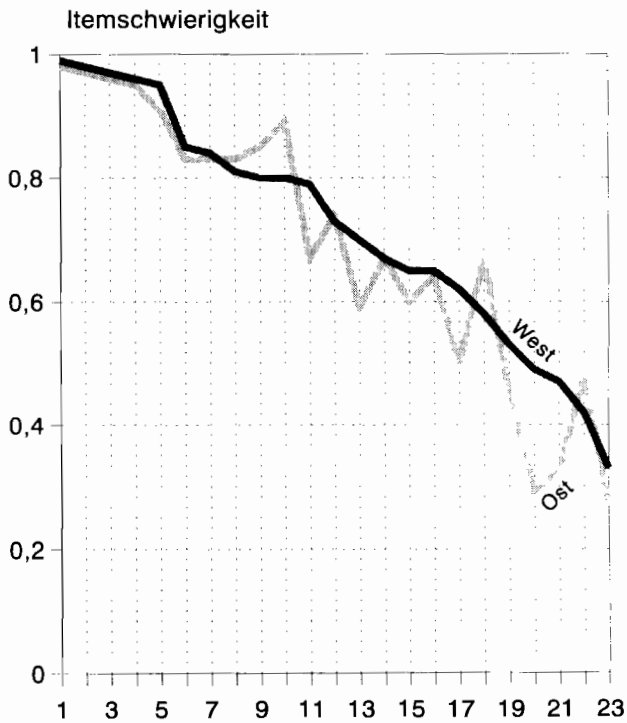

Abbildung 1
Interaktion zwischen der Schwierigkeit der Items und der Gruppenzugehörigkeit; Aufgabe AG
Anmerkungen: Untersuchung 1; N = 500 (255 Ost, 245 West).

aufbauendes computergestütztes Verfahren zur Differentiellen Itemanalyse haben Klieme und Stumpf (1991) entwickelt und programmiert. Mit Hilfe dieses Programms[8] wurde geprüft, ob sich einzelne Items identifizieren lassen, die einer der beiden Gruppen spezifische Schwierigkeiten bereiten. Differentielle Itemanalysen setzen voraus, daß der Gruppenunterschied im Gesamtniveau des (Sub-)Tests „ausgeschaltet" wird. Dies kann beispielsweise durch ein „matching" erzielt werden. Beachtet werden nur solche Differenzen, die auch bei den in bezug auf das Parallelisierungsmerkmal (Leistung in der Gesamtskala) weitgehend identischen Gruppen noch auftreten und die somit auch bei leistungsdisparaten Gruppen *unerwartet* sind. Bei dem Mantel-Haenszel Ansatz wird dies erreicht, indem für jede Teilpopulation entsprechend des Ergebnisses in der Gesamtskala „Niveaugruppen" gebildet werden und die „odds ratio" (p/(1-p)) für jede Niveaugruppe bestimmt wird. Die Hypothese, daß das Chancenverhältnis in den Niveaugruppen der ostdeutschen Teilpopulation sich um einen konstanten Faktor von dem entsprechenden Chancenverhältnis in der westdeutschen Teilpopulation unterscheidet, wird dann durch den von Mantel und Haenszel (im folgenden MH abgekürzt) entwickelten Chi-Quadrat-Test (MH^2 abgekürzt) geprüft (siehe Klieme & Stumpf, 1991).

Da es hier nicht um die Revision eines spezifischen Tests, sondern allgemein um eine Sensibilisierung für die Problematik der im innerdeutschen Vergleich möglicherweise unterschiedlichen Itemschwierigkeiten geht, soll das Ergebnis nur kurz berichtet werden. Der Bericht beschränkt sich auf die drei Items, die sich konsistent, d. h. *in beiden Untersuchungen* als auffällig erwiesen haben. Angesichts der relativ großen Sensibilität der zur Aufdeckung von Itemverzerrungen eingesetzten statistischen Verfahren bei der gleichzeitigen Gefahr von Stichprobenfehlern ist eine solche Replikation der Befunde an unabhängigen Stichproben besonders wichtig (Reynolds & Brown, 1984, S. 26). Bei zwei (knapp 9% der Gesamtitemzahl) der drei betroffenen Items schnitten die ostdeutschen Bewerber schlechter ab, als aufgrund ihres Gesamtleistungsniveaus in der Skala zu erwarten gewesen wäre. Bei einem Item war die Ostgruppe – relativ zu ihrem sonstigen Leistungshandicap - durchschnittlich erfolgreicher. Der deutlichste kulturspezifische Effekt zuungunsten der ostdeutschen Bewerber zeigte sich für die Analogie *„Präambel" verhält sich zu „Verfassung" wie „Prolog" zu „Drama"*, wobei das Lösungswort „Präambel" unter vier Distraktoren herauszufinden war (U1: MHα 1.92, $MH\chi^2$ 6.1, $p<.05$; U2: MHα 2.06, $MH\chi^2$ 12.1, $p<.01$). Bei dem anderen Item, bei dem der Lösungsvorsprung der Westgruppe sich nicht aus den – aufgrund der Leistungen in den übrigen Items formulierten – Erwartungen ableiten ließ, ging es darum zu erkennen, daß das Wortpaar „Dieb"-„Resozialisierung" dem Wortpaar „Trinker"-„Entziehungskur" analog ist (U1: MHα 1.90, $MH\chi^2$ 3.85, $p<.05$; U2: MHα 1.63, $MH\chi^2$ 4.26, $p<.01$).[9] Die Befunde sprechen dafür, daß bei zumindest 9% der Items des sprachgebundenen Tests der im innerdeutschen Vergleich aufgetretene Leistungsvorsprung der westdeutschen Bewerber mehr auf (für den Gesamtwert irrelevante) kulturspezifische Merkmale des Items, denn auf tatsächliche Differenzen in den verbalen Fähigkeiten der beiden Gruppen zurückzuführen waren. Um einen Indikator für die Größe des Effekts zu erhalten, wurden die beiden betroffenen Items bei einer nachträglichen Auswertung ignoriert. Während sich der Leistungsvorteil der Westgruppe in der U2 selbst bei dieser nachträglichen Auswertung auf dem 1% Niveau statistisch absichern ließ (Varianzanalysen mit der Subtestleistung in Rohwertpunkten als abhängige Variable), verfehlte der auf dieser Grundlage berechnete Gruppenunterschied in der U1 knapp die 5% Signifikanzgrenze. Dieser Befund ergab sich auch dann, wenn man nicht nur die beiden Items, die sich zuungunsten der ostdeutschen Bewerber auswirkten, sondern alle drei kulturell voreingenommenen Items bei der nachträglichen Bildung des Skalenwertes außen vor ließ.

Die für die ostdeutschen Bewerber ungewöhnlich schweren Items stellen ein psychometrisches Problem dar, da mit diesen Fragen wahrscheinlich etwas anderes gemessen wird als mit dem restlichen Test. Die abweichende Funktion *einzelner* (weniger) Items bedeutet aber nicht, daß der *Gesamtwert* aller Items dieser Skala eine inadäquate Messung der entsprechenden Leistungsausprägung ostdeutschen Bewerber darstellt. Auch wenn die Messung auf Skalenebene in beiden Gruppen funktioniert, kann es aber sinnvoll sein, kritische Einzelitems – z. B. mit Hilfe der hier beschriebenen Methodik – zu identifizieren.

4.4 Zum Einfluß der Repräsentation „ost-"/„west-" spezifischer Kenntnisse auf die Leistungen im Wissenstest

Aufgrund erster Analysen zu Ost-West-Unterschieden und aufgrund inhaltlicher Überlegungen wurden zum Ende des Jahres 1991 neun der insgesamt 40 Items des Tests „Gemeinschaftskunde" erneuert bzw. modifiziert. Sieben der neun Änderungen wurden mit dem Ziel vorgenommen, die nach Ansicht von Experten bislang überwiegend an den Schwerpunkten „westdeutscher" schulischer Curricula und an „westdeutscher" Allgemeinbildung orientierten Items durch Fragen nach solchen Wissensinhalten zu ersetzen, die sowohl in den alten als auch in den neuen Bundesländern repräsentiert sind. So wurde beispielsweise mit diesen modifizierten Items nach Ereignissen der jüngsten gemeinsamen deutschen Geschichte oder explizit nach Ereignissen in der Historie der DDR gefragt.

Die Testmodifikation fiel in den Erhebungszeitraum der ersten Untersuchung. Dies eröffnete die Möglichkeit, den

8 Für die Bereitstellung des Programms möchte der Verfasser den Programmautoren herzlich danken.

9 Es fällt leicht, im nachhinein plausible „Erklärungen" dafür zu (er)finden, *warum* gerade diese beiden Items in den neuen Bundesländern andere Schwierigkeitswerte erzielen als in den alten Bundesländern und die Notwendigkeit des statistischen „Aufwands" in Frage zu stellen. Versuche, „kulturell verzerrte Items" allein mit Hilfe des „gesunden Menschenverstandes" aufzudecken, erzielten aber selten über dem Zufall liegende Trefferquoten (siehe z. B. Jensen, 1984, S. 515 f.).

kulturspezifischen Effekt der Modifikation zu testen. Allerdings haben (aufgrund einer zeitlich späteren Eröffnung der „Testsaison") nur 43 ostdeutsche (gegenüber 175 westdeutschen) Bewerbern der U1 die „alte" Version des Kenntnistests bearbeitet. Erschwert wurde die Analyse des Modifikationseffekts auf die Messung des Kenntnisstandes auch dadurch, daß im Vergleich der beiden Zeitpunkte der ersten Untersuchung („vor" und „nach" der Einführung des überarbeiteten Kenntnistests) innerhalb der Westgruppe die später getesteten Bewerber insgesamt etwas testleistungsschwächer waren. Solche Veränderungen spiegeln oft „saisonale" Effekte wider und sind nicht ungewöhnlich. Oft sind Bewerber, die sich erst auf „den letzten Drücker" oder auf eine zweite Ausschreibung hin bewerben, durchschnittlich testleistungsschwächer. Beim Vergleich der vor und nach der Testmodifikation getesteten Bewerber aus den neuen Bundesländern ließen sich hingegen auf Gesamttestebene keine nennenswerten Leistungsunterschiede feststellen.

Es wurde eine 2×3 „between subjects"-Manova mit der Gruppenzugehörigkeit (mit den beiden Stufen „Ost" oder „West") und der verwandten Testversion / Untersuchungsgruppe (Stufen 1 und 2: „vor" und „nach" der Modifikation des Kenntnistests innerhalb der ersten Untersuchung, Stufe 3: die zweite Untersuchung mit dem modifizierten Test) als unabhängige Variablen und dem Kenntnisstand als abhängige Variable gerechnet. Um die Effekte des unterschiedlichen Leistungsausmaßes in den übrigen Tests zu kontrollieren, wurde die Gesamttestleistung (ohne den Kenntnistest) als Kovariate in die Analyse eingeführt.

Abbildung 2 zeigt den durchschnittlichen Kenntnisstand in gemeinschaftskundlichen Fragen für die drei Gruppen, jeweils nach „Ost-" und nach „Westbewerbern" aufgeteilt. Abgetragen sind die adjustierten Werte für die Anzahl der richtig beantworteten Kenntnisfragen in Prozent des maximal erreichbaren Punktwertes. Die hochsignifikanten Haupteffekte für die beiden Faktoren wurden durch eine Interaktion (F (2,846) = 8.93, p<.01) spezifiziert. Während die Modifikation des Wissenstests in der Westgruppe (siehe die obere, fast waagerechte Linie in der Abbildung) kaum Veränderungen der Testleistung nach sich gezogen hat, wußten innerhalb der Ostgruppe die mit der modifizierten Version des Wissenstests konfrontierten Bewerber der U1 10% mehr Fragen zu beantworten als diejenigen, denen die alte Testversion appliziert wurde. Die Ergebnisse sprechen dafür, daß der Aufgabenkontent einen gruppenspezifischen Einfluß auf die Testleistung hatte. Alternativ könnte man die Hypothese formulieren, daß der Leistungszuwachs der Ostgruppe nicht auf die Änderung des Tests, sondern auf einen allgemeinen Trend zu kenntnisreicheren Bewerbern in den neuen Bundesländern zurückzuführen ist. Um diese Gegenthese prüfen zu können, wurde auch die U2 – sozusagen als Kontrollgruppe – mit in die Analyse aufgenommen, wobei die Version des Kenntnistests zwischen den Faktorstufen „2" und „3" dieser unabhängigen Variablen konstant gehalten wurde. Wäre die Leistung der Ostgruppe auch ohne eine Änderung der Tests mit der Zeit immer besser geworden, so hätten sich die ostdeutschen Bewerber der zweiten Untersuchung wiederum deutlich in ihrer Leistung von den – ebenfalls mit der modifizierten Version des Wissenstests, aber zeitlich „früher" geprüften – ostdeutschen Bewerber der ersten Untersuchung abheben müssen. Die Graphik verzeichnet für diesen Vergleich aber eine geringere Leistungsdifferenz als für den Vergleich der mit wechselnden Testversionen geprüften ostdeutschen Bewerber.

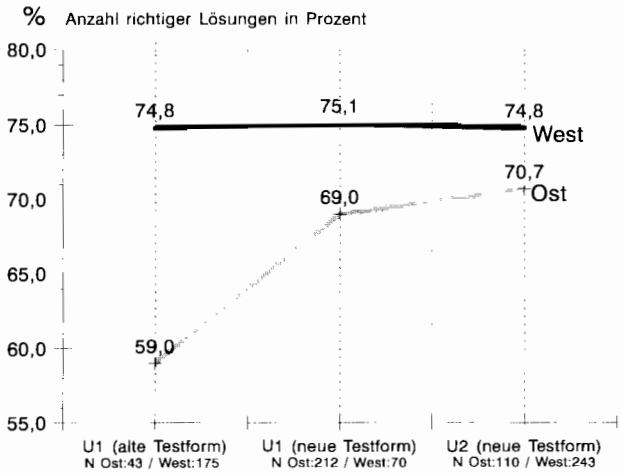

Abbildung 2
Gruppenspezifischer Kenntnisstand in Gemeinschaftskunde vor und nach der Testmodifikation

Innerhalb der Gruppe der ostdeutschen Bewerber war der Leistungsnachteil derjenigen, die mit dem alten, mehr auf westdeutsche Kenntnisse zugeschnittenen Wissenstest geprüft wurden, gegenüber denjenigen, die die modifizierte Version des Wissenstests bearbeitet hatten, mit großer Wahrscheinlichkeit auf die Veränderung des Tests und nicht auf Differenzen der Kenntnisse zwischen diesen beiden Gruppen zurückzuführen. Dies ist ein Indiz für einen kulturspezifischen Effekt von Testmerkmalen auf die Leistung. Allerdings blieben die durchschnittlichen Kenntisse der Bewerber aus den alten Bundesländern auch bei der Messung mit dem modifizierten Testverfahren im Licht innerdeutscher Konkurrenz noch vergleichsweise fundierter.

5 Diskussion

Das theoretische und empirische Material läßt es unwahrscheinlich erscheinen, daß bestimmte Merkmale „westdeutscher" Eignungstests – wie das Verhältnis von Itemanzahl und Testbearbeitungszeit (siehe Punkt 4.2) oder die Verwendung „westdeutscher" Sprach- und Wissensbestände – sich in identischer Weise auf die Testleistung ost- und westdeutscher Testanden auswirken. Der sich andeutende negative Effekt dieser Testmerkmale auf die von ostdeutschen Bewerbern in westdeutschen Berufseignungstests erzielten Ergebnisse muß aber insgesamt als gering eingestuft werden. Einerseits erwiesen sich nur bestimmte Aufgabentypen bzw. vereinzelte Items als kulturell voreingenommen. Andererseits blieb die Testleistungsdisparität zugunsten der

westdeutschen Bewerber zumeist auch dann noch bestehen, wenn der Effekt der kulturspezifisch diskriminierenden Testmerkmale reduziert wurde. Das Ergebnis läßt somit Raum für die Wirksamkeit testexogener Faktoren beim Zustandekommen des Gruppenunterschiedes. Zu nennen sind hier insbesondere Stichproben- und Personmerkmale.

5.1 Stichprobenmerkmale

Es soll explizit betont werden, daß Gruppen von Arbeits- und Ausbildungsplatzsuchenden keinesfalls *repräsentative* Stichproben ihrer jeweiligen Kultur darstellen. Vor dem Hintergrund (a) der Binnenwanderungsprozesse (siehe Maretzke & Möller, 1992) und (b) der ungleichen Verteilung der formellen Bewerbungsvoraussetzungen sowie (c) der eventuell unterschiedlich gelungenen Selbstselektion (siehe z. B. Stratemann, 1992 oder Weinert, 1987) ist es denkbar, daß 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 Untersuchungen mit weniger vorausgewählten Stichproben erklären, z. B. den von Strohschneider (1994) berichteten Gleichstand der allgemeinen intellektuellen Leistungsfähigkeit von Berufsschülern aus beiden Teilen Berlins. Die genannten Aspekte und die unterschiedliche Sozialisation (z. B. andere Bildungsschwerpunkte, kürzere Schulzeiten) können durchaus dazu führen, daß in den neuen Bundesländern geeignete Bewerber für weitgehend westlich geprägte Berufsbilder etwas schwerer zu finden sind. Ost-West Mittelwertsunterschiede in *anforderungsspezifischen* Tests sind somit keinesfalls a priori ein Indikator mangelnder Güte der Meßinstrumente. Hinsichtlich der Gelungenheit der Berufswahl und der Selbstselektion dürfte insbesondere den in Ost- und Westdeutschland unterschiedlichen Arbeitsmarktbedingungen und der subjektiven Wahrnehmung dieser Arbeitsmarktbedingungen eine herausragende Bedeutung zukommen. Jugendliche im Osten sehen in der Arbeitslosigkeit im Vergleich zu Westjugendlichen häufiger ein schwerwiegendes Problem (Heiliger & Kürten, 1992) und betonen in ihren Wertprioritäten (Krebs, 1992) sowie in Fragen der beruflichen Zukunft (Palentien, Pollmer & Hurrelmann, 1993) den Aspekt der Sicherheit bzw. des Unsicherheitsempfindens. Hille (1993) zufolge nannten ostdeutsche Jugendliche wesentlich häufiger instrumentelle und materielle Gründe (Verdienst, Sicherheit, Aufstieg) für die Wahl des Berufs als westdeutsche Jugendliche. Nach Schramm (1992) überschritt 1990 das Ausmaß der Arbeitsplatzunsicherheit im Osten für die Gesamtgruppe das im Westen für Problemgruppen gemessene Maß. Diese Faktoren könnten dazu führen, daß sich Ostdeutsche vergleichsweise häufiger als Westdeutsche auch dann um „sichere" Ausbildungen und Positionen bemühen, wenn diese ihren Interessen und Fähigkeiten weniger entsprechen. Dieses Bewerbungsverhalten könnte sich in den durchschnittlichen Leistungen in Berufseignungstests niederschlagen.

5.2 Personmerkmale

Auch die in zahlreichen Untersuchungen berichteten Ost-West Unterschiede in verschiedenen nicht-kognitiven Personmerkmalen lassen – sofern es sich dabei um Personmerkmale handelt, die im Zusammenhang mit der kognitiven Leistungsfähigkeit(smessung) stehen – keinen innerdeutschen (Test-)Leistungsgleichstand erwarten. Baumert (1994) sowie Oettingen und Little (1993) fanden bei ostdeutschen Schülern beispielsweise vergleichsweise pessimistischere Selbsteinschätzungen des eigenen Leistungspotentials als bei westdeutschen Schülern.

5.3 Konsequenzen

Psychologische Testverfahren sollten einer kontinuierlichen empirischen Kontrolle mit dem Ziel der Qualitätsoptimierung unterliegen. Dabei sind Daten von allen anwendungsrelevanten Personengruppen zu berücksichtigen. Für Berufseignungstests mit gesamtdeutschem Geltungsanspruch gilt somit, daß seit der Wiedervereinigung auch Daten aus den neuen Bundesländern in den Prozeß der Testkonstruktion, der Testpflege und der Testevaluation einzubeziehen sind. Dies bedeutet aber keineswegs, daß diese Qualitätsoptimierung unter dem Imperativ einer Mittelwertsgleichheit für verschiedene Gruppen steht. Die psychometrische Ebene der Testkonstruktion und -evalution ist von der gesellschaftlichen Ebene der Testanwendungen zu unterscheiden (siehe z. B. Walsh & Betz, 1985, S. 383). Auf den Bericht von Gruppenunterschieden in Leistungstests wird oft reflexartig mit der Forderung nach gruppenspezifischen Testaus- oder -bewertungen reagiert. Solche Forderungen sind zumeist gesellschaftlich-politischer Natur und sollten in diesem Kontext diskutiert werden. Dabei ist zu bedenken, daß mit einem solchen gruppenspezifischen Vorgehen Zuordnungsprobleme verbunden sind. Zahlreiche Personen könnten z. B. aufgrund bildungs-, schicht-, regional- und/oder geschlechtsspezifischer Merkmale einen eigenen Gruppenstatus fordern. Wie immer man sich entscheidet: die Entscheidung über die leistungsunabhängige Bevorzugung bestimmter Gruppen ist *außerhalb des diagnostischen Begründungszusammenhangs* zu treffen.

Den für die Qualitätsoptimierung Verantwortlichen sollte auch nicht das Mandat zur Schaffung eines „kulturell freien" Berufseignungstests erteilt werden. Abgesehen davon, daß die Realisierbarkeit „kulturell freier" Testverfahren grundsätzlich in Abrede gestellt werden kann (z. B. Poortinga & van de Vijver, 1987, S. 21), muß auch bezweifelt werden, daß ein kulturell „blinder" Eignungstest gute prognostische Validitäten aufweisen würde (Simons & Möbus, 1978, S. 192; Reynolds & Brown, 1984, S. 22). Eignungsdiagnostik bedingt spezifische Anforderungen, das Ziel sind brauchbare Prognosen des berufserfolgsrelevanten Verhaltens. Das vorherzusagende Verhalten findet in einem u. a. kulturell definierten Rahmen statt, für universelle/kulturell unspezifische Tests sind in diesem diagnostischen Zusammenhang eher geringere Validitätskoeffizien-

ten zu erwarten. Mit Bezug auf berufliche Anforderungen kann man auch dann auf bestimmten kulturspezifischen Testmerkmalen insistieren, wenn diese sich nachweislich negativ auf die Leistung bestimmter Gruppen auswirken. Die Eliminierung einzelner Items kann die Validität eines Testes erhöhen, birgt aber ebenso die Gefahr der Validitätsminimierung (Schmitt, 1989, S. 148). Schließlich ist auch der oft propagierte Ersatz der herkömmlichen westdeutschen Tests durch „neue" Testverfahren problematisch, wenn – wie so oft – für diese Verfahren die notwendigen Kennwerte und Validitätsnachweise fehlen.

Die vorgestellten Analysen sind weder Anlaß noch Rechtfertigung für drastische Verfahrens- oder Auswertungsänderungen westdeutscher Berufseignungstests mit gesamtdeutschem Geltungsanspruch. Unter der Voraussetzung, daß ein Test für den Einsatz in Ost- und Westdeutschland vorgesehen ist, sollte aber innerhalb der oben dargestellten Modifikationsgrenzen die kulturelle Ausgewogenheit spezifischer Aspekte des Aufgabenmaterials und der Aufgabendarbietung angestrebt werden. Voraussetzungen hierfür ist, daß die gruppenspezifische Wirkung bestimmter Testmerkmale – wie in anderen Ländern üblich – auch in Deutschland *standardmäßig* geprüft werden. Darüber hinaus werden Untersuchungen zum innerdeutschen Vergleich der prognostischen Validitäten von Berufseignungstests benötigt.

Unter dem Vorbehalt der Replikation der dargestellten Befunde zur kulturspezifischen Wirkung von Testmerkmalen sowie den noch ausstehenden Befunden zur Frage der Kriteriumsvalidität, lassen sich die folgenden vorläufigen Empfehlungen zur Optimierung westdeutscher Berufseignungstests mit gesamtdeutschem Geltungsanspruch formulieren:

- Für Niveautests ist zu prüfen, ob der Streuungsbereich der Testpunktwerte ohne Einbußen bei der Homogenität und Reliabilität eine – für beide Gruppen identische – Erhöhung der Testbearbeitungszeit erlaubt. Durch eine solche Maßnahme würde die Wahrscheinlichkeit erhöht, daß die Leistungsvarianz auf die Niveau- und nicht auf die in Ost- und Westdeutschland möglicherweise unterschiedliche Speedkomponente zurückzuführen ist.
- Bei differentiellen Itemschwierigkeiten sollte – sofern ein Austausch der betroffenen Items aus Gründen der Validität nicht in Frage kommt – versucht werden, Effekte der Darbietungsfolge durch eine gruppenspezifische Itemreihung zu vermindern. Für die so entstehenden neuen Testformen ist dann der Parallelitätsnachweis zu erbringen.
- Sofern die diagnostische Fragestellung es erlaubt, sollten bei der Modifikation von Kenntnistests die zeitweilig unterschiedlichen schulischen Curricula und Informationsgrundlagen in beiden Teilen Deutschlands Berücksichtigung finden. Empfehlenswert sind mit ost- und westdeutschen Experten besetzte Testentwicklungsteams. Vergleichbares gilt für das verwendete Sprachmaterial von Tests.
- Die möglicherweise geringere Vertrautheit der ostdeutschen Bewerber mit der Testsituation unterstreicht noch einmal die ohnedies notwendige ausführliche, ansprechende und anschauliche Vorabinformation *aller* Testanden.
- Mit umfangreichen realistischen und verständlichen Informationen über den im Rahmen des Auswahlverfahrens zu vergebenden Ausbildungs- oder Arbeitsplatz kann der Prozeß der Selbstselektion positiv beeinflußt werden.

Bei den ersten vier Empfehlungen ist zu beachten, daß Eingriffe in etablierte Testverfahren sorgfältig abzuwägen und in ihren Effekten empirisch zu kontrollieren sind. Die Modifikationen dienen dem Ziel, eine Diskriminierung aufgrund irrelevanter Gruppenzugehörigkeiten auszuschließen. Sie sollten aber nicht über das Ziel hinausschießen und die mit Berufseignungstests *intendierte Unterscheidung* zwischen potentiell erfolgreichen und weniger erfolgreichen Bewerbern vereiteln.

Literatur

American Psychological Association (1985). *Standards for Educational and Psychological Testing*. Washington, D. C.: American Psychological Association.

Angoff, W. H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R. A. Berk (Hrsg.), *Handbook of methods for detecting test bias* (S. 96–116). Baltimore: John Hopkins University Press.

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. 22–95). 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.

Dorans, N. J. & Holland, P. W. (1993). Differential item functioning detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Hrsg.), *Differential item functioning* (S. 507–586). New York: Plenum Press.

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.

Hartigan, J. A. & Wigdor, A. K. (1989). *Fairness in employment testing*. Washington, DC: National Academy Press.

Heiliger, C. & Kürten, K. (1992). Jugend '92: Ergebnisse der IBM-Jugendstudie. In Institut für Empirische Psychologie (Hrsg.), *Die selbstbewußte Jugend* (S. 68–155). Köln: Bund-Verlag.

Helms, J.E. (1992). Why is there no study of cultural equivalence in standardized cognitive ability testing? *American Psychologist, 47*, 1083–1101.

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.

Hensgen, A. & Blum, F. (1995). 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. 19. Arbeitsbericht* (S. 56–85). Bonn: Institut für Test- und Begabungsforschung.

Hille, B. (1993). Lebenssituation und Lebensperspektiven Jugendlicher im vereinten Deutschland. *Aus Politik und Zeitgeschichte, 24*, 14–20.

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.

Iseler, A. (1970). *Leistungsgeschwindigkeit und Leistungsgüte*. Weinheim: Beltz.

Jensen, A.R. (1980). *Bias in mental testing*. Cambridge: Cambridge University Press.

Jensen, A.R. (1984). Test bias. Concepts and criticisms. In C.R. Reynolds & R.T. Brown (Hrsg.), *Perspectives on bias in mental testing* (S. 507–586). New York: Plenum Press.

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

Klieme, E. (1991). Werden bestimmte TMS-Aufgaben geschlechts- oder interessentypisch bearbeitet? In G. Trost (Hrsg.), *Test für medizinische Studiengänge (TMS): Studien zur Evaluation. 15. Arbeitsbericht* (S. 148–170). Bonn: Institut f. Test- u. Begabungsforschung.

Klieme, E. & Stumpf, H. (1991). DIF: A computer programm for the analysis of differential item performance. *Educational and Psychological Measurement, 51*, 669–671.

Krebs, D. (1992). Werte in den alten und neuen Bundesländern. In Jugendwerk der Deutschen Shell (Hrsg.), *Jugend '92* (Bd. 2, Im Spiegel der Wissenschaften, S. 35–48). Opladen: Leske + Budrich.

Lipsey, M.W. & Wilson, D.B. (1993). The efficacy of psychological, educational and behavioral treatment. Confirmation from meta-analysis. *American Psychologist, 48*, 1181– 1209.

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

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.

Oettingen, G. & Little, T.D. (1993). Intelligenz und Selbstwirksamkeitsurteile bei Ost- und Westberliner Schulkindern. *Zeitschrift für Sozialpsychologie, 24*, 186–197.

Palentien, C., Pollmer, K. & Hurrelmann, K. (1993). Ausbildungs- und Zukunftsperspektiven ostdeutscher Jugendlicher nach der politischen Vereinigung Deutschlands. *Aus Politik und Zeitgeschichte, 24*, 3–13.

Poortinga, Y.H. & Vijver, F.J.R. van de (1987). Cultural bias and the interpretation of test scores. In C. Schwarzer & B. Seipp (Hrsg.), *Trends in european educational research* (Bd. 20, S. 13–21). Braunschweig: Braunschweiger Studien zur Erziehungs- und Sozialarbeitswissenschaft.

Reynolds, C.R. & Brown, R.T. (1984). Bias in mental testing. In C.R. Reynolds & R.T. Brown (Hrsg.), *Perspectives on bias in mental testing* (S. 1–39). New York: Plenum.

Sax, G. & Cromack, T. (1966). The Effects of various forms of item arrangements on test performance. *Journal of Educational Measurement, 3*, 309–311.

Schmitt, N. (1989). Fairness in employment selection. In M. Smith & I.T. Robertson (Hrsg.), *Advances in selection and assessment* (S. 133–153). New York: Wiley.

Schmitt, N. & Noe, R.A. (1986). Personnel selection and equal employment opportunity. In C.L. Cooper & I. Robertson (Hrsg.), *International Review of Industrial and Organizational Psychology* (S. 71–115). New York: Wiley.

Schramm, F. (1992). *Beschäftigungsunsicherheit. Wie sich die Risiken des Arbeitsmarktes auf die Beschäftigten auswirken – Empirische Analysen in Ost und West*. Berlin: Edition Sigma.

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.

Stratemann, I. (1994). Personalauswahl und -entwicklung in den neuen Bundesländern. *Zeitschrift für Arbeits- und Organisationspsychologie, 38*, 41–45.

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

Thissen, D. Steinberg, L. & Gerrard, M (1986). Beyond group-mean differences: The concepr of item bias. *Psychological Bulletin, 99*, 118–128.

Thomas, A. & Helfrich, H. (1993). Wahrnehmungspsychologische Aspekte im Kulturvergleich. In A. Thomas (Hrsg.), *Kulturvergleichende Psychologie* (S. 145–180). Göttingen: Hogrefe.

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.

Van der Molen, H.T., Te Nijenhuis, J. & Keen, G. (1995). The effects of intelligence test preparation. *European Journal of Personality, 9*, 43–56.

Van de Vijver, F.J.R. & Poortinga, Y.H. (1992). Testing in culturally heterogeneous populations: when are cultural loadings undesirable? *European Journal of Psychological Assessment, 8*, 17–24.

Walsh, W.B. & Betz, N.E. (1985). *Tests & assessment*. Englewood Cliffs, NJ: Prentice Hall.

Weinert, A.B. (1987). *Lehrbuch der Organisationspsychologie*. München: Psychologie Verlags Union.

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.

Anschrift des Verfassers: Dipl.-Psych. Martin Kersting, c/o Deutsche Gesellschaft für Personalwesen, Grassistr. 12, 04107 Leipzig.

Eingegangen: 15.10.1995

Revision eingegangen: 20.1.1996