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*, 44, Heft 53, S. 47-98.

Errata:

Im Artikel *"Personalauswahl in den neuen Bundesländern (...)"* muß es auf Seite 85 in der letzten Reihe der Tabelle und auf Seite 86 in den Zeilen 12 und 13 statt *"Anteil valider Positiven an allen Positiven" heißen:*
"Anteil der validen Positiven an allen potentiell Erfolgreichen".

Personalauswahl in den neuen Bundesländern: Tests ohne Grenzen oder Grenzen der Tests? Eine vergleichende Analyse der durchschnittlichen Testergebnisse in Ost und West. *(Teil II)*

Martin Kersting

Übersicht:

1. Einführung

Seit dem Beitritt der DDR zur Bundesrepublik sind nun mehr als drei Jahre vergangen. Mit der Zeit konnten sich die Beschreibung und Analyse der deutsch-deutschen Teilung und des Zusammenwachsens aus ***psychologischer*** Perspektive immer stärker etablieren. Mittlerweile liegt eine große Zahl an Zeitschriften- und Buchartikeln, zum Teil in Form von Schwerpunktsammlungen, zu diesem Thema vor, die -insbesondere soweit sie auf empirischem Material fußen- eine erste Basis für eine fundierte Diskussion und ein verantwortbares Handeln schaffen. Die *Deutsche Gesellschaft für Personalwesen (DGP)* sieht sich der spezifischen Situation der Personalarbeit in den fünf neuen Bundesländern besonders verpflichtet. In bezug auf die Personalauswahl bedeutet dies, daß die *DGP* ihre in den Altbundesländern entwickelten und bewährten Verfahren nicht unreflektiert auf die Beitrittsgebiete ausdehnt, sondern die Anwendung der Verfahren in den neuen Bundesländern wissenschaftlich begleitet und die Verfahren selbst gegebenenfalls modifiziert.

Von den ersten vergleichenden Ergebnissen des Einsatzes der *DGP*-Eignungstests in den neuen und in den alten Ländern handelte der erste Teil dieses Artikels im Heft 52 der *DGP*-Informationen (Kersting,1993). Als Ergebnis eines Vergleichs zwischen den Testleistungen von 1377 Bewerber(inne)n für die Laufbahn des allgemeinen gehobenen Verwaltungsdienstes in den alten und in den neuen Bundesländern wurde konstatiert, daß es statistisch signifikante Leistungsunterschiede zugunsten der westdeutschen Bewerber(innen) gab. Diese Unterschiede betrugen im Durchschnitt beim Vortest 2,6 Einheiten auf der von 70-130 variierenden Leistungsskala und wurden daher als quantitativ eher geringfügig eingestuft. Auch andere Organisationen auf dem Gebiet der Eignungsdiagnostik stellten in entsprechenden Analysen höhere Erwartungswerte in dieser Größenordnung für die Westdeutschen fest. So etwa das Personalstammamt der Bundeswehr bezüglich der Testergebnisse von 1052 ostdeutschen Offiziersanwärtern im Sommer 1992 (persönliche Mitteilung Dr. *Melter*[1]), das Institut für Test- und Begabungsforschung bezüglich der Ergebnisse der ostdeutschen Teilnehmer(innen) am Test für medizinische Studiengänge (1989: ca. 250, 1990: 1438 und 1991: 3305 Personen) (persönliche Mitteilung Dr. *Trost*[1]; Fay, 1991; Hensgen & Blum, 1992; Blum & Hensgen, 1993) sowie das Institut für Wirtschaftspsychologie bezüglich der Ergebnisse von ca. 500 ostdeutschen Bewerber(inne)n für anspruchsvolle Führungsfunktionen im Vertriebsbereich (persönliche Mitteilung *Hübbe*[1]; Stratemann, 1992, S.70 ff.)). Bei allen genannten Untersuchungen handelt es sich um den Vergleich von Leistungen, die -ungeachtet der spezifischen Ausrichtung der Tests- dem Bereich der Intelligenz zuzuordnen sind.

[1]) Für die Informationen und für die Kontaktbereitschaft möchte ich Herrn Dr. *Melter* , Herrn Dr. *Trost* und Herrn Dipl.-Psych. *E.Hübbe* (Institut für Wirtschaftspsychologie) herzlich danken.

Gerade die Diskussion um "Intelligenz" ist aus verschiedenen Gründen heraus besonders belastet (siehe Jäger, 1984, S.21-22). Der Bericht von Intelligenzunterschieden zwischen Gruppen -und seien sie noch so geringfügig- weckt unweigerlich eine Reihe von "Abwehrreaktionen". So wird häufig Validität der vergleichenden Untersuchung in Frage gestellt und/oder es werden Maßnahmen gefordert, die dazu geeignet sind, die Unterschiede "verschwinden" zu lassen. So leitet Stratemann (1994, S.43) aus ihren ost-west-vergleichenden Analysen ab, daß sich *"die Normwerte bei einem Testeinsatz unter Fairness-Gesichtspunkten als inadäquat"* erweisen, ohne daß ihr Artikel einen empirischen Nachweis der Inadäquatheit (etwa interne Indikatoren in Form von vergleichenden Testkennwerten oder externe Indikatoren in Form von vergleichenden Bewährungsdaten) bereithält.

Reaktionen auf den Bericht von Leistungsunterschieden sind oft ideologisch begründet. Die Diskussion wird dann dadurch erschwert, daß die zugrundeliegende ideologische Position implizit bleibt. Eine solche ideologische Position ist beispielsweise das "Identitätskonzept". Konflikte entstehen durch die Diskrepanz zwischen unserem Glauben in die Gleichheit von Individuen und Gruppen und der Tatsache, daß zwischen Einzelnen und Gruppen zahlreiche Unterschiede bestehen (Kong, 1989). Der Nachweis von Gruppenunterschieden in Leistungstestverfahren allein ist aber noch kein Beweis für die Unangemessenheit des Testes. Ebenso wie es ohne weiteres wissenschaftlich keine a priori Setzungen darüber geben kann, daß eine Gruppe in ihren Testleistungen einer anderen Gruppe im Durchschnitt überlegen ist, kann es auch keine a priori Setzung des Leistungsgleichstandes geben (Reynolds und Brown, 1984, S. 24). Jensen (1980) hat diese Art des Mißverständnisses in der Diskussion um mögliche Testverzerrungen als *"egalitarian fallacy"* bezeichnet (S.370).
Ein anderes Mißverständnis, *"standardization fallacy"* bezeichnet (Jensen, 1980, S.372), basiert auf der Annahme, daß ein Test, der von Angehörigen einer bestimmten Gruppe entwickelt und an Stichproben dieser Gruppe "genormt" wurde, *zwangsläufig* Mitglieder anderer Gruppen benachteiligen **muß**. Eine solche Gesetzmäßigkeit gibt es nicht. Jensen (1984, S.516 und 564) nennt zahlreiche Beispiele für Gruppen, die im Durchschnitt bei einem Test höhere Leistungen erzielen als Angehörige der Kultur in der der Test hergestellt und normiert wurde. So übertreffen beispielsweise Japaner(innen) in einem amerikanischen nonverbalen Intelligenztests im Durchschnitt die Testleistungen der Amerikaner(innen). Nicht selten ist dieses Standardisierungs-Mißverständnis auch auf fehlende Kenntnisse über Normierungsprozesse zurückzuführen. Bezieht man bei der Normierung eines Tests leistungsdisperate Gruppen in gleichem Umfang gleichzeitig ein, wird dadurch keinesfalls die relative Distanz der Gruppen zueinander aufgehoben. Lediglich die numerischen "labels" verändern sich. (Vernon, 1979, S. 307). Um etwas anderes handelt es sich bei der Forderung nach ***gruppenspezifischen*** Normen, auf die weiter unten im Text (Punkt 5) noch eingegangen werden wird.
Das dritte Mißverständnis (*"culture-bound fallacy"*; Jensen, 1980, S.371) in der Diskussion um mögliche Testverzerrungen bezieht sich auf die unzulässige Gleichsetzung von "kulturellen Ladungen" von Tests und einer "kulturellen

Verzerrung". Am Beispiel von Intelligenztests wird deutlich, wie schwer es ist, einen *"cultur-fair"* oder einen *"cultur-free"* Test zu konstruieren. Zum einen sind die Vorstellungen über Intelligenz immer sozio-kulturell geprägt (Samuda, 1989, S.33), zum anderen ist es fraglich, ob ein kulturell-"blinder" Intelligenztest in der Lage wäre, intelligentes Verhalten im "wirklichen Leben" vorherzusagen. Intelligentes Verhalten findet ja stets innerhalb eines bestimmten kulturellen Umfeldes statt (Reynolds & Brown, 1984, S.23). Die Tatsache, daß ein gegebener Test einen bestimmten kulturellen Background hat, bedeutet nicht, daß er beim Einsatz für Angehörige eines anderen kulturellen Backgrounds zwangsweise zu kulturell verzerrten Ergebnissen führt. Dies ist -ebenso wie bei den anderen angesprochenen Punkten- eine empirische Frage.

Schlußfolgerungen nach der Maxime *"also schloß er messerscharf / das nicht sein kann / was nicht sein darf"* sind als unangemessen zurückzuweisen. Ebenso wie aus der Tatsache von Gruppenunterschieden allein keine Testverzerrung abgeleitet werden kann, gilt vice versa aber natürlich auch, daß ein Meßinstrument nicht a priori für beliebig viele Gruppen Gültigkeit beanspruchen kann. Der Geltungsanspruch eines Meßinstrumentes zur Personalselektion muß empirisch ausgelotet werden. Nach Schuler und Funke (1989, S.317) gilt ein Verfahren unter dem Aspekt der Fairness *"prima vista als suspekt, wenn der relative Anteil Ausgewählter aus einer unterprivilegierten Gruppe geringer ist als 80 % des Anteils der Nichtminorität"*. Dies würde -wie weiter unten noch genauer ausgeführt wird- für den selten vorkommenden Fall, daß Bewerber(innen) aus West- und Ostdeutschland um ein und dieselbe Stelle konkurrieren, auf die Anwendung der *DGP* -Testverfahren zutreffen.

"Suspekt" wäre das *DGP* -Auswahlverfahren beispielsweise, wenn der Verdacht besteht, daß die nachgewiesene durchschnittliche Unterlegenheit der Bewerber(innen) aus den neuen Bundesländern nicht auf tatsächliche Fähigkeitsunterschiede in der anvisierten Leistungsdimension, sondern auf fehlerhafte oder unangemessene psychometrische Methoden zurückgeführt werden muß. Die Anwendung der Testergebnisse in der Personalauswahl könnte folglich zu einer fehlerhaften Selektion führen. Dieser Verdacht der kulturellen Verzerrung von Testergebnissen kann wissenschaftlich als die statistische Frage nach einer systematischen Fehleinschätzung der wahren Werte spezifischer Gruppen formuliert werden. Dies würde bedeuten, daß die gleichen Testergebnisse für Angehörige verschiedener Gruppen unterschiedliche Bedeutungen hätten. Eine derart statistisch formulierte Frage nach der kulturellen Verzerrung von Testergebnissen läßt sich nicht am grünen Tisch oder anhand bestimmter ideologischer Grundpositionen entscheiden, sondern bedarf der empirischen Forschung. Gesucht wird dabei nach internen und externen Indikatoren für eine solche Testverzerrung. Darüber hinaus gilt es, die Ursachen für die Testverzerrung oder -im Falle einer nachweislich unverzerrten Diagnose- für die Gruppenleistungsunterschiede selbst zu beschreiben.

Im Heft 52 der *"DGP*-Informationen" wurden unterschiedliche Erklärungshypothesen für den Befund der niedrigeren Testerwartungswerte für die Bewerber(innen) aus den neuen Bundesländern referiert. Der vorliegende zweite Teil des Artikels knüpft an den ersten Teil an und liefert empirische Prüfungen zu einigen Hypothesen, die im ersten Teil des Artikels entwickelt wurden. Insbesondere wird nach internen Indikatoren einer möglichen kulturellen Testverzerrung zuungunsten der Neubundesbürger(innen) gesucht. Diese Suche nach internen Indikatoren für eine Kulturverzerrung wird mit einer Prüfung der Faktorenstruktur des Tests eröffnet. Im darauffolgenden Abschnitt wird der Frage nachgegangen, ob die Testergebnisse dadurch kulturell verzerrt wurden, daß die Bewerber(innen) in den neuen Bundesländern trotz der Zeitbegrenzung der Tests langsamer und sorgfältiger arbeiteten. Die geringere Leistungsgüte dieser Personen wäre dann auf eine geringere Geschwindigkeit bei der Testbearbeitung zurückzuführen, die Diagnose entsprechend zu modifizieren. Mit verschiedenen Methoden wird dann drittens der Frage nachgegangen, ob die Leistungsunterschiede eventuell auf eine Überrepräsentation von westdeutschen Kenntnisfragen und/oder auf eine unterschiedliche Wirkung der "westdeutsch" formulierten Items bei "ostdeutschen" Bewerber(inne)n zurückgeführt werden können. Auch zu der Frage, ob es möglicherweise Beurteilereffekte in dem Sinne gegeben hat, daß West-Psycholog(inn)en Ost-Bewerber(innen) bei gleicher Leistung schlechter beurteilen als Ost-Psycholog(inn)en, wird ein empirischer Beitrag geliefert. Der empirische Teil des Artikels bleibt auf die Prüfung potentieller interner Indikatoren einer Testverzerrung begrenzt. Untersuchungen zu externen Indikatoren, d.h. z.B. korrelative Beziehungen der Tests mit testunabhängigen Variablen, liegen nicht vor. Von besonderer Bedeutung sind hier vergleichende Untersuchungen zur prädiktiven Validität. Auch ohne empirische Daten zur Kriteriumsvalidität sollen in einem weiteren Abschnitt einige Probleme der Vorhersage in leistungsdisperaten Gruppen angesprochen werden. Modellrechnungen erlauben eine erste Abschätzung des Effekts der unterschiedlichen Gruppenleistungen auf die Vorhersage. Potentielle Konsequenzen aus den Befunden -z.B. die Einführung von "Ost-Normen"- werden diskutiert. Zum Abschluß des Artikels wird noch einmal eine Hypothese aufgegriffen, die erklären könnte, warum die gemessenen Leistungsunterschiede womöglich ein unverzerrtes Bild tatsächlicher Leistungsunterschiede zwischen den Bewerber(innen) aus den neuen und den alten Bundesländern darstellen. In dieser Hypothese werden die gefundenen Unterschiede in der Testleistung darauf zurückgeführt, daß sich in den neuen Bundesländern z.T. ein anderer, den Westbewerber(inne)n in vielerlei Hinsicht unvergleichbarer Personenkreis bei den Verwaltungen bewirbt. Unter den Stichworten "nicht-kognitive Faktoren", "Selbstselektion" und "Stichproben-Bias" werden Gründe für diese Annahme referiert.
Der vorliegende zweite Teil der Darstellung knüpft unmittelbar an den ersten Teil im vorherigen Heft an und setzt diesen als bekannt und verfügbar voraus. Angaben zur Ausgangsstichprobe, zum Vorgehen der stufenweisen Selektion (Haupttest nach Vortest) und zu den eingesetzten Meßinstrumenten sind diesem Teil der Darstellung zu entnehmen. Auf Wunsch schickt Ihnen Ihr zuständiges *DGP*- Büro gerne einen Abdruck des ersten Teils des Artikels zu.

2. Interne Indikatoren für eine kulturabhängige Testverzerrung

2.1 Faktorenanalyse

Eine Möglichkeit, den Generalitätsanspruch eines Meßinstrumentes abzusichern, besteht im Nachweis übereinstimmender Strukturen in den unterschiedlichen Gruppen. Der Nachweis einer über verschiedene Gruppen hinweg invarianten Struktur stärkt das Vertrauen, daß der Test das, was er mißt, in den verschiedenen Gruppen auf die gleiche Art und Weise mißt und daß es mit einiger Wahrscheinlichkeit in den unterschiedlichen Gruppen das gleiche Konstrukt ist, das gemessen wird (Reynolds und Brown, 1984, S.27).
Eine solche strukturanalytische Überprüfung wurde mit Hilfe der Faktorenanalyse für den Datensatz des Haupttests (Beschreibung der Stichprobe und der Tests, siehe Kersting 1993, S.64) vorgenommen. Im Heft Nr. 52 der "*DGP* Informationen" wurden für den Vergleich der Haupttestergebnisse nur diejenigen 245 Personen berücksichtigt, welche mindestens 90 Punkte im Vortest erzielt hatten. Diese Einschränkung war notwendig, da für die neuen und alten Bundesländer eine unterschiedliche Zulassungspraxis zum Haupttest nachgewiesen werden konnte. In den neuen Bundesländern wurden auch solche Personen zum Haupttest eingeladen, die sich im Vortest bereits als leistungsschwächer erwiesen hatten. Für die Faktorenanalyse wurden nun alle 548 zum Haupttest zugelassenen Personen berücksichtigt, da ja gerade die Struktur unterschiedlich leistungsstarker Gruppen miteinander verglichen werden sollte. Die Vortestdaten eignen sich aufgrund der geringen Zahl an Tests wenig für faktorenanalytische Prüfungen.

Das Ergebnis einer Faktorenanalyse liefert Informationen darüber, welche Variablen (in diesem Falle die einzelnen Tests) gemeinsame und welche Variablen verschiedene Informationen erfassen. (Zur Faktorenanalyse siehe z.B. Bortz, 1989, S. 615-682.) Tabelle 1 zeigt die Ergebnisse einer Faktorenanalyse (Hauptachsenanalyse) mit Kommunalitäten Iteration und anschließender Varimax-Rotation. Die Faktorenzahl wurde aufgrund des Eigenwertverlaufes auf drei begrenzt. Die Berücksichtigung eines vierten Faktors führt zu einer "overextraction" und zur Herausbildung eines schwer interpretierbaren Singulärfaktors, der im wesentlichen auf das Diktat zurückgeführt werden kann. Faktoren sind "synthetische", wechselseitig voneinander unabhängige Variablen, die die Zusammenhänge zwischen den einzelnen Variablen (hier: den einzelnen Tests) ordnen. Sie bilden eine Grundlage der Datenreduktion, der Skalenbildung oder der Zusammenfassung einzelner Ergebnisse. So geht die *DGP* etwa davon aus, daß es den Tests "Ähnliche Wortbedeutungen", "Analogien", "Schlußfolgerungen" und "Textanalyse" gemeinsam ist, Anforderungen an die intellektuelle Verarbeitungskapazität im Umgang mit sprachlichem Material zu stellen. Die in der Tabelle 1 abgedruckten Resultate der Faktorenanalyse rechtfertigen die identische Vorgehensweise bei der Auswertung und Interpretation der Testergebnisse in den neuen und in den alten Bundesländern. In beiden Gruppen lassen sich das numerische Denken, das sprachliche Denken und das Arbeitsverhalten als Faktoren identifizieren.

Lediglich in der Weststichprobe läßt der Verbalfaktor Prägnanz vermissen. Obwohl auf die Berechnung von Ähnlichkeitskoeffizienten zum Ver-gleich der Faktorenstrukturen verzichtet wurde, scheint die Inspektion der Faktorenladungen die Annahme der weitgehenden Invarianz der Faktorenstruktur über die beiden Stichproben zu rechtfertigen.

Tabelle 1: Faktorenanalyse Haupttest. Hauptachsenanalyse mit Kommunalitäten-iteration und anschließender Varimax Rotation (Vorgabe: Drei Faktoren)

	Gesamt (N=548)				Ost (N=265)				West (N=283)			
	Faktor				Faktor				Faktor			
Test	num.	Arb.	verb.	h^2	num.	Arb.	verb.	h^2	num.	Arb.	verb.	h^2
ÄW	-.19	.05	**.37**	.18	-.24	.07	**.46**	.27	-.12	.01	**.26**	.08
AG	.07	.04	**.44**	.20	.06	-.03	**.52**	.27	.11	.06	**.30**	.11
SL	.20	.04	.34	.16	.17	.07	.47	.26	.25	-.01	**.17**	.09
TA	.11	.10	**.49**	.26	.10	.20	.47	**.27**	.17	-.05	**.41**	.20
VB	**.30**	-.04	-.03	.09	**.32**	-.04	-.04	.11	**.33**	.02	-.04	.11
ZZ	**.43**	.18	-.07	.22	**.32**	.15	-.11	.14	**.50**	.16	-.11	.29
TX	**.66**	-.04	.16	.46	**.62**	-.07	.13	.40	**.69**	-.06	.10	.50
ES	**.52**	.14	.14	.31	**.52**	.11	.15	.30	**.51**	.10	.09	.28
TS	**.56**	.04	.19	.35	**.53**	.07	.20	.33	**.62**	.04	.12	.39
KT	-.12	**.70**	.10	.52	-.13	**.76**	.05	.59	-.12	**.61**	.12	.40
CA	.22	**.56**	.04	.36	.15	**.54**	.14	.34	.26	**.61**	-.09	.44
PA	.30	**.53**	.06	.38	.25	**.56**	.02	.37	.36	**.50**	.00	.38
Diktat	-.12	.22	.30	.15	-.07	.25	.28	.15	-.13	.17	.39	.20
GR	.54	.14	.03	.32	.56	.18	.01	.35	.54	.10	-.02	.30
Gem-K	.09	-.07	.26	.08	.16	-.06	.19	.07	.05	-.09	.31	.11

Legende:
num. - Verarbeitungskapazität / numerisches Aufgabenmaterial
verb. - Verarbeitungskapazität / verbales Aufgabenmaterial
Arb. - Arbeitsverhalten
ÄW - Wortbedeutungen
AG - Analogien
SL - Schlußfolgerungen
TA - Textanalyse
VB - Verschiedene Beziehhungen
ZZ - Zahlenmatrizen
TX - Textrechnen
ES - Ergebnisse schätzen
TS - Tabellen & Statistiken
KT - Kontrolltest
CA - Computerausdruck
PA - Postaufgabe
GR - Grundrechnen
Gem-K - Gemeinschaftskunde

2.2 Stichprobe und Meßinstrumente für weitere Analysen

Für die unter Punkt 2.3 aufgeführten Analysen zum Einfluß der Testbearbeitungsgeschwindigkeit und für die in den Punkten 2.4.1 und 2.4.3 dargestellten Analysen zur Möglichkeit von kulturabhängigen Itemverzerrungen mußten die in Teil I dieses Artikels dargestellten Ergebnisse einer weiteren Auswertung unterzogen werden. Wurden für die ersten Auswertungsschritte lediglich Angaben über die auf der Anzahl der richtigen Lösungen pro Aufgabe basierenden "Standardwerte" benötigt, mußten nun alle Antworten der Bewerber(innen) zu jeder einzelnen Frage kodiert und ausgewertet werden (Itemanalyse). In diese weitergehenden Analysen wurden nur ein Teil der ursprünglichen Stichprobe von 1377 und nur ein Teil der Aufgaben einbezogen. Es handelt sich dabei um eine in bezug auf wesentliche demographische Merkmale (Geschlecht / Alter) und in bezug auf die zentralen Leistungsvariablen zur Vorteststichprobe weitgehend parallelisierte Teilstichprobe. Der Umfang der Teilstichprobe beträgt 500 Bewerber(innen) aus dem Vortest. 202 Personen aus dieser Gruppe nahmen am Haupttest teil. Demographische Angaben zu der Vortestteilstichprobe können der Tabelle 2 entnommen werden. Neben den absoluten Häufigkeiten sind dort zusätzlich noch die Tabellenprozente für die jeweiligen Subbereiche der Tabelle angegeben.

Von den neun Vortestaufgaben wurden fünf in die Itemanalyse einbezogen. Bezüglich des gedanklichen Umgangs mit sprachlichem Material wurden beide Aufgaben aus dem Vortest, nämlich die "Ähnlichen Wortbedeutungen" und die "Analogien", auf Itemebene analysiert. Aus dem Bereich des gedanklichen Umgangs mit Zahlen wurden die "Textrechenaufgaben" und die "Zahlenmatrizen" für die Analyse berücksichtigt. Schließlich wurden noch die Antworten auf die "gemeinschaftskundlichen Wissensfragen" im Detail untersucht.

2.3 Zum Einfluß der Testbearbeitungsgeschwindigkeit auf die Leistungsgüte

In der Psychometrie unterscheidet man sogenannte *"speed"* von sogenannten *"power"* Tests. Im Psychologischen Wörterbuch von Dorsch (1982) sind "*speed* Tests" erläutert als *"Tests, bei denen unter Wegfall einer Zeitbegrenzung sämtliche Aufgaben von allen Pbn gelöst bzw. richtig beantwortet werden"* (S.636). Demgegenüber werden "*power* Tests" als *"Tests der Leistungshöhe"* beschrieben, *"bei denen eine möglichst hochwertige Lösung gefordert wird. (...) Die Tests haben entweder keine oder eine großzügig bemessene Zeitbegrenzung"* (S. 499). (Zur genaueren Begriffsbestimmung und zur Kritik am typologischen Beschreibungsansatz siehe Eberle, 1980.) Die in der Praxis verwendeten Tests siedeln sich auf diesem, durch die beiden extremen Pole *"speed"* und *"power"* aufgespannten, Kontinuum an. Die Tests werden mit einer zeitlichen Begrenzung dargeboten, aber die Zeit ist -mit Ausnahme expliziter Messungen der Bearbeitungsgeschwindigkeit- so reichlich

Tab. 2: Demographische Daten zur Stichprobe, N = 500

		OST (51 %)			WEST (49 %)			GESAMT, N = 500		
		♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂
Anzahl		142 55,7 %	113 44,3 %	255	133 54,3 %	112 45,7 %	245	275 55 %	225 45 %	500
Alter	Median	18	20	19	20	22	20	19	21	20
	Mittelwert	19,5	22,1	20,6	20,3	23,5	21,8	19,9	22,8	21,2
SCHULE	EOS / Gymnas.	114 44,7 %	67 26,3 %	181 71 %	98 40 %	61 24,9 %	159 64,9 %	212 42,4 %	128 25,6 %	340 68 %
	Fach(hoch) -schule	12 4,7 %	13 5,1 %	25 9,8 %	35 14,3 %	51 20,8 %	86 35,1 %	47 9,4 %	64 12,8 %	111 22,2 %
	BMA	15 5,9 %	29 11,4 %	44 17,3 %				15 3 %	29 5,8 %	44 8,8 %
	Sonstige	1 0,4 %	4 1,6 %	5 2,0 %				1 0,2 %	4 0,8 %	5 1,0 %
HERKUNFT	Brandenburg	80 31,4 %	62 24,3 %	142 55,7 %				80 16 %	62 12,4 %	142 28,4 %
	Mecklenburg-V.	31 12,2 %	17 6,7 %	48 18,8 %				31 6,2 %	17 3,4 %	48 9,6 %
	Sachsen-Anhalt	31 12,2 %	34 13,3 %	65 25,5 %				31 6,2 %	34 6,8 %	65 13,0 %
	Niedersachsen				70 28,6 %	61 24,9 %	131 53,5 %	70 14,0 %	61 12,2 %	131 26,2 %
	Stadt Hannover				63 25,7 %	51 20,8 %	114 46,5 %	63 12,6 %	51 10,2 %	114 22,8 %

bemessen, daß die Mehrheit der Testteilnehmer(innen) ihre persönliche Leistungsgrenze erreicht. Die Bestimmung der Laufzeit eines Tests wird durch aufwendige Variation der Laufzeiten einer Aufgabe und anschließende Ergebnisanalysen festgelegt. Als Kriterium einer vertretbaren Zeitbegrenzung kann z.B. gelten, daß sich die Leistungs**rang**folge der getesteten Personen bei einer deutlichen Verlängerung der Zeitvorgaben nicht verändert (die nominelle Leistung pro Person kann dabei vernachlässigt werden). Die Leistungsgeschwindigkeit und ihr Verhältnis zur Leistungsgüte stellt ein grundsätzliches Problem der Psychometrie dar (siehe z.B. Iseler, 1970). Zu einem Problem der kulturbedingten Testverzerrung kann die Testbearbeitungsgeschwindigkeit dann werden, wenn sie in unterschiedlichen Gruppen unterschiedlich ausgeprägt ist. Die Laufzeiten der *DGP* -Tests wurden für Weststichproben entwickelt und kontrolliert. Der durchschnittliche Leistungsvorteil der Bewerber(innen) aus den alten Bundesländern könnte nun so interpretiert werden, daß die Zeitbegrenzung der Tests möglicherweise die schnell und oberflächlich arbeitenden Westdeutschen gegenüber den gründlich und zuverlässig arbeitenden Ostdeutschen in Vorteil setzt. Die nachgewiesenen Leistungsunterschiede würden dann Unterschiede in der Testbearbeitungsgeschwindigkeit oder allgemein eine unterschiedliche Vertrautheit mit Antwortprozeduren (siehe van de Vijver und Poortinga, 1992, S. 19) und nicht Unterschiede in der intellektuellen Verarbeitungskapazität reflektieren.

Die Annahme von Ost-West-Unterschieden in der Testbearbeitungsgeschwindigkeit ist keineswegs abwegig. Die Geschwindigkeit in der Testbearbeitung ist vielfältig determiniert, wobei neben Persönlichkeitsdimensionen wie z.B. der Motivation, auch die Kultur als eine potentielle Determinante diskutiert wird (Iseler, 1970, S.234). Gerade die Vorstellung und Bedeutung von "Zeit" und "Geschwindigkeit" dürften über die Kulturen differieren, und gerade hinsichtlich der "Zeit" dürfte die ehemalige DDR ein wohlhabenderes Land gewesen sein als ihr hektischer westlicher Pedant. Helms betont die Gefahr, daß ein unreflektiertes Zugrundelegen von Werten (wie z.B. Wertvorstellungen über Zeit) zu einem "Eurozentrismus" bei kognitiven Leistungstests führen könnte. Darunter versteht sie *"a perceptual set in which European and/ or European American values, customs, traditions and characteristics are used as exclusive standards against which people and events in the world are evaluated and perceived"* (Helms, 1989, zitiert nach Helms, 1992, S. 1093).

Ein weiterer Faktor, der zu einem unterschiedlichen Erleben und Umgang mit der Zeitgrenze der Tests führen könnte, ist die möglicherweise in den Gruppen unterschiedlich ausgeprägte (Test-)angst (zur "Angst" siehe weiter unten, Punkt 4.).

Darüber hinaus spielt beim Umgang mit der Zeitbegrenzung von Tests auch die Testerfahrung eine wesentliche Rolle. Gerade durch die Zeitbegrenzung werden Tests anfällig für Trainingseffekte (Hartigan & Wigdor, 1989, S. 282). Die *DGP* fordert die Bewerber(innen) bei der Testinstruktion zwar ausdrücklich dazu auf, sich bei Aufgaben, die einem besonders schwerfallen, nicht "festzubeißen", sondern zur nächsten Aufgabe überzugehen. Dennoch dürfte die rasche Testbearbeitung bei testerfahrenen Personen verbreiteter sein als bei testunerfahrenen Personen. Bewerber(innen) mit Testerfahrung oder Testvorbereitung könnten ihre Leistung erhöhen, indem sie sich bei der Testbearbeitung zunächst in einem ersten oberflächlichen "Schnelldurchgang" ihrem "ersten Eindruck" entsprechend und unter Ausnutzung der Ratewahrscheinlichkeit für eine Lösung entscheiden. Die Bewerber(innen) aus den neuen Bundesländern sind vermutlich im Durchschnitt weniger testerfahren als ihr Gegenüber aus den Altbundesländern. Zwar gab es auch in der ehemaligen DDR Leistungstests, sie fanden aber als *">Überbleibsel der bürgerlichen Psychologie<"* charakterisiert infolge des *"Testverbots in der UdSSR von 1936, dem sog. Pädologie-Dekret der ZK der KPdSU"* eine ungleich geringere Anwendung (Ettrich & Guthke, 1991, S. 18). Dabei konnten sich mit den sogenannten "Olympiaden" z.B. für Mathematik und Russisch andere Selektionssysteme etablieren (ebd, S. 33).

Die vorgenommenen Itemanalysen erlauben eine erste Untersuchung möglicher kulturabhängiger Effekte der Zeitbegrenzung der Tests auf die durchschnittliche Leistung in Ost und West. Ein Problem liegt dabei darin begründet, daß man einer auf dem Antwortbogen nicht ausgefüllten Antwortzeile nicht ansieht, ob die getestete Person diese Frage ausgelassen hat, nachdem sie sie begutachtet hat (vielleicht war die Frage der Person zu schwer), oder ob sie überhaupt nicht zur Bearbeitung der Aufgabe gekommen ist. Diese unterschiedlichen Fälle werden in der Folge nicht mehr unterschieden. Die Untersuchung von Laufzeiteneffekten setzt eigentlich eine systematische Variation der Laufzeiten voraus, die hier nicht geleistet wurde. Die hier vorgestellten Analysen sind somit als Behelfslösung anzusehen.

Der Vergleich der durchschnittlichen Testbearbeitungshäufigkeit von Bewerber(inne)n aus den alten und aus den neuen Bundesländern läßt sich am Beispiel der Aufgabe "Zahlenmatrizen" veranschaulichen. Diese Aufgabe stellt eine Anforderung an das logische Denkvermögen dar. Dabei geht es darum, die Regeln zu erkennen, nach denen Zahlenprogressionen weiterzuführen sind. Auf der Abzisse des Flächendiagramms in Abbildung 1 sind die Items entsprechend der Darbietungsfolge abgetragen. Da sich die Reihenfolge bei verschiedenen Testformen unterscheidet, kann von der vorliegenden Graphik nicht auf ein bestimmtes inhaltliches Item geschlossen werden. Vom Iteminhalt her betrachtet, steht das Item an Position 3 in Form "1", in Testform "2" vielleicht an Position 5 usw. Es geht in dieser Graphik nur um Effekte der Itemreihung. Fragen nach möglichen inhaltlichen Problemen, die mit den einzelnen Items in unterschiedlichen Kulturen verbunden sein könnten, werden im nächsten Abschnitt behandelt (Punkt 2.4.1). Auf der Ordinate ist die Bearbeitungshäufigkeit in Prozent abgetragen, die schwarze Fläche im Hintergrund verzeichnet die Werte für die West-Gruppe, die schraffierte Fläche im Vordergrund steht für die Ostgruppe. Das an siebter Position stehende Item der Aufgabe Zahlenmatrizen wurde also beispielsweise von 91 % aller West-, aber nur von 84 % aller Ostdeutschen bearbeitet. Das entsprechend der Darbietungsfolge auf Position "dreizehn" stehende Item wurde nur noch von knapp 55 % der Bewerber(innen) aus den alten und nur noch von knapp 43 % der Bewerber(innen) aus den neuen Bundesländern bearbeitet usw. Dabei bedeutet "bearbeitet", daß bei diesem Item auf dem Antwortbogen von den Bewerber(inne)n irgendetwas angekreuzt wurde, egal ob diese "Lösung" falsch oder richtig war.

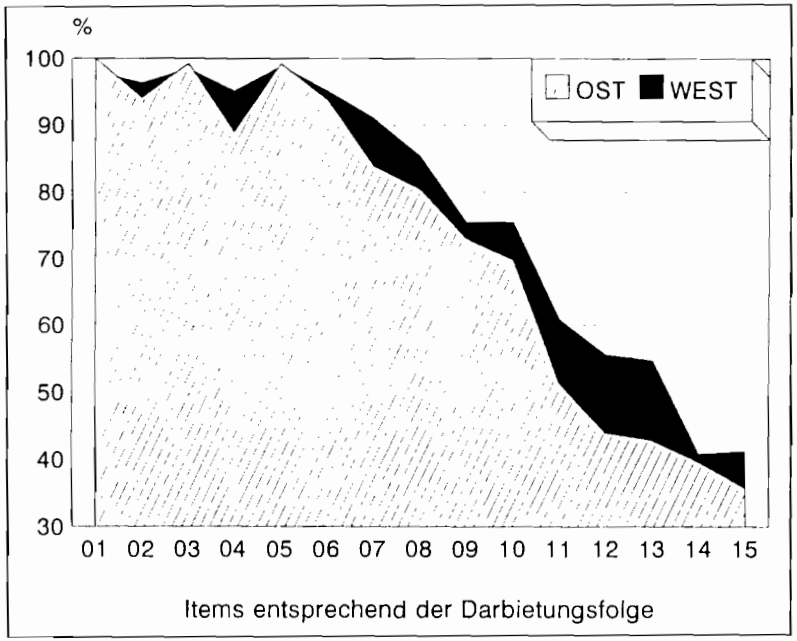


Abb. 1: Bearbeitungshäufigkeit (in Prozent) pro Gruppe; Aufgabe: ZZ

Analog zu dieser Beschreibung wurde nun für jede der fünf Aufgaben die durchschnittliche Bearbeitungshäufigkeit pro Gruppe (Alte versus Neue Bundesländer) als einfaches Aggregat über alle einzelnen Items der jeweiligen Aufgabe bestimmt. Aufgrund der extrem schiefen Verteilungen (siehe beispielhaft Abbildung 1) sind die Verteilungskennwerte "Mittelwert" oder "Median" für den Ost-West-Vergleich der durchschnittlichen Bearbeitungshäufigkeiten unbrauchbar. Mit dem U-Test von Mann-Whitney wurde ein parameterfreies Prüfungsverfahren gewählt. Vier der fünf Überprüfungen erwiesen sich auf dem Ein-Prozent-Niveau als statistisch signifikant, bei der sprachlichen Aufgabe "ähnliche Wortbedeutungen" ergab sich nur ein borderline Effekt auf dem 10 % Niveau. Grundsätzlich bearbeiten die "Wessis" in der vorgegebenen Zeit also mehr Items als die "Ossis". Allerdings handelte es sich auch hier um quantitativ eher geringfügige Unterschiede. Da angesichts der Verteilungen weder der Mittelwert noch der Median sinnvoll interpretierbar sind, kann zur Quantifizierung nur die durchschnittliche Differenz im mittleren Rang angegeben werden. Sie betrug über die fünf Aufgaben gemittelt 45,3. Ein anschaulicherer Wert ergibt sich durch die Bestimmung der Anzahl an Bewerber(inne)n aus den neuen Bundesländern, die unter dem für einen "power Test" als magische Grenze definierten Wert von 90 % bearbeiteter Items bleiben. (Zum Kriterium "90 %" siehe z.B. Hartigan et al., S. 101.) Von je

hundert "Ost-" und "Westdeutschen" bleiben im Durchschnitt über die fünf analysierten Aufgaben hinweg sechs "Ostdeutsche" mehr als "Westdeutsche" unter der 90 % Hürde. Interessant ist auch das Ankreuzverhalten bei dem eingesetzten Kenntnistest. Neben den üblichen Distraktoren gibt es hier auch die Möglichkeit, *"ich weiß nicht"*, anzukreuzen und die Bewerber(innen) werden gebeten, bei Unkenntnis diese Alternative zu wählen. Unter der Voraussetzung, daß die Kenntnisse in beiden Gruppen gleich sind, kann die Wahl dieser Antwortalternative als eine Art "Test-compliance", als ein Indikator für die Instruktionsfolgsamkeit der Bewerber(innen) gewertet werden. Innerhalb der Gruppe von Bewerber(inne)n, bei denen diese Antwortkategorie überhaupt eine größere Rolle spielt (definiert als die Gruppe aller Personen, die bei mehr als 10% der Fragen diese Antwortkategorie "ich weiß nicht" wählten), waren die Personen aus den neuen Bundesländern 3,6 mal so häufig vertreten wie Personen aus den alten Bundesländern.

Die in Ost und West geringfügig unterschiedliche Bearbeitungshäufigkeit "stört" natürlich den diagnostischen Prozeß. Als Folge dieser unterschiedlichen Bearbeitungshäufigkeiten können sich neben Reliabilitätseinschränkungen auch gruppenspezifische Verzerrungen in der Leistungsmessung ergeben. Um diese Effekte zu kontrollieren, wurde nun eine nachträgliche Auswertung vorgenommen, bei der die Anzahl der Rohwertpunkte auf die Anzahl der überhaupt bearbeiteten Items relativiert wurde. Diese Auswertung berücksichtigt nun nur noch die "Trefferquote" -unabhängig von der Anzahl der bearbeiteten Items. Eine Person, die von 12 bearbeiteten Items sechs richtig beantwortet hat, bekommt dieser nachträglichen Auswertung der Trefferquote zufolge ebenso den Wert von 50 % wie eine Person, die zwar insgesamt neun richtige Antworten gegeben hat, aber auch 18 Items bearbeitet hatte.

Die Effekte sollen hier wieder am Beispiel der Aufgabe "Zahlenmatrizen" veranschaulicht werden. Die beiden linken Balken der Abbildung 2 zeigen zunächst die konventionelle Auszählung der Rohwerte (Anzahl der richtigen Antworten, abgetragen in Prozent des maximal möglichen Wertes) für die Ostgruppe (schraffierter Balken) und die Westgruppe (schwarzer Balken). Wendet man nun statt dessen die beschriebene Auswertungsmethode der "Trefferquote" an (die beiden rechten Balken in Abbildung 2) und eliminiert somit den Effekt der unterschiedlichen Bearbeitungshäufigkeit, so wird der Ost-West-Unterschied zwar geringer, bleibt aber nach wie vor statistisch signifikant. Dabei ergibt sich bei der hier ausgewählten Aufgabe der Zahlenmatrizen insgesamt noch der größte Effekt. Tabelle 3 zeigt, daß bei allen fünf Aufgaben der Ost-West- Unterschied zugunsten der Westgruppe auch bei dieser nachträglich angewandten Auswertungsstrategie statistisch signifikant blieb. Zum Vergleich sind in dieser Tabelle auch die Ergebnisse der entsprechenden statistischen Tests für den Vergleich der "Rohwerte" angegeben. Man kann also konstatieren, daß die in Ost und West unterschiedliche Testbearbeitungshäufigkeit zwar zu den Leistungsunterschieden beiträgt, diese aber nicht vollständig erklärt.

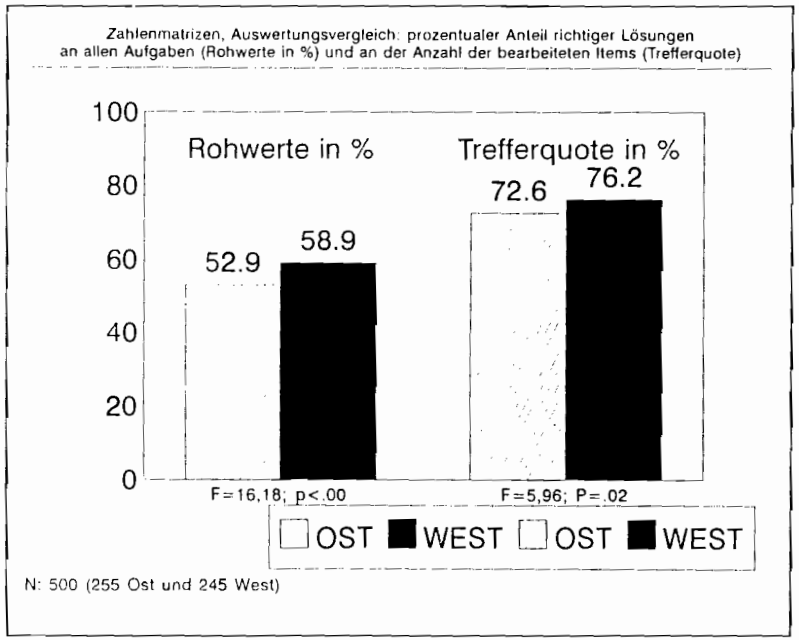


Abb 2: Vergleich verschiedener Auswertungsmodi; Aufgabe ZZ

Test	Auswertung	Ost	West	F	p
ÄW	Anzahl korrekter Lösungen	11.69	13.03	28.69	<.00
	Trefferquote in Prozent	62.69	67.52	16.27	<.00
AG	Anzahl korrekter Lösungen	15.89	16.56	6.76	.01
	Trefferquote in Prozent	72.04	74.78	6.76	.01
ZZ	Anzahl korrekter Lösungen	7.93	8.84	16.18	<.00
	Trefferquote in Prozent	72.62	76.26	5.96	.02
TX	Anzahl korrekter Lösungen	6.80	7.87	20.45	<.00
	Trefferquote in Prozent	65.28	70.36	10.40	<.00
Gem.	Anzahl korrekter Lösungen	26.73	30.43	53.27	<.00
	Trefferquote in Prozent	67.82	76.50	47.88	<.00

Tabelle 3: Vergleich verschiedener Auswertungsmodi; Legende siehe Tab.1

2.4 Indikatoren für kulturbedingte Itemverzerrungen

Es ist denkbar, daß einzelne Items in kulturell unterschiedlichen Gruppen unterschiedliche Wirkungen provozieren. Jensen (1980, S. 432) benutzt hierfür den Ausdruck *"groups x items interaction"*. Als potentielle Ursachen für kulturabhängige Itemverzerrungen 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 der Minderheitengruppe unvertraut ist. Diese Unverständlichkeit der Fragestellung könnte dann die Nennung der Lösung verhindern, obwohl diese Lösung vielleicht durchaus bekannt war.

Auf die DDR bezogene Beispiele für die erstgenannte Kategorie liefern Ettrich und Guthke (1991, S. 17) *"Zum Beispiel erwies sich die Frage im HAWIK* (Erläuterung: HAWIK = Hamburg-Wechsler Intelligenztest für Kinder) *> 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 meist ohne Religionsunterricht erzogenen Kinder konnten auch die andere HAWIK-Frage: > Warum feiert man Ostern? < nicht beantworten. Andere Items wirkten -bezogen auf die bis 1989 obwaltenden gesellschaftlichen Umstände- absurd oder verhöhnend (z.B. wenn in einem Interessentest gefragt wird, ob man lieber eine Reise nach Italien organisieren oder lieber Gebrauchtwagen verkaufen möchte?)"*. Die Autoren nennen zugleich auch ein interessantes Beispiel für die Schwierigkeit, Tests an den gegebenen kulturellen Rahmen anzupassen. *"Beispielsweise wurde im Peabody Picture Vocabulary Test (PPVT) der > Sheriffstern < durch das > Pionierabzeichen < ersetzt, mit der Folge, daß kaum noch ein Vorschulkind das Item richtig löste, weil es zwar im > West-fernsehen < schon den Sheriffstern kennengelernt hatte, aber noch nicht mit dem erst für die Schulzeit relevanten Pionierabzeichen nähere Bekanntschaft gemacht hatte."* (Ebd. S. 16.)

Beispiele für die zweite Gruppe von Itemverzerrungen sind schwerer zu finden. Helms (1992, S.1087) referiert ein umstrittenes Beispiel, welches ebenfalls aus dem Wechsler Intelligenztest für Kinder stammt. Einigen Autoren zufolge sollen schwarze Kinder auf die Frage *"What is the thing to do if a fellow (girl) much smaller than yourself starts to fight with you?"* aus ihren Lernerfahrungen heraus typischerweise die falsche Antwort *"hit him back"* wählen. Andere Autoren weisen diese Behauptung dem Referat von Helms zufolge aber aufgrund empirischer Untersuchungen zurück. Die Autorin liefert auf S. 1097 in ihrem Artikel ein weiteres Beispiel für diese Kategorie. In diesem Beispiel geht es um die in unterschiedlichen Kulturen unterschiedlich hoch bewertete "Kreativität", die Angehörige einer bestimmten Kultur im Test zu Lösungen verleiten kann, die in der Kultur der Testkonstrukteur(inn)e(n) als "falsch" angesehen werden.

Die letzte Kategorie von kulturabhängigen Itemverzerrungen kann schließlich anhand eines Beispiels von van de Vijver et al. (1992, S. 19) veranschaulicht werden. In diesem Artikel wird von einer Überprüfung der Fähigkeit zur Ein-

schätzung von konstanten Mengen in unterschiedlich voluminösen Behältern berichtet, die bei Kindern eines bestimmten Volkstammes eher ungewöhnliche Ergebnisse erbrachte. Diese Ergebnisse waren möglicherweise darauf zurückzuführen, daß das in der Fragestellung verwendete Wort *"more"* in der Sprache der getesteten Kinder mehrdeutig ist.

Für die hier interessierende Suche nach internen Indikatoren für kulturverzerrende Itemeffekte bei einem gemeinsamen Test für Bewerber(innen) aus den neuen und aus den alten Bundesländern dürften vor allem die aus der Sozialisation (z.B. unterschiedliche Curricula) heraus in den Gruppen eventuell unterschiedlich entwickelten Kenntnisse von Bedeutung sein.
Aber auch die Wortbedeutungen und die Sprachgestaltung innerhalb des vereinigten Deutschlands könnten sich über die 40 Jahre der Trennung hinweg zumindest in Nuancen verändert haben. Die Wörter *"Broiler"*, *"Sättigungsbeilage"* und *"Transparenzpudding"* waren hüben ebenso fremd wie drüben die Verwendung des Wortes *"Teil"* in der Werbung *"jedes Teil fünf Mark"* oder die Anrede *"sehr geehrte"* statt dem vertrauten *"werte"*. Es ist also die Frage, ob die Tests in beiden Teilpopulationen Deutschlands "linguistisch äquivalent" (siehe z.B. Helms, 1992, S. 1092) sind. Diese Frage soll zunächst mit Hilfe der Itemanalysen der beiden sprachlichen Tests geklärt werden. In einem nächsten Schritt wird dann über eine Untersuchung mit einer Aufgabe berichtet, die explizit "ostdeutsches Sprachmaterial" verwendete. Abschließend wird dann die Kulturspezifität der Kenntnisfragen thematisiert, indem über den Effekt einer aus kultureller Perspektive vorgenommenen Modifikation berichtet wird.

2.4.1 Itemanalysen der beiden Tests zum sprachlichen Denken

Abbildung 3 repräsentiert die Interaktionen zwischen der Gruppenzugehörigkeit ("Ost" = gestrichelte Linie und "West" = durchgezogene Linie) und den Items. Auf der Ordinate ist die Itemschwierigkeit (Prozentsatz richtiger Lösungen des Items in der Gruppe) abgetragen, die Bezeichnung des Items findet sich auf der Abzisse. Diesmal sind die Items entsprechend ihres Inhaltes bezeichnet, wobei sich die Numerierung an der Testform "1" orientiert. Mit "Item Nummer 1" ist bei der Aufgabe "Ähnliche Wortbedeutungen" also beispielsweise das in der Testform "1" an erster Position stehende Item gemeint, ungeachtet dessen, daß dieses Item in Testform "2" an zwölfter Position steht. Die nachstehenden Analysen können also nicht unter dem Gesichtspunkt der Darbietungsfolge ausgewertet werden.

Den Graphiken läßt sich entnehmen, daß die Items in bezug auf ihre Schwierigkeit in den beiden Gruppen eine unterschiedliche Rangordnung aufweisen. Schaut man sich z.B. den Rangplatz an, den das Item Nr. 13 der Analogieaufgaben in der jeweiligen Schwierigkeitshierachie der Ost- und Westgruppe belegt, so gibt es hier beispielsweise neun Items, die den Bewerber(inne)n aus den neuen Bundesländern im Durchschnitt relativ gesehen mehr Schwierigkeiten bereiten als dieses Item Nr. 13.

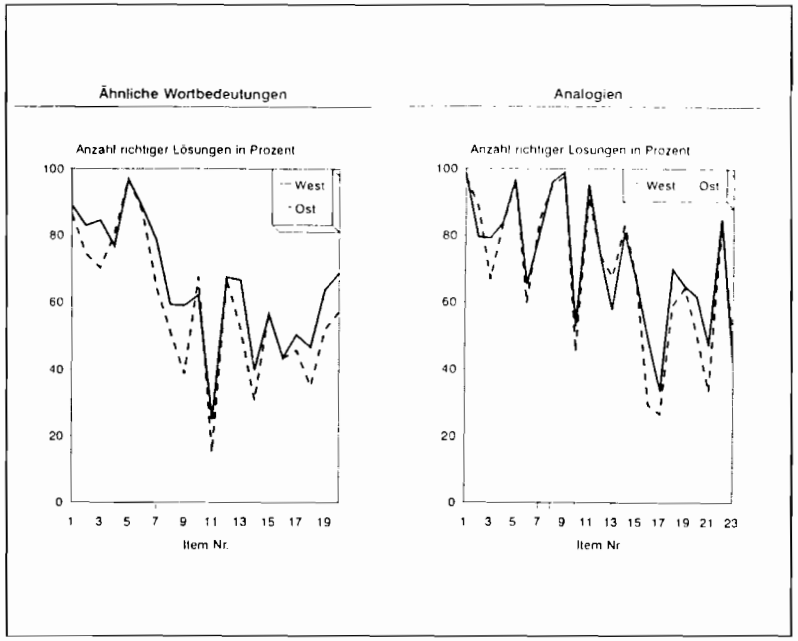


Abb 3.: Interaktionen zwischen Items und der Gruppenzugehörigkeit; N = 500

Item Nr. 13 steht in der Schwierigkeitshierarchie für die Bewerber(innen) aus den neuen Bundesländern insgesamt auf einen mittlerem Rangplatz. Demgegenüber fällt dieses Item den getesteten Altbundesbürger(inne)n vergleichsweise schwerer. In dieser Gruppe übertreffen nur fünf andere Items dieses Item Nr. 13 an Schwierigkeit. Andererseits kommt etlichen Items beider Subtests der identische Rangplatz in beiden Gruppen zu, bei anderen Items ereignen sich nur minimale Rangplatzverschiebungen. Als ein Indikator für das tatsächliche Ausmaß an Interaktionen zwischen Gruppen und Items läßt sich die Korrelation zwischen den Schwierigkeitsrangreihen der Items pro Gruppe (ungeachtet ihrer absoluten Höhe) bestimmen (der Schwierigkeitswert pro Item wird zunächst innerhalb der Gruppen in einen Rangwert transformiert). Der Rangkorrelationskoeffizient betrug für den Subtest "Ähnliche Wortbedeutungen" .93 und für den Subtest "Analogien" .94. Dieser Wert kann als "befriedigend" angesehen werden, wenn man bedenkt, daß auch für eine zufällige Aufteilung innerhalb einer Gruppe oder für eine Aufteilung nach anderen Kriterien, z.B. "Frauen und Männern", keine perfekte Übereinstimmung der Rangfolgen zu erwarten ist. Insgesamt sind Analysen auf der Ebene von Single-Items nur mit Zurückhaltung zu interpretieren, da sich gerade bei solchen einmaligen Messungen Zufallseffekte etablieren können. Gerade deshalb interpretiert ja kein(e) verantwortungsbewußt(e)r Psychologe-

(in) die Antworten auf einzelne Items, sondern verläßt sich nur auf die zuverlässigeren Skalen- oder noch besser: Faktorwerte.[2] Einzelne Items stellen stets eine äußerst unsichere Datenbasis dar. Es bleibt also abzuwarten, ob die sich hier andeutenden geringen kulturellen Unterschiede der Itemstimuli sich an anderen Stichproben erneut aufzeigen lassen, bevor man die Befunde als gesichert ansieht und Konsequenzen daraus zieht. Interpretationen sind solange als vordergründig zu werten, Interpret(inn)en müssen ihre Phantasie zwischenzeitlich zügeln. Dies mag angesichts der sich hier andeutenden Befunde, daß es den Bewerber(inne)n aus den alten Bundesländern z.B. vergleichsweise schwerer fällt als ihrer im real existierenden Sozialismus großgewordenen Kontrastgruppe, die genaue Bedeutung des Wortes *"geborgen"* aus fünf anderen Wörtern herauszufinden (Item Nr.4); oder daß die Analogie Nr. 22 *"Arbeit verhält sich zu Werkzeug wie Musik zu ..."* gerade in den neuen Bundesländern relativ gesehen seltener aufgeht, natürlich manchmal schwerfallen. Solchen "Entdeckungen" kann aber ohne die notwendige Replikation nur der Charakter von Anekdötchen zukommen. Es hat sich in mehreren Überprüfungen immer wieder gezeigt, daß die Einschätzung potentieller Kulturverzerrungen von Items mit Hilfe des "gesunden Menschenverstandes" anstelle von empirischen Daten meist nicht glückt. So berichtet Jensen (1984, S.515) daß sich Personen bei der Inspektion von Items zwar meistens recht schnell einig wurden, welche Items als besonders "kulturell geladen" anzusehen sind, daß sie mit ihren Vorhersagen aber keine Trefferquote erzielten, die über dem Zufall lag. Die Notwendigkeit der Replikation der Befunde an unabhängigen Stichproben betonen Reynolds und Brown (1984, S.26) gerade auch angesichts der großen Sensibilität der zur Aufdeckung von Itemverzerrungen eingesetzten statistischen Verfahren. Die *DGP* wird den hier referierten Befund zum Anlaß nehmen, die gruppenspezifische Itemkennwerte demnächst erneut zu untersuchen. Über solche nachträglichen Untersuchungen hinaus kommt der Tatsache, daß die *DGP* bei den in nächster Zukunft anstehenden umfassenden Testrevisionen und Testneukonstruktionen die Ost-West-Thematik von Anfang an mitberücksichtigen wird, eine größere Bedeutung zu. Die Äquivalenz der Kennwerte für Ost und West wird dabei ein Kriterium des Testaufbaus sein.

Neben der Überprüfung an weiteren Stichproben ist auch eine kulturspezifische Überprüfung der Attraktivität der "Distraktoren", also der falschen Antwortalternativen, notwendig. Klieme und Stumpf (1991) haben ein spezielles Programm zur "Differentiellen Item-Analyse" entwickelt, welches neuere Methoden der Itemanalyse berücksichtigt und dabei zwischen Schwierigkeits-,

[2]Am Rande sei bemerkt, daß gerade pseudoprofessionelle Testkritiker sich diesen Fehler der Betrachtung von "single items" bei ihren Argumentationen bzw. bei ihren Versuchen, negative Stimmungen gegen Tests zu schüren, geschickt zunutze machen. In einem ersten Schritt suchen sie sich mißlungene Items aus *verschiedenen* Tests heraus. Diese Fleißarbeit -unbestritten ein hübsches Kuriositätenkabinett- präsentieren sie dann ihren Leser(inne)n oder Zuhörer(inne)n. Dem folgt dann die Diskreditierung des überhaupt nicht thematisierten Gesamttests, indem dem Gegenüber suggeriert wird, die Entscheidungen würden aufgrund dieser einzelnen Items -und nicht, wie in Testpraxis, auf Gruppen von Items- beruhen (siehe z.B.Hesse und Schrader, 1988,S.151 f.)

Fehler- und Trennschärfeanalysen unterscheidet. Auf den Bericht der Trennschärfekennwerte, die ebenfalls zur Information über die möglicherweise kulturspezifische Güte der Items herangezogen werden können, wurde hier verzichtet, da die Trennschärfen von der Schwierigkeit abhängig sind und man somit Gefahr läuft, ein und dasselbe Ergebnis mehrmals zu werten.

Die Gefahr einer Überinterpretation von singulären Itemstatistiken wird von vielen Fachleuten betont. Poortinga und van de Vijver (1987) beklagen ganz allgemein, daß den Itemstatistiken im Zusammenhang mit der kulturellen Verzerrung von Testinterpretationen im Vergleich zum Generalitätsanspruch des Tests zuviel Gewicht beigemessen wird. Guldin (1991, S. 218 f.) verweist spezifisch auf die Unangemessenheit der Verwendung von Itemschwierigkeiten als Indikatoren und empfiehlt in Hinblick auf die Identifikation von Meßfehlern auf der Itemebene das auf probabilistische Testkonzepte aufbauende Konzept der Äquivalenz von Messungen. Zentraler Ansatzpunkt der Kritik an Interpretationen von Itemverzerrungen ist immer wieder die "Ebene der Messung". Während der Test auf der Ebene mehrerer integrierter Items interpretiert wird, konzentriert sich der Nachweis von Itemverzerrungen auf die Item-Ebene. Ein Test mit vier Items müßte demnach als "verzerrt" gelten, wenn jeweils zwei Items der einen Gruppe deutlich leichter fallen würden als der anderen Gruppe -obwohl beide Gruppe in diesem hypothetischen Fall denselben Testwert erzielen würden. Der Begriff *"Item-bias"* wird deshalb mittlerweile häufig durch den stimmigeren Begriff *"differential item functioning"* ersetzt (z.B. Hartigan et al., 1989, S. 107). Die nachfolgenden Abschnitte handeln von Versuchen, den gruppenspezifischen Effekt der kulturellen Ladung des Tests auf der Aufgabenebene zu bestimmen.

2.4.2 Zum Einfluß der Vertrautheit mit der Sprache auf die Rechtschreibleistungen: Variation des sprachlichen Aufgabenmaterials

Als eine mögliche Quelle für kulturverzerrte Testleistungen wird immer wieder vor allem die Sprache genannt, wobei zumeist sowohl Schwierigkeiten der sprachlichen Formulierungen und Anforderungen im Test als auch die verbale Kommunikation mit den testdurchführenden Personen thematisiert werden (z.B. *"examiner and language bias"* bei Reynolds et al., 1984, S.17; *"linguistic equivalence"* bei Butcher, 1982, S. 283, sowie Helms, 1992, S. 1092 f.; *"Tester-Testee interaction / communication failure"* bei van de Vijver et al, 1992, S.19). Dieser Punkt betrifft natürlich insbesondere Tests, die von einer Sprache in die andere übersetzt und/oder bei Nichtmuttersprachler-(inne)n angewandt werden. Gleichwohl diskutiert z.B. Helms (1992, S. 1093) explizit auch verschiedene linguistische "Versionen" ***einer*** Sprache. Mit der nachfolgenden Untersuchung sollte geprüft werden, ob das in der *DGP* verwendete Diktat möglicherweise Eigenheiten eines westdeutschen Sprachgebrauchs widerspiegelt und dadurch den Bewerber(inne)n aus den neuen Bundesländern vergleichsweise mehr Schwierigkeiten bereitet. Der Ost-West-

Vergleich der Rechtschreibleistung mit Hilfe des Lückendiktates der *DGP* hatte für die Bewerber(innen) aus den neuen Bundesländern einen höheren Fehlererwartungswert nachgewiesen. Daß sich gerade im Bereich der Rechtschreibung mit die deutlichsten Ost-West-Unterschiede zugunsten der Testteilnehmer(innen) aus den alten Bundesländern zeigten (siehe Kersting, 1993, S. 65), war ein besonders erwartungswidriges Ergebnis. Schließlich wurde in der ehemaligen DDR dem Rechtschreibunterricht ein größerer Stellenwert beigemessen als in der Bundesrepublik allgemein üblich. Nach Becker (1991) wurde in der ehemaligen DDR ein Grundwortschatz von im Alltag häufigen Begriffen intensiv geübt und nach der Vorbereitung mit Diktaten überprüft. Der Autor zitiert eine Ost-West vergleichende Untersuchung mit knapp 5000 Erst- bis Viertkläßler(inne)n, bei der die West-Schüler(innen) sich im Diktat doppelt soviele Fehler leisteten wie die Ost-Schüler(innen). Nur beim Schreiben freier Texte stimmte die Fehlerquote überein.

Um den so genährten Verdacht einer kulturellen verzerrten Rechtschreibüberprüfung der *DGP* zu untersuchen, wurden aus ehemaligen Schulbüchern der DDR vier Texte herausgesucht und zu einem neuen "Ost-Lückendiktat" zusammengestellt.[3] Ausgewählt wurden vier Texte, die im Deutschunterricht der DDR als Diktate (wenngleich nicht als Lückendiktate) zur Lernzielkontrolle eingesetzt wurden.

Dieses Ost-Diktat wurde nun zusätzlich zur normalen "Vortestbatterie" bei 59 Bewerberinnen und 61 Bewerbern für den allgemeinen gehobenen Verwaltungsdienst in den Altbundesländern (insgesamt 120) und bei 89 Bewerberinnen und 63 Bewerbern für die gleiche Laufbahn in den neuen Bundesländern (insgesamt 152) appliziert, wobei die Darbietungsreihenfolge (Ostdiktat vor Westdiktat und vice versa) kontrolliert wurde (und keinen Effekt zeitigte).[4]

Zunächst wurde deutlich, daß es nicht einfach ist, ein Diktat mit angemessenem Schwierigkeitsgrad zu konstruieren (siehe Weber, 1986). Für die Gesamtgruppe fiel das "neue" "Ost-Diktat" leichter aus als das etablierte Diktat der *DGP*. Während die Fehlerzahl im *DGP*-Diktat zwischen eins und 22 variierte (Mittelwert 10.45, Standardabweichung 3.52), leisteten sich die Bewerber(innen) im "Ost-Diktat" nur maximal fünfzehn Fehler. Fünf Kandidat(inn)en überstanden den Test sogar fehlerfrei. Der Mittelwert lag hier bei 4.46 Fehlern, die Standardabweichung betrug 2.23. Für die weiteren Analysen wurden diese Verteilungsunterschiede zwischen den Tests durch eine Standardisierung (Z-Transformation) innerhalb der Gesamtgruppe entfernt.

[3]Diese Konstruktion wäre ohne die maßgebliche Hilfe mehrerer Mitarbeiter(innen) (z.T. Lehrer(innen) bzw. Lehramtsstudent(inn)en) des *DGP*-Büros Leipzig nicht möglich gewesen.

[4]Die Ergebnisse des Ost-Diktats blieben bei der Ernstfallbeurteilung des Vortests selbstverständlich unberücksichtigt. Die Teilnehmer(innen) wurden nach dem "Ostdiktat" über den experimentellen Charakter dieses Teils des Vortests informiert.

Mit der Untersuchung, an der 275 Personen teilnahmen, sollte geklärt werden, ob die im ursprünglichen "West-Diktat" und in dem übrigen Test unterlegene Gruppe der Neubundesbürger(innen) bei dem Diktat aus der eigenen Kultur eventuell gleich gute oder gar bessere Leistungen als die Westdeutschen erzielte. Diese Prüfung konnte zunächst nicht vorgenommen werden, weil sich in dieser Stichprobe überraschenderweise kaum Ost-West-Leistungsunterschiede zeigten. Lediglich bei den Zahlenmatrizen zeigte sich die aus den bisherigen Analysen vertraute durchschnittliche Überlegenheit der Bewerber(innen) aus den alten Bundesländern ($F(3,268) = 4.36$, $p = .04$). Ein gleich gerichteter Effekt bei den Wissensfragen ließ sich nicht gegen den Zufall absichern ($F(3,268) = 3.44$, $p = .07$).
Weder im herkömmlichen Diktat, noch im neukonstruierten "Ost-Diktat" zeigten sich signifikante Leistungsunterschiede zwischen den beiden Gruppen. Beim Ost-Diktat zeigte sich aber eine signifikante Interaktion zwischen den beiden unabhängigen Variablen "Geschlecht" und "Kultur": während in den alten Bundesländern die Rechtschreibfehler erwartungsgemäß häufiger in den von Männern ausgefüllten Diktaten zu finden waren, blieben die Ost-Männer mit ihren Fehlererwartungswerten unter dem Wert ihrer Mitstreiterinnen.

Der Befund des ausbleibenden Ost-West-Unterschieds an einer Stichprobe von immerhin fast 300 Personen kann als Beleg dafür angesehen werden, daß der kulturelle Effekt auf die Testleistungen insgesamt sehr schwach ist. Man kann hierin auch die Hoffnung bestärkt sehen, daß der Effekt insbesondere Startschwierigkeiten widerspiegelt und mit der Zeit "rauswächst". Leider sprechen Auswertungen anderer aktueller Untersuchungen (siehe unten, Punkt 5) gegen diese These der "Spontanremission der Test- oder der Selbstselektionskrankheiten".

Zur Erklärung des ausbleibenden Effekts können auch Stichprobenspezifika angeführt werden. Bedeutung könnte hier z.B. dem Untersuchungstermin zukommen. Die Untersuchung wurde in der 47ten, 48ten und 49ten Woche des Jahres 1992 durchgeführt. Da die "Bewerbungs- und Testsaison" in den alten und neuen Bundesländern zu unterschiedlichen Terminen beginnt und endet, geht mit dieser zeitlich identischen Messung ein unterschiedlicher Saisonabschnitt einher. Während für die Altbundesländer die Saison zu diesem Zeitpunkt schon im vollen Gange war, hatte die Saison in den neuen Bundesländern gerade erst begonnen. Der Saisonabschnitt kann aber systematisch mit der Leistungsstärke der Gruppe zusammenhängen (siehe unten, Punkt 2.4.3). Ungewöhnlich ist auch, daß in der Weststichprobe der Anteil an Frauen geringer war als der Anteil an Männern. Ein weiterer gravierender Unterschied zur Erstuntersuchung bestand darin, daß diesmal nur jeweils **ein** westliches und **ein** östliches Bundesland in die Untersuchung einbezogen waren. Dabei handelte es sich um ein neues Bundesland, welches im "ostinternen" Leistungsvergleich der Ergebnisse des ersten Ost-West Vergleichs tendenziell eher testleistungsstärkere Kandidat(inn)en aufweisen konnte.

So erfreulich das Verschwinden der Ost-West Leistungsunterschiede im Test sonst ist, bezogen auf die vorliegende Fragestellung war es natürlich ein unerfreuliches Ergebnis. Um den kulturspezifischen Effekt des "Ostdiktates" auf die Rechtschreibleistung doch noch prüfen zu können, wurde eine systematische Reduktion der Oststichprobe vorgenommen. 17 Männer und 15 Frauen aus den neuen Bundesländern wurden aus der Oststichprobe ausgeschlossen. Selektiert wurden dabei solche Personen, die erstens ein gutes Testergebnis (gemessen am statistischen Aggregat über alle Tests) aufzuweisen hatten und zweitens in der experimentellen Bedingung "Westdiktat vor Ostdiktat" zu finden waren (da diese Zelle des Versuchsplans vergleichsweise überbesetzt war). Soweit es mit diesen beiden Kriterien vereinbar war, wurde dabei solange jede zweite Person ausgeschlossen, bis die Sollgröße von 120 verbleibenden Bewerber(inne)n aus den neuen Bundesländern erreicht war.

In dieser reduzierten Stichprobe zeigten sich nun die Testleistungsunterschiede zugunsten der Westgruppe in einem mit der Erstuntersuchung vergleichbaren Ausmaß. Wie in dieser "Erstuntersuchung" (Kersting, 1993) zeigten sich bei einer Aufgabe zum logischen Denken keine Leistungsunterschiede zwischen den Gruppen. In der "neuen" Stichprobe verfehlte außerdem der kulturspezifische Leistungsunterschied in der Aufgabe zum Arbeitsverhalten und im Grundrechentest die Signifikanzgrenze. Bei allen übrigen sechs Tests blieben die Erwartungswerte für die künstlich reduzierte Stichprobe an Bewerber(inne)n aus den neuen Bundesländern unter denen für die Bewerber(innen) aus den alten Bundesländern zurück, dies gilt auch für das herkömmliche *DGP* -Diktat ($\underline{F}(3,236) = 4.18$, $\underline{p} = .04$). Im neukonstruierten Ostdiktat zeigten sich hingegen nach wie vor keine Leistungsunterschiede zwischen den Gruppen ($\underline{F}(3,236) = .00$, $\underline{p}$ = n.s.). Zu erwarten gewesen wäre ja eigentlich, daß durch die Reduzierung der Stichprobe um die leistungsstarken Personen nicht nur die durchschnittliche Leistung im herkömmlichen *DGP* -Diktat betroffen gewesen wäre (hier führte die Oststichprobenreduktion nach dem Kriterium der Gesamttestleistung ja zu kulturellen Leistungsunterschieden zugunsten der Altbundesbürger(innen)), sondern auch die durchschnittliche Leistung im "Ost-Diktat". Offensichtlich können aber auch Personen, die sich in der Mehrzahl der übrigen Tests als leistungsschwach erweisen, in diesem Diktat gleich gute Leistungen erzielen wie ihr im Durchschnitt testleistungsstärkeres Gegenüber. Der Vergleich der Diktatleistungen *innerhalb* der Gruppen zeigt in der Tendenz, daß den Bewerber(inne)n aus den alten Bundesländern beim "heimischen" Diktat weniger Fehler unterliefen als beim "Ostdiktat", wobei sich dieser Effekt aber zufallskritisch nicht ausreichend absichern ließ ($\underline{t} = -1.72$, $\underline{p} = .09$).

Bilanziert man die Ergebnisse, muß man konstatieren, daß die Befunde der Untersuchung zum Einfluß der Vertrautheit mit der Sprache insgesamt wenig aufschlußreich sind. Vor allem die nur unzureichend kontrollierten Stichprobenspezifika und die notwendige Reduktion der Stichprobe stellen die Befunde in Frage. Hinzu kommt, daß der Effekt nicht notwendigerweise auf die Variation des Sprachmaterials zurückgeführt werden muß, sondern auch

auf Erinnerungseffekte bei den Bewerber(inne)n aus den neuen Bundesländern zurückgeführt werden kann. Während das Standarddiktat als unveröffentlichte Eigenkonstruktion der *DGP* für die Testteilnehmer(innen) neu war, ließen die Kommentare der Bewerber(innen) aus den neuen Ländern bei der Anweisung des Ostdiktates Rückschlüsse auf die Erinnerung an Schulzeiten zu. Will man die Untersuchung der Fragestellung in Zukunft wiederholen, muß man aus den hier angesprochenen Fehlern bei der Untersuchungsdurchführung lernen.

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

Besonders bei Kenntnistests besteht die Gefahr, daß bildungsspezifische Eigenheiten der Kultur der Testkonstrukteur(inn)e(n) in unzulässiger Weise auf andere Kulturen übertragen werden. In der Literatur werden solche Effekte unter den Stichworten *"inappropriate content"* (Reynolds et al., S.17) oder *"content bias"* (Walsh und Betz, 1985, S.380) thematisiert.

Dieser Aspekt wird auch von den Bewerber(inne)n und Auftraggeber(inne)n am häufigsten betont. So wurde den *DGP* -Psycholog(inn)en nach der Anweisung des Kenntistests von verärgerten Bewerber(inne)n schon einmal der Rat erteilt, bei einem Test in Hannover die Teilnehmer(innen) aufzufordern, alle Republiken der ehemaligen UdSSR aufzulisten... Die *DGP* hat zu Beginn ihrer Arbeit in den neuen Bundesländern den Test mit gemeinschaftskundlichen Fragen allerdings überwiegend zur Gewinnung von Daten "mitlaufen" lassen. Allen getesteten Neubundesbürger(inne)n wurde für die Testinterpretation zumindest der Durchschnittswert der Westgruppe zuerkannt, auch wenn sie tatsächlich schlechter abgeschnitten hatten. Lediglich Bewerber(innen), die im "West-Kenntnistest" fundierte Kenntnisse zeigten, konnten durch die dann vorgenommene Verrechnung des "tatsächlichen" Wertes Bonuspunkte sammeln.[5] Erst nach einer Überarbeitung der Gemeinschaftskundefragen, die in Zusammenarbeit mit Lehrer(inne)n und Psycholog(inn)en aus der ehemaligen DDR erfolgte, wurden die Testergebnisse auch in den neuen Bundesländern "normal" ausgewertet.

Bei dieser gegen Ende des Jahres 1991 vorgenommenen Modifikation des Tests wurden bei neun Items des insgesamt 40 Items umfassenden Tests die Fragen oder Antwortmöglichkeiten verändert. Sieben dieser Veränderungen standen explizit im Zeichen einer Anpassung an die Kenntnisse der Bewerber(innen)gruppe aus den neuen Ländern. So wurde beispielsweise in diesen Fragen nach Ereignissen der jüngsten gemeinsamen deutschen Geschichte oder sogar explizit nach Ereignissen in der Historie der DDR gefragt. Bei

[5] Im ersten Teil des Artikels im letzten Heft der "*DGP* Informationen" wurden unter der Rubrik "Wissen" diese "ostkorrigierten" Werte berichtet, so daß das tatsächliche Ausmaß der Ost-West Unterschiede im damaligen Wissenstest deutlich unterschätzt wurde (Kersting 1993, S. 65).

Fragen, die z.B. auf bestimmte Merkmale der Bundesländer zielten, wurden nun auch die neuen Bundesländer als Antwortalternativen verwendet usw. Neben diesen Modifikationen wurde der Test insgesamt sprachlich überarbeitet, die Fragen wurden beispielsweise geschlechtsneutral formuliert.

Die Tatsache, daß diese zweite Modifikation innerhalb des Erhebungszeitraumes der vorliegenden Ost-West-Untersuchung vorgenommen wurde, eröffnete die Möglichkeit, den kulturspezifischen Effekt der Modifikation zu testen. Allerdings haben nur 43 der insgesamt 255 in der Stichprobe (siehe Tab. 1) enthaltenen Bürger(innen) aus den neuen Bundesländern die "nicht-modifizierte" Version des Kenntnistests bearbeitet. Die Modifikation sollte ja gerade zum "Saisonbeginn" in den neuen Ländern erfolgen. In den Altbundesländern war die Saison zum Stichtag des Testwechsels (1. Februar 1992) hingegen schon im vollen Gange. Dementsprechend bearbeiteten 175 der 245 Westdeutschen in der Stichprobe die nicht-modifizierte Version des Kenntnistests.

Erschwert wurde die Analyse des Modifikationseffekts auf die Messung des Kenntnisstandes dadurch, daß im Vergleich der beiden Zeitpunkte ("vor" und "nach" der Einführung des überarbeiteten Kenntnistests) die Bewerber(innen) aus den alten Bundesländern sich insgesamt in ihrer durchschnittlichen Testleistung etwas verschlechterten. Beim Vergleich der vor und nach der Testmodifikation getesteten Bewerber(innen) aus den neuen Bundesländern konnten hingegen auf Gesamttestebene keine nennenswerten Leistungsunterschiede festgestellt werden. Für ein statistisches Gesamtmaß über alle neun Tests ergab sich für die westdeutschen Bewerber(innen), die vor der Modifikation getestet wurden, ein durchschnittlicher Standardwert von 100,2. Für die westdeutschen Bewerber(innen), die nach dem 1.Februar getestet wurden, ergab sich ein Wert von 97,7 in diesem Gesamtmaß. Die entsprechenden Werte der ostdeutschen Kontrastgruppe lagen bei 96,7 und bei 97,0. Zwischen den Gruppierungsvariablen "Kultur" und "Testzeitpunkt" (vor und nach dem 1. Februar 1992) ereignete sich eine statistisch bedeutsame Interaktion ($F(3,496) = 6.08$, $p = .01$).

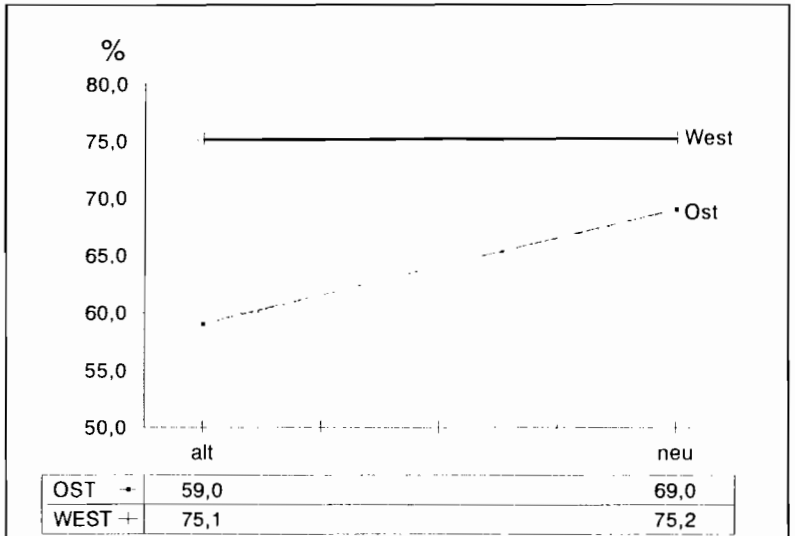
Solche Veränderungen spiegeln oft "saisonale" Effekte wider und sind nicht ungewöhnlich. Oft sind Bewerber(innen), die sich erst auf "den letzten Drücker" oder auf eine zweite Ausschreibung hin bewerben, durchschnittlich leistungsschwächer. Dabei kann es sich beispielsweise um Personen handeln, die sich erst bewerben, nachdem sie in Auswahlverfahren für ihren "Traumberuf" ausgeschieden sind. Diese Gruppe stellt somit eine Negativauslese dar, die letzten Testtermine werden in vulgo oft "Lumpensammler" genannt. Diese Ausführungen sind in bezug auf die vorliegende Stichprobe natürlich Spekulationen, die hier im konkreten Fall nicht zutreffen müssen. Im nachhinein sind solche Effekte nur schwer zu deuten.

Die oben genannte Fragestellung, ob die "kulturelle Adaption" des Kenntnistests an die Lernerfahrungen der ostdeutschen Bewerber(innen) sich positiv auf deren Testleistungen auswirkt, wird durch den Befund der Leistungsveränderungen in einigen anderen -nicht modifizierten- Tests verkompliziert. Interessiert man sich -wie im vorliegenden Fall- für den Effekt der Testmodifikation auf das Ausmaß an Kenntnissen, sollte man möglichst jenen anderen Effekt, welcher von dem unterschiedlichen Leistungsausmaß in anderen Tests zum Zeitpunkt vor der Testmodifikation auf das Leistungsausmaß zum Zeitpunkt nach der Testmodifikation ausgehen kann, kontrollieren. Eine solche Kontrolle wird durch eine sogenannte Kovarianzanalyse ermöglicht. Bezogen auf den vorliegenden Fall wird bei diesem statistischen Vorgehen der Fehlerwert dadurch reduziert, daß eine Beziehung zwischen der interessierenden Variablen, den Kenntnissen und der Gesamttestleistung als Kovariate hergestellt wird (siehe z.B. Tabachnick & Fidell, 1989). Die Mittelwerte der abhängigen Variablen "Kenntnisse" werden dabei sozusagen an die Werte adjustiert, die sie annehmen würden, wenn alle Personen den gleichen Wert auf der Kovariaten, dem Ausmaß der sonstigen Testleistung, hätten. Aufgeklärt wird bei dieser Betrachtungsweise jene Varianz, die über die allgemeine Leistungsveränderung zwischen den beiden Zeitpunkten hinaus existiert.[6]

Es wurden 2x2 *"between subject"* Kovarianzanalysen mit der Gruppenzugehörigkeit ("Ost" oder "West") und dem Zeitpunkt ("vor" oder "nach" der Modifikation des Kenntnistests) als unabhängige Variablen und dem Kenntnisstand als abhängige Variable durchgeführt, wobei die Ergebnisse durch die Einführung einer Kovariaten (Leistungen in dem Aggregat der übrigen acht Tests) korrigiert wurden.

Abbildung 4 verzeichnet den durchschnittlichen Kenntnisstand in gemeinschaftskundlichen Fragen der vor und nach der Testmodifikation getesteten Gruppen, aufgeteilt nach "Ost-" und nach "Westbewerber(innen)". Abgetragen sind die adjustierten mittleren Rohwerte. Die obere, fast waagerechte Linie für die Bewerber(innen) altbundesdeutscher Verwaltungen zeigt, daß die Testmodifikation hier keine signifikanten Änderungen nach sich gezogen hat. Anders verhält es sich bei den Bewerberinnen und Bewerbern aus den neuen Bundesländern (untere Linie). Konnte die mit der älteren Version des Wissenstests befragte Ostgruppe im Durchschnitt nur knapp 59 % der Fragen lösen, so wuchs dieser Prozentsatz für die mit der modifizierten Testform untersuchte Ostgruppe um 10 % auf 69 Prozent. Trotz dieser Steigerung bleibt die durchschnittliche Leistung der Bewerber(innen) aus den neuen Bundesländern aber immer noch unter dem Wert der Westgruppe, die im Durchschnitt 75 % der modifizierten Wissensfragen zu beantworten wußte.

[6] Problematisch ist diese Analyse allerdings deshalb, weil die Testleistungen in den anderen Tests mit der Leistung im Kenntnistest zu .35 (part-whole korrigiert) korrelierten.



	alt	neu
OST •	59,0	69,0
WEST +	75,1	75,2

Anzahl richtiger Lösungen in Prozent (adjustierte Mittelwerte)

N=500 (1.alte Form: Ost=43, West=175; 2.neue Form: Ost=212, West=70)

Abb 4.: Kenntnisstand in Gemeinschaftskunde; Ost-West-Vergleich vor und nach einer Teständerung (jeweils andere Gruppen)

Auch bezüglich des Einflusses der Repräsentation "Ost-" oder "West-" spezifischer Kenntnisse auf die Leistungen im Wissenstest muß das Fazit also lauten: ja, es gibt einen Einfluß des Aufgabenkontents, aber auch nach der Kontentanpassung bleiben die Gruppenleistungsunterschiede bestehen.

2.4.4 Kontrolle möglicher Beurteiler(innen)effekte

Bei der Beurteilung lassen sich Menschen leicht von Stereotypen beeinflussen. Während Stereotypbeurteilungen in bestimmten Alltagssituationen aufgrund ihrer komplexitätsreduzierenden Funktion sinnvoll sein können, müssen sie im diagnostischen Prozeß als Fehler gewertet werden. Diagnostiker(innen) sollen ihre Subjektivität so weit wie möglich kontrollieren. Diese Kontrolle der Subjektivität wird bei den Eignungsuntersuchungen der *DGP* vor allem durch die "mechanisch" ablaufende Registrierung und Auswertung der im Test gewonnenen Informationen gewährleistet. Allerdings gibt es auch im Testteil des diagnostischen Prozesses der *DGP* einen Moment, wo das Urteil der Psycholog(inn)en mehr Freiheitsgrade hat: Die mechanisch erzeugten "Testprofile" der Kandidat(inn)en werden von den Psycholog(inn)en in einer sogenannten "klinischen" Verarbeitung zu einem "Empfehlungsgrad" zusammengefaßt. Stereotype gegenüber West- und Ostdeutschen seitens der Psycholog(inn)en könnten sich an dieser Stelle des diagnostischen Prozesses manifestieren. Dies gilt insbesondere dann, wenn die Testleistungen ostdeutscher Bewerber(innen) von westdeutschen Psycholog(inn)en eingestuft werden und umgekehrt.

Auf den Einfluß von Stereotypen auf die Beurteilung weisen Stepper und Strack (1993) in ihrem Bericht über ein sozialpsychologisches Experiment hin. 35 im Westen Deutschlands geborene Student(inn)en sollten sich gedanklich in die Lage einer Personalchefin / eines Personalchefs versetzen, die/der aufgrund von Bewerbungsunterlagen die Qualifikation eines Kandidaten beurteilen sollte. Die Hälfte der Gruppe hatte über einen Kandidaten aus Ostdeutschland zu urteilen, die andere Hälfte beurteilte einen westdeutschen Kandidaten. Außer Geburts- und Studienort waren die Bewerbungsunterlagen der Kandidaten aber identisch, es gab somit keine Unterschiede in der Qualität des ost- und westdeutschen Bewerbers. Vor der Beurteilung wurde die Hälfte der urteilenden Personen in eine positive, die andere Hälfte in eine negative Stimmung versetzt. Dabei zeigte sich, daß in guter Stimmung offensichtlich bei der Beurteilung Stereotype angewandt wurden, die bei den Westdeutschen trotz der gleichen Qualifikation der Kandidaten zu einer Abwertung (Nicht-Einstellung) des Ostdeutschen führte. (Bei den Versuchspersonen in negativer Stimmung zeigte sich dieser Effekt hingegen nicht. Einige Autoren gehen davon aus, daß Menschen in guter Stimmung zur oberflächlichen Informationsverarbeitung neigen, so daß die Anwendung von Stereotypen in diesem emotionalen Zustand wahrscheinlicher wird).

Anhand der vorliegenden Daten läßt sich die These einer stereotyp-verzerrten Beurteilung der Testleistungen zumindest annäherungsweise prüfen. 202 Personen der in Tabelle 1 aufgeführten 500 Personen wurden zum Haupttest zugelassen. Darunter befanden sich 105 aus den neuen Bundesländern. Die Testprofile von 21 Bewerber(inne)n aus den neuen Bundesländern wurden von West-Psycholog(inn)en beurteilt. Die Testprofile der übrigen 84 Bewerber(innen) aus dieser Gruppe wurden hingegen von Psycholog(inn)en der gleichen kulturellen Gruppe beurteilt. Bei der Beurteilung der Testprofile orientieren sich die Psycholog(inn)en insbesondere an den Leistungen auf dimensionaler Ebene, damit sind die Bereiche "sprachliches Denken", "rechnerisches Denken", "Arbeitsverhalten" und "Kenntnisse" gemeint. Betrachtet man zunächst die objektiveren, mechanisch erhobenen Daten für diese Bereiche, so zeigte sich in keiner Dimension ein statistisch bedeutsamer Leistungsunterschied zwischen den von West- und den von Ost-Psycholog(inn)en beurteilten Neubundesbürger(inne)n. Dementsprechend dürften sich die Gruppen auch im weniger objektiven klinischen Urteil über das Testprofil nicht unterscheiden, wenn die Psycholog(inn)en sich an die Daten halten und keine Stereotype in die Beurteilung miteinfließen lassen. Tatsächlich ergab sich für beide Gruppen im Durchschnitt ein fast identischer Empfehlungsgrad für das Testprofil. Es spielte keine Rolle, ob die Testleistungen der ehemaligen DDR-Bürger(innen) von Ost- oder von West-Psycholog(inn)en beurteilt wurden ($\underline{t}$=.42, n.s.). Die Gegenprobe ließ sich leider nicht realisieren: Die Testprofile aller 97 zum Haupttest zugelassenen Bewerber(innen) aus den alten Bundesländern wurden von Psycholog(inn)en aus den alten Bundesländern beurteilt.

3. Testleistungsunterschiede und ihre Bedeutung bei Auswahlentscheidungen

Im dritten Abschnitt des Artikels soll sich die Betrachtungsweise von der internen Analyse der Tests abkehren und der Bedeutung von Testleistungsunterschieden bei Auswahlentscheidungen zuwenden. Dabei soll die laut Jäger (1970, S. 647) notwendige Unterscheidung zwischen dem Eignungsurteil und der Ausleseentscheidung beachtet werden. Zunächst werden allgemeine Überlegungen zu den Begriffen "Fairness" und "Testverzerrungen" referiert. Anschließend wird dann ein Versuch unternommen, anhand von Modellrechnungen abzuschätzen, welche Bedeutung den Gruppenleistungsunterschieden bei Auswahlentscheidungen überhaupt zukommen kann, wenn sie in einer so geringfügigen Größe auftreten wie beim vorliegenden Ost-West Vergleich.

3.1 Testfairness und Testverzerrungen im allgemeinen

Die Frage nach der Angemessenheit des Tests, nach der *"Fairness"* der mit Hilfe des Tests getroffenen Entscheidungen, kann nicht aufgrund von internen Kriterien, wie z.B. Mittelwertsunterschieden, Faktorladungen, Itemkennwerten usw., entschieden werden. Dazu bedarf es externer Kriterien. Der Begriff *"fair"* setzt immer die *normative* Setzung irgendeines externen Kriteriums voraus, nach dem die Entscheidungen *fair* oder *unfair* sein sollen (siehe Trost, 1985, S. 47). Dieses Kriterium wird in Abhängigkeit von der jeweiligen Perspektive, aus der die Fairnessfrage betrachtet wird, gewählt: *"fair"* (1.) aus Sicht der einzelnen Bewerber(innen), (2.) der Gesamtgruppe an Bewerber(inne)n, (3.) der Untergruppen innerhalb der Bewerber(innen), also z.B. der Ostdeutschen oder der Frauen, *"fair"* (4.) aus Sicht der Institution, also z.B. der einstellenden Verwaltung, oder *"fair"* (5.) aus Sicht der Gemeinschaft, 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, unterschiedliche Fairnesskonzepte schließen sich wechselseitig aus. Nach Hunter und Schmidt, 1976, lassen sich die zahlreichen unterschiedlichen Fairnesskonzepte auf drei ethische Grundpositionen zurückführen. Den ersten beiden Positionen des *"qualified"* und des *"unqualified individualism"* (eingeschränkter und uneingeschränkter Individualismus) ist die Absicht gemeinsam, Personen mit gleichen Erfolgsaussichten auch gleiche Auswahlchancen zu sichern. Die Positionen unterscheiden sich dahingehend, daß der Position des eingeschränkten Individualismus zufolge, Indikatoren der Gruppenzugehörigkeit keinen Einfluß auf die Selektionsentscheidung nehmen dürfen. Die dritte ethische Grundposition des *"fair share"* basiert auf der Forderung, bei Auswahlentscheidungen gesellschaftlich relevante Gruppen nach der Vorgabe von Quoten zu berücksichtigen. Aus politischen Gründen geforderte Quotensysteme *"entziehen sich einer fachwissenschaftlichen Betrachtung"* (Wottawa und Amelang, 1980, S.211). Eine fachliche Diskussion verlangt die Forderung nach Quoten dann, wenn sie aus der impliziten Annahme hervorgeht, daß verschiedene Gruppen im Prädiktor- (z.B. im Eignungstest) wie im Kriteriumsverhalten (z.B. im Beruf) in Wirklichkeit gleich gut sind und die unterschiedliche Leistungsfähigkeit nur duch den Test "vorgetäuscht" wird (Bartussek, 1982, S. 3).

Insofern es sich um eine ethische Fragestellung handelt transzendiert das Problem der Fairness den Bereich der Psychometrie (siehe z.B. Poortinga, 1983, S.240). Psycholog(inn)en sind keine Fachkräfte für die Setzung ethischer Wertvorstellungen. Entsprechend haben viele Psycholog(inn)en immer wieder gefordert, zwischen der ethischen Fairnessfrage -zu der sie eher weniger als mehr beitragen können als z.B. Philosoph(inn)en, Theolog(inn)en und Angehörige anderer Berufsgruppen- und der statistischen Frage nach Test- und Vorhersageverzerrungen, der Frage nach der differentiellen Validität eines Tests, zu unterscheiden (z.B. Jensen, 1980, S.375; 1984, S.520; Reynolds et al., 1984, S. 21; Walsh et al., 1985, S.383). Eine Vermischung der beiden Konzepte erscheint nicht sinnvoll, da die Verwendung von Tests, die bezüglich ihrer Vorhersageleistung unverzerrt sind, keine Garantie dafür ist, daß die mit Hilfe dieses Tests getroffene Entscheidung unter allen Perspektiven "fair" ist -und vice versa (siehe z.B. Hunter, Schmidt & Hunter, 1979, S.733). "Fairness" und "Test-Bias" sind unterscheidbare Konzepte, gleichwohl beide Konzepte in Relation zueinander gestellt werden können und müssen. Diese Verdeutlichung der Relationen zwischen den beiden Konzepten gehört beispielsweise zum Aufgabenfeld der Psycholog(inn)en. Ihr Beitrag zur Fairnessfrage sollte grundsätzlicher formuliert darin bestehen, die unterschiedlichen Fairnessmodelle und ihre Implikationen transparent werden zu lassen und die dahinterstehenden Kriterien zu explizieren, die Schlüssigkeit und Widerspruchsfreiheit der einzelnen Modellkomponenten zu evaluieren, die jeweilige Definition und Bedeutung von Testverzerrungen innerhalb der Fairnessmodelle darzustellen und die Zielvorgaben des Modells in Handlungsregeln umzusetzen sowie Schwachstellen der Modelle aufzudecken. Diesem Auftrag sind zahlreiche Fachvertreter(innen) mit Überblicksarbeiten zu der großen Zahl an Fairnessmodellen, die sich den oben genannten drei ethischen Grundpositionen zuordnen lassen, gerecht geworden (z.B. Bartussek, 1982; Gösslbauer, 1977; Hunter, Schmidt & Rauschenberger, 1977 und 1984; Möbus, 1983; Petersen und Novick, 1976; Simons & Möbus, 1978; Wottawa und Amelang, 1980). Eine normative Setzung darüber, welche allgemein gültige Vorstellung von Fairness zu herrschen hat und welches Kriterium in welchem Fall "richtig" ist, steht weder der Psychologie noch irgendeiner anderen Fachrichtung zu. Dies ist Aufgabe der betroffenen Individuen, Institutionen und der Gemeinschaft. Die von der Psychologie geleistete Explikation erleichtert dabei aber deutlich die Entscheidungsfindung. Die Idee der Trennung der Konzepte "Fairness der Auswahlentscheidung" und "Test-Bias" wurde dahingehend kritisiert, daß die Tests durch diesen "Kunstgriff" gegenüber wissenschaftlicher Kritik immunisiert würden, die Testkonstrukteur(inn)e(n) sich entsprechend aus der Verantwortung stehlen wollten (siehe z.B. Hilliard III, 1984, S.144). Diese Kritik ist schwer nachvollziehbar, da ja gerade diese Vorgehensweise es erlaubt, einen Test unter dem Gesichtspunkt *bestimmter* Vorstellungen über Fairness als ungeeignet abzulehnen.

3.2 Testfairness und Testverzerrungen bei dem von der *DGP* durchgeführten Auswahlverfahren in den neuen Bundesländern

Die große Zahl unterschiedlicher, teils widersprüchlicher, teils aufeinanderaufbauender Fairnessmodelle wurde bereits angesprochen, Darstellungen sind den zitierten Quellen zu entnehmen.

Für die folgenden Überlegungen soll zunächst von dem nach Bartussek (1980, S. 5) zur empirischen Untersuchung von Testfairnessfragen am häufigsten angewandten *"(Regressions)Modell der fairen Vorhersage"* nach Cleary ausgegangen werden. Ein Selektionsverfahren 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"*. (Bartussek, ebd. S. 4 f.) Bei identischen "cutoffs" (z.B. kritischer Punktwert für die Zulassung) verlangt das Modell für zwei zu vergleichende Gruppen (z.B. "West" und "Ost") **identische** Regressionsgeraden zur Vorhersage des Kriteriums (also z.B. des Ausbildungs- oder Berufserfolges).

Die Definition macht deutlich, daß sich die Frage nach der Relevanz der mittleren Ost-West-Testleistungsunterschiede für die Personalauswahl unter diesem Gesichtspunkt erst dann beantworten läßt, wenn die prädiktiven Validitäten des Tests für beide Gruppen bestimmt sind. Würden sich die im Durchschnitt testleistungsschwächeren ehemaligen DDR-Bürger(innen) auch in der Ausbildung und im Beruf als leistungsschwächer erweisen, so hätte der Test diesen Personen "zu Recht" eine etwas geringere Bewährungschance prognostiziert -unabhängig davon, daß es unter anderen Gesichtspunkten sinnvoll erscheinen mag, auch leistungsschwächere Personen einzustellen. Aus dieser Sicht liegt das Problem darin, *"ob bestimmte Tests zu subgruppenspezifischen Fehleinschätzungen der Kriteriumswerte führen, und nicht darin, ob es Unterschiede der Testmittelwerte an sich gibt"* (Wottawa und Amelang, 1980, S.202; siehe z.B. auch Vernon, 1979, S.318). Abschätzungen über die Treffsicherheit der *DGP* -Tests in bezug auf z.B. den Ausbildungserfolg von ostdeutschen Fachhochschulstudent(inn)en liegen zur Zeit noch nicht vor. Aussagekräftige Bestimmungen der Vorhersagekraft sollten möglichst an einer Gruppe von *gemeinsam* ausgebildeten und beurteilten Personen aus West und Ost vorgenommen werden, da sich anderenfalls separate Gruppeneffekte (z.B. Beurteilung in beiden Gruppen gemäß der Normalverteilung, obwohl die/der leistungsmäßig mittelmäßige Kandidat/in der einen Gruppe eventuell in der anderen Gruppe als leistungsstark gelten würde) nicht kontrollieren lassen. Solange noch keine Validierungsergebnisse vorliegen, kann man sich bezüglich der Erwartung über die Gleichheit bzw. Unterschiedlichkeit der prognostischen Qualität in unterschiedlichen Gruppen an den zahlreichen Ergebnissen anderer empirischer Untersuchungen zur differentiellen oder singulären Kriteriumsvalidität orientieren. Dabei konnte für Gruppen, die kulturell deutlich disperater 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).

Des weiteren muß betont werden, daß die Bewerber(innen) in den von der *DGP* durchgeführten Auswahlverfahren fast ausschließlich *innerhalb* einer Kulturgruppe um einen Ausbildungsplatz konkurrieren. Die Anzahl der Fälle, in denen sich Westdeutsche bei Verwaltungen in den neuen Bundesländern um einen Ausbildungsplatz bewerben und umgekehrt, sind statistisch vernachlässigbar klein. Ausnahmen stellen spezifische Laufbahnen (z.B. Aufstiegsbewerbungen zum höheren Polizeivollzugsdienst in den neuen Bundesländern) und spezifische Verwaltungen, wie z.B. das Auswärtige Amt, dar. Diese Ausnahmen eignen sich im übrigen am besten für die Durchführung von vergleichenden Bewährungskontrollen. Das bedeutet, daß sich im allgemeinen die Gefahr einer kulturell "unfairen" Selektionsentscheidung aufgrund des *DGP* -Tests praktisch nicht stellt.

3.3 Auswirkungen der kulturspezifischen Testleistungsunterschiede auf den Selektionsprozeß

Mit den folgenden Ausführungen wird versucht abzuschätzen, welche Auswirkungen die geringen Testleistungsunterschiede zwischen Bewerber(inne)n aus den beiden Teilen Deutschlands auf den Selektionsprozeß der jeweiligen Gruppe haben. Die Überlegungen gehen dabei der Einfachheit halber von vier idealisierten Annahmen aus:

1. Es wird unterstellt, daß die Verwaltungen in beiden Teilen Deutschlands sich an die Empfehlungen der *DGP* halten und somit Eignungsurteil und Ausleseentscheidung identisch sind.
2. Es wird unterstellt, daß die Empfehlung/Entscheidung nur aufgrund der Testergebnisse (ohne Berücksichtigung der Verhaltensbeobachtungen) erfolgt.
3. Es wird unterstellt, daß in den neuen und in den alten Bundesländern alle Personen mit einem Testempfehlungsgrad von "4 -" zur Ausbildung zugelassen werden und alle Personen mit einem schlechteren Empfehlungsgrad abgewiesen werden.
4. Es wird unterstellt, daß die Vorhersagekraft der *DGP* -Tests für die beiden deutschen Bewerber(innen)gruppen identisch ist.

Die *DGP* legt bei der Vergabe ihrer Empfehlungsgrade im Osten und im Westen Deutschlands die gleichen Maßstäbe an. Unter der Gültigkeit der oben ausgeführten vereinfachenden Annahmen bedeuten Mittelwertsunterschiede im Prädiktor (im Test) automatisch auch Unterschiede in der Selektionsrate (Anteil der Angenommenen unter den Bewerber(inne)n). Eine Änderung der Selektionsrate hat aber Konsequenzen für den Auswahlprozeß. Bevor diese Konsequenzen beschrieben und quantifiziert werden können,müssen zunächst einige Grundbegriffe eingeführt werden. Zu diesem Zweck werden einige Ausführungen von Wiggins (1973, S.223-274) zu diesem Thema referiert.

Während man die Qualität von Tests und Vorhersagen aufgrund von psychometrischen Kriterien beurteilt, werden Personalentscheidungen an ihren Konsequenzen für die Individuen, Institutionen und für die Gemeinschaft gemessen. Die Personalentscheidung basiert auf irgendwelchen Informationen, in unserem Falle auf den Testergebnissen, die nach irgendeiner Entscheidungsregel verarbeitet werden. Im vorliegenden vereinfachten Fall lautet die Regel beispielsweise: *"Hat ein(e) Bewerber(in) im Auswahlverfahren mindestens einen Empfehlungsgrad von "4-" erreicht, biete ihr/ihm einen Ausbildungsplatz an. Ist der Empfehlungsgrad schlechter ausgefallen, schließe sie/ihn vom Zulassungsverfahren aus."* Der Nutzen der Personalauswahl ist nun unter anderem abhängig von der Entscheidung (hier: Zulassung / Ablehnung) und dem Ergebnis (hier: Erfolg oder Mißerfolg in der Ausbildung). Grundsätzlich resultieren folgende vier Ergebniskategorien:

A. Es wurde eine Zulassung zur Ausbildung empfohlen, und die Person war in der Ausbildung erfolgreich.
   Valide positiv; Erfolg vorhergesagt, Erfolg tritt ein.
B. Es wurde eine Zulassung zur Ausbildung empfohlen, aber die Person war in der Ausbildung nicht erfolgreich.
   Falsch positiv; Erfolg vorhergesagt, Mißerfolg tritt ein.
C. Die Zulassung zur Ausbildung wurde nicht empfohlen, und die Person war in der Ausbildung nicht erfolgreich.
   Valide negativ; Mißerfolg vorhergesagt, Mißerfolg tritt ein.
D. Die Zulassung zur Ausbildung wurde nicht empfohlen, aber die Person war in der Ausbildung erfolgreich.
   Falsch negativ; Mißerfolg vorhergesagt, Erfolg tritt ein.

Der Begriff der *Selektionsrate* bezieht sich auf alle Personen, die zur Ausbildung zugelassen sind, unabhängig von ihrem Ausbildungserfolg. Also alle validen Positiven und alle falsch Positiven (*Selektionsrate* SR = A + B).
Eine andere wichtige Größe in der Entscheidung ist die *Basisrate*. Diese bezieht sich auf alle Personen, die die Ausbildung bestehen würden, unabhängig davon, ob ihre Zulassung zur Ausbildung empfohlen wurde. Also alle valide Positiven und alle falsch Negativen (*Basisrate* BR = A + D).

Für den Fall, daß die Basisrate in den neuen Bundesländern geringer ist, soll zunächst im allgemeinen der der Effekt der Basisrate auf die Vorhersageergebnisse betrachtet werden. Angenommen zwei Verwaltungen, z.B. eine in den neuen Bundesländern und eine in den alten Bundesländern, wollen jeweils 25 Ausbildungsplätze besetzen. Bei beiden Verwaltungen haben sich jeweils 100 Bewerber(innen) auf die Ausbildungsplätze beworben und die Verwaltung wird auf jeden Fall 25 aus diesen 100 Personen für die Ausbildung auswählen (konstante Selektionsrate). Der Anteil der geeigneten Bewerber(innen) an der Gesamtbewerber(innen)gruppe (die Basisrate) betrage aber in der einen Verwaltung fünf von 100 und in der anderen Verwaltung 50 von 100. Der nachstehende Vergleich verdeutlicht, in welchem unterschiedlichen Ausmaß die Qualität der Personalentscheidung dieser Verwaltungen vom Einsatz des gleichen hypothetischen Tests mit einer für beide Gruppen gleichen mittleren

Kriteriumsvalidität von .35 gegenüber einer Auswahl nach Zufall profitieren würde. Dazu werden die Vorhersageergebnisse (A, valide positiv; B, falsch positiv; C, valide negativ, und D, falsch negativ) als Wahrscheinlichkeiten dargestellt: P(A), P(B), P(C) und P(D). Die Wahrscheinlichkeit für die erwünschte qualitativ hochwertige Entscheidung, nämlich für die "valide Positiven", läßt sich nun nach folgender Formel (Wiggins, 1973, S.255) bestimmen:

$$P(A) = BR * SR + \varphi_{yy'} \sqrt{BR(1-BR)\ SR(1-SR)}$$

wobei BR für Basisrate (im ersten Fall .05, im zweiten Fall .50), SR für Selektionsrate (in beiden Fällen .25) und $\Phi_{yy'}$ für den Validitätskoeffizienten der Vorhersage (im Beispiel beim Test = .35, bei einer Zufallsauswahl = .00) stehen.

Entscheidungs-ergebnis	Basisrate = .05		Basisrate = .50	
	Zufalls-auswahl	**Test-auswahl**	**Zufalls-auswahl**	**Test-auswahl**
Valide positiv (A)	1,2	4,5	12,5	23,3
Falsch positiv (B)	23,8	20,5	12,5	1,7
Valide negativ (C)	71,2	74,5	37,5	48,3
Falsch negativ (D)	3,8	0,5	37,5	26,7
Korrekte Entscheidungen (A+C)	72,4	79	50	71,6
Prozentanteil der valide Positiven an allen Ausgewählten (A / Selektionsr.)	4,8 %	18 %	50 %	93,2 %

Tabelle 4: Proportion der Entscheidungen in vier hypothetischen Fällen

Tabelle 4 zeigt die Proportionen der Entscheidungen für alle vier hypothetischen Fälle. Die Werte in der Tabelle bilden die Basis für eine Antwort auf die Ausgangsfrage: "Welchen Effekt hat die Basisrate auf die Vorhersageergebnisse?". Anders formuliert kann man auch fragen: "Wann lohnt sich der Einsatz eines Tests gegenüber der Auswahl per Zufall?" Die Antwort auf diese Frage fällt in Abhängigkeit des Begriffs "lohnen" jeweils anders aus. Sieht man es als besonders lohnenswert an, möglichst viele korrekte Entscheidungen zu treffen, dann bringt der Test unter der Bedingung besonders viel, daß die Basisrate im mittleren Bereich liegt. Dies sieht man in der

vorletzten Reihe der Tabelle. Für alle vier Fälle wurde die Anzahl der korrekten Entscheidungen als Addition der validen Positiven und der validen Negativen (A+C) bestimmt. Vergleicht man nun die Vorhersageleistung aufgrund des Tests mit der Vorhersageleistung aufgrund einer zufälligen Auswahl, ergibt sich folgendes Bild. Bei einer extremen Basisrate von nur fünf geeigneten Personen unter 100 Bewerber(inne)n, kann der Testeinsatz die Anzahl der korrekten Entscheidungen gegenüber der Zufallsstrategie nur geringfügig steigern (von 72,4 auf 79).
Für den anderen Fall einer mittleren Basisrate von 50 geeigneten Personen unter 100 Bewerber(innen) ergibt sich hingegen ein deutlich anderer Befund. Hier bewirkt der Testeinsatz gegenüber der Zufallsauswahl einen beachtlichen Zuwachs an korrekten Entscheidungen (von 50 auf 71,6). Unter dieser Perspektive muß die Antwort also lauten, daß der Einsatz eines Tests um so bessere Ergebnisse bringt, je weniger extrem die Basisrate ist.

Bei der Personalauswahl legt man aber oft einen ganz anderen Begriff von "lohnenswert" zugrunde. Bei dieser Betrachtungsweise fallen nur diejenigen ins Gewicht, die ausgewählt wurden. Personen, die nicht ausgewählt wurden, spielen für die Leistungsfähigkeit einer Verwaltung keine Rolle. Daran ändert auch der hypothetische Fall nichts, daß sie die Ausbildung erfolgreich absolviert hätten, wenn sie zugelassen worden wären. (Diese Beschreibung meint die falsch Negativen). Der Nutzen für die Institution (und im Falle der Verwaltung somit der Nutzen für das Gemeinwohl) entsteht durch die erfolgreichen Personen, die eine Zulassung erzielt haben; der Schaden entsteht durch die Zulassung nicht-erfolgreicher Personen. Abgelehnte Personen bringen der Verwaltung so oder so keinen Schaden und keinen Nutzen. Man betrachtet deshalb unter dieser Perspektive nur die "Trefferquote", also den prozentualen Anteil der positiven Fälle unter allen ausgewählten Personen (A /Selektionsrate). Dieser Wert ist in der untersten Zeile der Tabelle angegeben. Man sieht, daß der Testeinsatz unter dieser Perspektive in jedem Fall eine deutliche Verbesserung der Entscheidungen bedingt und gerade auch bei einer extrem niedrigen Basisrate für die Auswahl eines überzufällig hohen Prozentsatzes an positiven Fällen sorgt. Bei der extrem geringen Basisrate ist die so definierte Trefferquote beim Test viermal so hoch als bei einer Zufallsauswahl, bei der mittleren Basisrate wird die Trefferquote hingegen nur noch verdoppelt. Eine mittlere Basisrate bedeutet ja, daß es relativ viele gute Bewerber(innen) in der Gesamtgruppe an Bewerber(inne)n gibt. In dieser Situation findet aber auch das blinde Huhn "Zufall" gelegentlich ein Korn. Wäre die Basisrate ganz extrem, d.h. wären alle Bewerber(innen) geeignet, würde man mit jeder Methode -also auch durch graphologische oder astrologische Untersuchungen oder anderen Hokuspokus- eine geeignete Person herausfinden. Unter der Maßgabe, daß es allein darum geht, unter den Ausgewählten möglichst viele erfolgreiche zu haben, ist ein Test um so nützlicher, je geringer die Basisrate ist!

Der Einfluß der Selektionsrate auf das Vorhersageergebnis ist unter mathematischer Betrachtung identisch mit dem Einfluß der Basisrate auf das Vorhersageergebnis. Für die Personalauswahl ist der Einfluß der Selektionsrate auf das Vorhersageergebnis aber bedeutsamer, da die Selektionsrate -im Gegensatz zur Basisrate- beeinflußt werden kann. Bei konstanter Basisrate bedeutet eine strengere Selektion einen Rückgang der "falsch Positiven" und einen Anstieg der "falsch Negativen" (konservatives Vorgehen). Eine milde Selektion führt zu einer Erhöhung des Anteils an "falsch Positiven" und einem Rückgang der "falsch Negativen". Auch hier gilt: verfolgt man das Ziel, den Anteil an korrekten Entscheidungen zu erhöhen, ist ein Test bei mittleren Selektionsraten besonders effektiv. Betrachten wir zunächst ein Beispiel, indem die Basisrate der Selektionsrate entspricht. Von 100 Bewerber(inne)n seien beispielsweise 60 "in Wirklichkeit" geeignet (Basisrate) und es stehen 60 Ausbildungsplätze zur Verfügung (Selektionsrate). In diesem Falle besteht theoretisch die Chance, daß z.B. mit Hilfe eines Tests aus den 100 Personen genau die sechzig geeigneten Personen herausgefunden werden; die Vorhersagbarkeit wäre dann unter dem Kriterium eines möglichst hohen Anteils korrekter Entscheidungen perfekt. Hat die Verwaltung aber beispielsweise im nächsten Jahr wieder 100 Bewerber(innen) und darunter sind wieder 60 "in Wirklichkeit" geeignet (gleiche Basisrate), es steht aber diesmal nur ein Ausbildungsplatz zur Verfügung (andere Selektionsrate), dann wird die Entscheidung *zwangsläufig* 59 ebenfalls geeigneten, aber nicht zugelassenen Personen nicht gerecht werden *können*- auch wenn der Test ebenso gut ist wie im ersten Beispiel. Dies verdeutlicht den Effekt der Selektionsrate. Eine perfekte Vorhersagbarkeit unter dem Kriterium der Anzahl korrekter Entscheidungen ist nur dann möglich, wenn die Basisrate der Selektionsrate entspricht. Um so mehr die Basisrate und die Selektionsrate auseinanderdriften, um so weniger kann ein Test zu einer optimalen Anzahl korrekter Entscheidungen beitragen. Hat eine Verwaltung beispielsweise 100 geeignete Bewerber(innen) für **einen** Ausbildungsplatz, kann der Test die Anzahl korrekter Entscheidungen ebensowenig positiv beeinflussen, als wenn die Verwaltung keine(n) geeignete(n) Bewerber(in) für 100 Ausbildungsplätze hat.
Legt man hingegen besonderen Wert darauf, daß der Anteil erfolgreicher Personen an der Gesamtzahl ausgewählter Personen möglichst hoch ist, erhöht sich die Effizienz des Tests mit zunehmender Strenge der Selektion. Ein Test ist unter dieser Perspektive um so nützlicher, je strenger die Selektion ist. Betrachtet man die beiden oben genannten Fälle unter dem Aspekt des prozentualen Anteils der Erfolgreichen an allen Ausgewählten, so wird jedesmal der Nutzen des Tests auch gerade in Extremfällen unter Beweis gestellt: In dem einen Falle hat die Verwaltung durch den Testeinsatz eine erfolgreiche Person gefunden, in dem anderen Falle ist sie vor dem Schaden der Einstellung einer nicht-erfolgreichen Person bewahrt worden.

Nach dieser Einführung kann der Effekt der geringen Testleistungsunterschiede zwischen Bewerber(inne)n aus den beiden Teilen Deutschlands auf die Auswahlentscheidung (1.) vom Prinzip her dargestellt und (2.) auch modellhaft quantifiziert werden. Gruppenunterschiede in den Testleistungen bedeuten unter der vereinfachenden Annahme einer Gleichsetzung von Auslese-entscheidung und Eignungsurteil nichts anderes als Unterschiede in der Selektionsquote. Der relative Anteil der zur Ausbildung zugelassenen Personen an der Gesamtbewerber(innen)gruppe ist in den neuen Bundesländern kleiner

als in den alten Bundesländern. Abbildung 5 stellt den Effekt für zwei deutlich leistungsdisperate Gruppen dar. Die horizontale Linie trennt zwischen den im Kriterium, also z.B. in der Ausbildung, erfolgreichen und nicht-erfolgreichen Personen. Die vertikale Linie repräsentiert den "cut-off" des Tests: alle Personen mit einer schlechteren Leistung liegen links der vertikalen Linie. Bei allen diesen Personen ist es dem Test zufolge empfehlenswert, die Ausbildungszulassung zu verweigern. Die vertikale Ausrichtung der Ellipsen verdeutlicht, daß für beide Gruppen ein gleich enger Zusammenhang zwischen Test und Ausbildungserfolg angenommen wird. Die Ellipsen selbst stellen die beiden Gruppen dar. Die Gruppe in der rechten (schraffierten) Ellipse hat im Durchschnitt sowohl bessere Testleistungen als auch einen größeren Ausbildungserfolg. Die horizontale und die vertikale Linie teilen die Fläche in die vier bekannten Felder:
Im Feld "A" sind alle Personen, die ausreichende Testergebnisse erzielt haben und die in der Ausbildung erfolgreich waren (*valide positiv*).
Im Feld "B" sind alle Personen, die ausreichende Testergebnisse erzielen konnten, aber in der Ausbildung nicht erfolgreich waren (*falsch positiv*).
Im Feld "C" sind alle Personen, die keine ausreichenden Testergebnisse erzielt haben und die in der Ausbildung nicht erfolgreich waren (*valide negativ*).
Im Feld "D" sind alle Personen, die keine ausreichenden Testergebnisse erzielt haben, aber in der Ausbildung erfolgreich waren (*falsch negativ*).

Betrachtet man nun die Anteile der beiden Ellipsen an den vier Flächen, so sieht man, daß die Personen in der leistungsstärkeren Gruppe (schraffiert) häufiger dem Feld "A" zugeordnet werden. Das ist nach der gängigen Fairnessauffassung auch "in Ordnung", es handelt sich ja um die durchschnittlich im Test und im Kriterium (z.B. in der Ausbildung) erfolgreichere Gruppe. Entscheidend ist nun, daß im Vergleich zur leistungsschwächeren Gruppe ein größerer Anteil dieser leistungsstarken Gruppe (schraffiert) in das Feld "D" fällt, während die leistungsschwächere Gruppe relativ häufiger im Feld "B" landet. Die Felder "B" und "D" repräsentieren die beiden Fehlertypen in der Entscheidung. Rein aus entscheidungstheoretischer Sicht sind beide Felder als gleichwertige "Fehler" zu werten. Aus der Sicht des Individuums gibt es aber durchaus einen Unterschied zwischen den beiden Fehltertypen: In der leistungsstärkeren Gruppe (z.B. der Westdeutschen) können relativ mehr Personen von dem Entscheidungsfehler "B" "profitieren", sie werden als falsch Positive zur Ausbildung zugelassen, obwohl sie es nicht "verdient" haben, obwohl sie sich in der Ausbildung also nicht bewähren. (Wobei im Ernstfall eine zum Scheitern verurteilte Ausbildung ein zwiespältiger Profit ist.) In der leistungsschwächeren Gruppe müssen relativ mehr Personen den aus der Sicht des Individuums unangenehmsten Fehler "D" erleben: sie werden zu Unrecht zurückgewiesen, obwohl sie sich in der Ausbildung bewährt hätten. Aus der Sicht der Institution hat der Fehlertyp "B" folgende Bedeutung: Verwaltungen mit durchschnittlich testleistungsschwächeren Bewerber(inne)n, also ostdeutsche Verwaltungen, sind relativ häufiger damit konfrontiert, daß ihnen potentiell erfolgreiche Bewerber(innen) "durch die Lappen" gehen (Fehlertyp "D", falsch negativ). Die testbasierte Auswahl erweist sich bei dieser Gruppe als weniger sensibel für die potentiell Erfolgreichen unter den Bewerber(inne)n. Das heißt, die Trefferquote "Anzahl aller ausgewählten Positiven an der Gesamtzahl aller Positiven" fällt vergleichsweise ungünstiger aus. Dies gilt -wie oben explizit ausgeführt-, obwohl der Test in beiden Gruppen den Ausbildungserfolg gleich gut vorhersagt (Annahme 4).

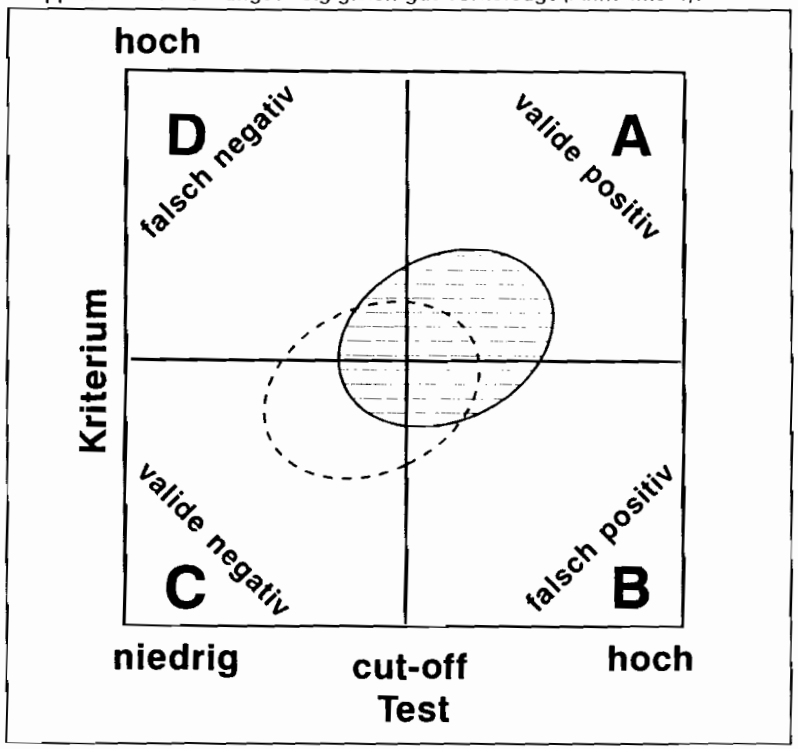


Abb 5.: Prinzipieller Effekt von Gruppenleistungsunterschieden auf die Vorhersageergebnisse bei der Verwendung fehlerbehafteter Prädiktoren (Abb. nach Hartigan et al., 1989, S.256)

Soweit das Prinzip. Wie groß ist dieser Effekt bei den hier ermittelten geringfügigen Gruppenmittelwertsunterschieden in den Testleistungen der Bewerber(innen) aus den alten und neuen Bundesländern aber tatsächlich? Die Testleistungsunterschiede zwischen Ost und West sind ja wesentlich geringer als die Unterschiede zwischen den Gruppen in dem allgemeinen Schaubild in Abbildung 5. Um diesen Effekt abzuschätzen, kann man das theoretische Bild aus Abbildung 5 zu einem Modellfall weiterentwickeln.
Ausgangspunkt für den Modellfall ist ein empirisches Ergebnis, nämlich die unterschiedliche Selektionsquote in Ost- und Westdeutschland. Bezogen auf die in Tabelle 2 dargestellte Gesamtgruppe von 500 Personen konnten sich

aufgrund der Testergebnisse 16,5 % (gerundet: 17 %) der für die Verwaltungen in den neuen Bundesländern getesteten und 26,9 % (gerundet: 27 %) der für die Verwaltungen in den alten Bundesländern getesteten Bewerber(innen) qualifizieren. Als "cut-off" wurde für die Berechnung der Selektionsquote ein Empfehlungsgrad von mindestens "4 -" angesetzt. Unberücksichtigt blieben in beiden Gruppen solche Personen, die sich im Vortest für den Haupttest qualifizieren konnten, zum Haupttest aber nicht erschienen sind.

Neben dieser empirischen Selektionsquote wird für die beabsichtigten Modellrechnungen nun noch der Wert für den Zusammenhang zwischen dem Testergebnis (dem Prädiktor) und dem Kriterium (dem Ausbildungs- und Berufserfolg) benötigt. Diese Größe müßte empirisch durch kulturvergleichende Bewährungskontrollen ermittelt werden (siehe oben). Da solche Ost-West-Untersuchungen der prognostischen Validität von Eignungsuntersuchungen zur Zeit nicht vorliegen, wurde hier für die Modellrechnungen auf die in Literaturzusammenstellungen berichtete mittlere Kriteriumsvalidität von kognitiven Fähigkeitstests rekurriert. Eingesetzt wurde der von Hunter und Hunter (1984, S. 81) für den Durchschnitt über alle Berufe berichtete Validitätskoeffizient für die Vorhersage des Ausbildungserfolgs von .54. Diese Korrelation wurde dann mit Hilfe der von Rosenthal (1984) vorgestellten Formel zur binominalen Effekt-Größendarstellung in ein vier-Felder-Schema transformiert. Die (gerundeten) Ergebnisse der Modellrechnung sind in Abbildung 6 dargestellt. In der Tabelle am Ende der Abbildung werden die einzelnen Indikatoren der Entscheidung miteinander verglichen. Da für beide Gruppen derselbe Korrelationskoeffizient zugrundegelegt wurde, resultiert zwangsläufig für beide Gruppen eine gleich hohe Anzahl an korrekten Entscheidungen (Anzahl der korrekt vorhergesagten erfolgreichen und der korrekt vorhergesagten nicht-erfolgreichen Personen). Mit 78 % bzw. 77 % ist auch die prozentuale Anzahl der erfolgreichen Personen unter allen ausgewählten Personen für beide Gruppen annähernd gleich. Diese Gleichheit entspricht einem fairen Auswahlverfahren nach dem Fairnessmodell der *"gleichen Wahrscheinlichkeiten"* von Linn (1973, zitiert nach Möbus, 1983) oder eben nach dem schon angesprochenen *"(Regressions)Modell der fairen Vorhersage"* nach Cleary, welches einen Spezialfall des Modells der gleichen Wahrscheinlichkeiten darstellt (siehe Möbus, ebd.). Der Begriff *"fair"* bezieht sich hier darauf, daß jede(r) voraussichtlich bessere Bewerber(in) den voraussichtlich weniger guten Bewerber(inne)n vorgezogen wird. Unter der Maßgabe, daß der Prozentsatz der Erfolgreichen unter den Akzeptierten in beiden Gruppen möglichst groß sein soll, kann man sagen, daß sich die Effekte der im Vergleich höhere Basisrate im Westen und der strengeren Selektion im Osten gegenseitig aufheben. Das bedeutet, daß der Testeinsatz den Verwaltungen in den neuen Ländern und in den alten Ländern mit gleicher Treffsicherheit garantiert, daß sich die empfohlenen Bewerber(innen) in der Ausbildung bewähren. Dies gilt aber nur solange, wie die Verwaltungen in den neuen Ländern ihre Bewerber(innen) trotz der ungünstigeren Basisrate an den gleichen Maßstäben wie die Verwaltungen in den Altbundesländern messen und somit eine strengere Selektion in Kauf nehmen. Ein gleicher Anteil an validen Positiven in den beiden deutschen Bewerber(innen)gruppen bedeutet angesichts der ungünstigeren Basisrate, daß die Verwaltungen in den neuen Ländern relativ gesehen mehr Kanidat-

Abb. 6: Beispiel für unterschiedliche Häufigkeiten der Vorhersagefehlertypen bei identischer Kriteriumsvalidität der in Ost und West eingesetzten Testverfahren

Beispiel aufgrund des an einer Stichprobe von 500 Personen ermittelten Verhältnisses zwischen abgelehnten und angenommenen Bewerber(innen) einerseits und einem (nach Hunter und Hunter, 1984) auf .54 geschätzten Zusammenhang zwischen Test und Ausbildungserfolg andererseits.

1. WEST

(von je 100)		TEST: abgelehnt (73)	TEST: angenommen (27)
KRITERIUM	Erfolg	17	21
	Mißerfolg	56	6

2. OST

(von je 100)		TEST: abgelehnt (83)	TEST: angenommen (17)
KRITERIUM	Erfolg	19	13
	Mißerfolg	64	4

3. Vergleich

	WEST	OST
Basisrate	38 %	32 %
Selektionsrate	27 %	17 %
Proportion korrekter Entscheidungen	77 %	77 %
Anteil valider Positiven an allen Selektierten	78 %	77 %
Anteil valider Positiven an allen potentiell Erfolgreichen	55 %	41 %

(inn)en testen lassen müssen als die Verwaltungen in den Altbundesländern. Um eine valide Positive, also eine Person mit einem ausreichenden Testergebnis, die sich in der Ausbildung bewähren wird, zu finden, müssen die Verwaltungen in den neuen Bundesländern den Zahlen der Modellrechnung entsprechend fast acht, die Verwaltungen in den Altbundesländern hingegen nur fünf Personen testen lassen. Die Annahme gleicher Korrelationskoeffizienten für beide Gruppen sichert identische Proportionen der *korrekten* Entscheidungen für die Bewerber(innen)gruppe aus Ost und West. Dennoch sorgen die Testleistungsunterschiede für Unterschiede in der Qualität der Entscheidungen: die beiden Fehlertypen der Entscheidung sind unterschiedlich häufig besetzt. Dies sieht man in der letzten Zeile der Tabelle, wo der Wert für den prozentualen Anteil der validen Positiven an allen berechnet ist: Während in den alten Bundesländern dieser Modellrechnung zufolge 55 % aller potentiell erfolgreichen Bewerber(innen) im Auswahlverfahren erkannt werden, sind es in den neuen Bundesländer nur 41 %. Dieser Fehlertyp spielt -wie oben erläutert wurde- aus der Sicht der Verwaltung kaum eine Rolle. Durch die fehlerhafte Zurückweisung entsteht der Verwaltung im Gegensatz zur fehlerhaften Zusage kein Schaden. Allerdings müßen mehr Personen getestet werden, finanziell stehen hier ca. 250.- DM pro getesteter Person im Falle einer fälschlichen Absage den fehlinvestierten Ausbildungskosten von ca. 120.000 DM pro Person im Falle einer fälschlichen Zusage entgegen.

Aus der Perspektive der Bewerber(innen) ist der unterschiedliche Fehlertypus hingegen relevant. Während für die westdeutschen Bewerber(inne)n eine im Vergleich zu ihren Brüdern und Schwestern aus den neuen Bundesländern erhöhte Gefahr besteht, **über**schätzt zu werden (Ausbildungszusage und späterer Mißerfolg), droht den ostdeutschen Bewerber(inne)n eher das "Schicksal", vergleichsweise **unter**schätzt zu werden (Absage, obwohl sie die Ausbildung geschafft hätten). Dies 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* in den beiden Gruppen gleich hohe Wahrscheinlichkeiten der Auswahl bestehen.

Die Modellrechnungen hatten zum Ziel, ein quantitatives Maß für diese potentielle Unfairness der Anwendung gleicher "cut-off"-Werte im Test für die Bewerber(innen) aus beiden Teilen Deutschlands zu gewinnen. Der in der untersten Tabelle der Abbildung 6 berichtete Unterschied in der Auswahl von 55 % aller potentiell erfolgreichen Bewerber(innen) im Westen und von 41 % im Osten beträgt 14 Punkte auf der Prozentskala. Dieser Wert relativiert sich, wenn man ihn in absolute Häufigkeiten überträgt. Dies liegt daran, daß es unter den Bewerber(inne)n der neuen Bundesländern insgesamt -dem Test zufolge- weniger potentiell erfolgreiche Personen gibt. Im Westen werden 45 % von 38, das sind **17**, im Osten 59 % von 32 potentiell erfolgreiche Bewerber(innen), das sind **19** Personen, ***nicht*** erkannt. (Diese unterschiedlich besetzten Felder sind in Abbildung 6 grau hinterlegt.) Es bleibt jeder/jedem Leser/in überlassen, zu urteilen, ob diese Unterschiede in der Häufigkeit des Fehlertyps als *kulturelle* Unfairness anzusehen sind. Dabei gilt es zu bedenken, daß für die **zu** Unrecht abgewiesene Person in der überwiegenden Zahl der Fälle nicht etwa Personen aus den alten Bundesländern herangezogen werden, sondern die Verwaltung -ggf. nach weiteren Ausschreibungen und Tests- jeweils eine andere Person aus der gleichen kulturellen Gruppe auswählt. Wie immer man urteilen mag: von dem Effekt der konservativen Vorhersage sind unter je 100 Personen pro Gruppe im Osten nur **zwei** Personen mehr als im Westen betroffen.

4. Überlegungen für den Fall, daß der Test Unterschiede in den wahren Werten wiedergibt: nicht-kognitive Variablen, Stichprobeneffekte und Probleme der Selbstselektion

Die Überlegungen und Analysen im zweiten Abschnitt dieses Artikels hatten zum Ziel, Anhaltspunkte für den Verdacht zu finden, daß der Test die tatsächliche durchschnittliche Leistungsfähigkeit der Bewerber(innen) aus den neuen Bundesländern unterschätzt. Auch in den Fällen wo Hinweise für eine solche Testverzerrung zu finden waren (bei der Testbearbeitungsgeschwindigkeit (Punkt 2.3), bei den Kenntnisfragen (Punkt 2.4.3)), ließen sich die kulturspezifischen Leistungsunterschiede nicht vollständig aufklären. Der vierte Abschnitt des Artikels soll nun Überlegungen für den Fall anbieten, daß der Test eine **unverzerrte** Messung darstellt. Das heißt für den Fall, daß die untersuchte Stichprobe an Bewerber(inne)n für Verwaltungen in den neuen Bundesländern tatsächlich im geringfügigen Maße testleistungsschwächer war als die zum Vergleich herangezogene westliche Gruppe. Die Ausführungen werden unter die Kategorien "nicht-kognitive Variablen", "Selbstselektion" und "Stichprobenbias" geordnet.

4.1 Nicht-kognitive Faktoren

Eine Erklärung für den Befund niedrigerer Testerwartungswerte in den neuen Bundesländern könnte z.B. im Nachweis unterschiedlicher Ausprägungen in solchen nicht-kognitiven Variablen liegen, die mit der Leistungsfähigkeit(smessung) zusammenhängen. Bereits im Heft 52 wurde in diesem Zusammenhang auf die Befunde von Wottawa (in Vorb.) hingewiesen, der berichtete, daß sich ost- und westdeutsche Führungskräfte bezüglich der Selbstdarstellung ihrer intellektuellen Effizienz (niedrigere Werte in den neuen Bundesländern) unterschieden. Auch Oettingen und Little (1993) fanden bei Ostberliner Schüler(inne)n vergleichsweise pessimistischere Selbsteinschätzungen des eigenen Leistungspotentials als bei Westberliner Schüler(inne)n. Selbst(wirksamkeits)einschätzungen stehen im Zusammenhang mit der Leistungsfähigkeit. Das Erleben von Belastungen kann die Leistungsfähigkeit ebenso beeinträchtigen. Pollmer und Hurrelmann (1992) zufolge war das schulische Belastungserleben bei den Schüler(inne)n in Sachsen größer als in Nordrhein-Westfalen. Über im deutsch-deutschen Vergleich höhere Angst- und Ärgerwerte bei Ostdeutschen berichten Hänsgen, Kasielke, Schmidt und Schwenkmezger (1992).

Es gibt eine Reihe weiterer nicht-kognitiver Faktoren, die -sofern sie zwischen der Bewerber(innen)gruppe "Ost" und "West" unterschiedlich ausgeprägt sind- die (Test-)Leistungsfähigkeit beeinträchtigen können. Die Wirkung dieser Variablen muß nicht direkt sein. Nicht-kognitive Faktoren können z.B. für ein unterschiedliches Bewerbungsverhalten in Ost- und Westdeutschland verantwortlich sein. Davon handelt der nächste Abschnitt.

4.2 Selbstselektion

Die Selbstselektion ist der Selektion logisch vorgeordnet. Bevor eine Person sich auf einen Ausbildungs- oder Arbeitsplatz bewirbt, wird sie Informationen und Eindrücke über die Ausbildung, den Beruf, über Arbeitsplätze und Organisationen sammeln, um sich die für sie "passenden" Verhältnisse, Bedingungen und Gegebenheiten ihrer zukünftigen Arbeit herauszusuchen (siehe z.B. Weinert, 1987).

Die Selbstselektion ist multidimensional determiniert. Dabei wird es beim Bewerbungsverhalten besonders auf das Interesse, auf berufsbezogene Werte und auf die Selbsteinschätzung eigener Fähigkeiten eines Individuums ankommen. Selbst wenn die unterschiedlichen Sozialisationsbedingungen in Ost- und Westdeutschland nicht zu unterschiedlichen Interessen geführt haben sollten -wie Todt (1992) aus der vergleichenden Auswertung einer Befragung zu allgemeinen und spezifischen orientierungsbezogenen Interessen schließt- so dürfte das Bewerbungsverhalten sich dennoch aufgrund zahlreicher anderer Komponenten in beiden Teilen Deutschlands unterscheiden. Neben einem eventuell schlechterem Informationsstand der Ostdeutschen in bezug auf "neue" Ausbildungen und Berufe wie z.B. "Diplom-Verwaltungswirt(in)" oder "Inspektoranwärter(in)", könnte vor allem auch das in Ost- und Westdeutschland offensichtlich unterschiedliche Arbeitsplatzsicherheitserleben und -bedürfnis entscheidend auf die Selbstselektion einwirken. Eine Ausbildung oder einen Beruf gemäß seinen Fähigkeiten und Interessen wählen kann nur der(die)jenige, der(die) es sich erlauben kann, "wählerisch" zu sein, der(die)jenige, der(die) sich existientiell sicher fühlt. Aus verschiedenen Ost-West-Vergleichsuntersuchungen können aber folgende Ergebnisse berichtet werden: Jugendliche im Osten sahen in der Arbeitslosigkeit im Vergleich zu Westjugendlichen häufiger ein schwerwiegenderes Problem (Heiliger und Kürten, 1992) und betonten in ihren Wertprioritäten den Aspekt der Sicherheit (Krebs, 1992). Über die subjektive Unsicherheit ostdeutsche Auszubildender in Fragen der beruflichen Zukunft berichten auch Palentien, Pollmer und Hurrelmann (1993). Hille (1993) zufolge nannten Jugendliche in den neuen Bundesländern wesentlich häufiger instrumentelle und materielle Gründe (Verdienst, Sicherheit, Aufstieg) für die Wahl des Berufs als die westdeutschen Jugendlichen. Im Zusammenhang mit den Werten von Jugendlichen kommt auch der Untersuchung von Boehnke (1993) eine wichtige Bedeutung zu, der die Werte von Lehrer(inne)n als "Wertmultiplikatoren" analysierte: von den ostdeutschen Lehrer(inne)n wurden Sicherheitswerte stark präferiert, im Westen standen dagegen Selbstbestimmungswerte im Vordergrund. Während Macharzina und Wolf (1994) für die Ostdeutschen keine vergleichsweise höhere Materialismus-Ausprägung nachweisen konnte, berichten Maier, Rappensperger, Rosenstiel und Zwarg (1994) für Examenskandidat(inn)en aus den neuen Bundesländern höhere Werte auf der Materialismusdimension als für die westdeutsche Vergleichsgruppe. Die Sicherheit des Arbeitsplatzes war dieser Gruppe vergleichsweise wichtiger als der Westgruppe. Über eine vergleichende Untersuchung zu beruflichen Werten berichten auch Borg, Braun und Häder (1993). Auch dieser Untersuchung zufolge ist die Sicherheit des Arbeitsplatzes den Ostdeutschen wichtiger als den Westdeutschen. Nach Schramm (1992) überschritt schon 1990 das Ausmaß der Arbeitsplatzunsicherheit im Osten für die Gesamtgruppe das im Westen für Problemgruppen gemessene Maß. Wilpert und Maimer (1993) stellen fest, daß der gesamte Bereich "Arbeit" im östlichen Teil Deutschlands eine größere Zentralität für die Menschen besitzt. Über die Ergebnisse einer repräsentativen Befragung in Mainz und Dresden, die u.a. das Konstrukt "Kontrollablehnung" im Licht innerdeutscher Vergleiche untersuchte, berichten Frese, Erbe-Heinbokel, Grefe, Rybowiak und Weike (1994). Unter "Kontrollablehnung" verstehen die Autor(inn)en die Ablehnung von Einflußmöglichkeiten auf Arbeitstätigkeiten und -bedingungen und interpretieren sie u.a. als Ausdruck von Resignation. Im Osten wiesen mehr als doppelt so viele Personen eine hohe Kontrollablehnung auf als im Westen.

Sicherlich ist diese Aufzählung von Befunden nur beschränkt aussagekräftig, sie ist außerdem bewußt selektiv vorgenommen worden. Es sollten Hinweise auf Faktoren gegeben werden, die (1.) in Ost und West unterschiedlich ausgeprägt sein können und die (2.) den Prozeß der Selbstselektion negativ beeinflussen können. Ohne die Befunde im Einzelnen zu werten, scheint die Annahme unterschiedlicher Prozesse der Selbstselektion in Ost und West (siehe Stratemann, 1992) berechtigt. Das größere Ausmaß an Arbeitslosigkeit und/oder die möglicherweise größere Existenzangst der Ostdeutschen kann in den neuen Ländern zu einem mehr oder minder wahllosen interessen- und begabungsunabhängigen Bewerbungsverhalten führen. Unterschiede in der Selbstselektion können dafür verantwortlich sein, daß die Bewerber(innen) in Ost und West nicht ohne weiteres miteinander vergleichbar sind, da sich hüben und drüben jeweils andere Gruppen der Bevölkerung bewerben. Sollten sich aber in den neuen Bundesländern im größeren Ausmaß Personen bewerben, die weniger gut zu der ausgeschriebenen Ausbildung/Stelle passen, wäre es verständlich, daß diese Personen in einem *berufsspezifischen* Test vergleichsweise schlechtere Durchschnittswerte erzielen als eine besser selbstselektierte Gruppe. Um diese Erklärung für Leistungsunterschiede geht es im nächsten Abschnitt.

4.3 Stichproben-Bias

Der Befund von Testleistungsunterschieden zwischen Bewerber(inne)n aus den neuen und den alten Bundesländern kann nur dann als allgemeiner "Ost-West"-Leistungsunterschied interpretiert werden, wenn die Stichproben je-

weils repräsentativ für die Grundgesamtheit sind (*"Sampling equivalence"*, z.B. bei Butcher, 1982). Es spricht vieles dafür, daß dies nicht so ist. Neben den oben genannten Überlegungen zur Selbstselektion, sind hier die deutschen Binnenwanderungsprozesse zu nennen, die z.B. bei Maretzke und Müller (1992) beschrieben und im letzten Heft der "*DGP*-Informationen" referiert wurden. 1991 sind demnach über 200.000 Personen aus den neuen in die alten Länder gezogen. Diese Abwanderungen wurden nur zu ca. 25% durch Zuzüge aus den alten Ländern in die neuen Länder kompensiert.

Einem anderen Aspekt wurde im ersten Teil des Artikels noch keine Rechnung getragen, nämlich der in den neuen und alten Bundesländern möglicherweise unterschiedlichen Anzahl potentieller Bewerber(innen). Gegenstand der Untersuchung waren Bewerber(innen) für die Ausbildung des gehobenen Verwaltungsdienstes. Als Einstellungsvoraussetzung gilt hier u.a. die Fachhochschulreife oder eine andere zum Hochschulstudium berechtigende Schulbildung oder ein als gleichwertig anerkannter Bildungsstand. Die Frage ist nun, ob der prozentuale Anteil der Bevölkerung, die diese Einstellungsvoraussetzungen erfüllt, in den neuen und in den alten Bundesländern gleich groß ist. Leider lagen dem Autor bei Redaktionsschluß keine verläßlichen und differenzierten Zahlen vor[7]. Ein erster Hinweis geht aber in die folgende Richtung: Während in der alten Bundesrepublik das Abitur den "Normalfall" der gehobenen Schulbildung darstellte und zur Zeit nach Angaben von Apel (1992, S. 361) ca. ⅓ aller Jugendlichen eines Jahrgangs das Abitur erwerben, erwarben in den letzten Jahren vor der Vereinigung nur ca. 12-13 Prozent der DDR-Jugendlichen die Hochschulreife (Apel, ebd.). Dies würde bedeuten, daß der Anteil der Personen, die die Eignungsvoraussetzung erfüllen an der Gesamtbevölkerung in den neuen Bundesländern prozentual gesehen kleiner ist als in den alten Bundesländern. Geht man von der Gleichverteilung der Leistungsfähigkeit innerhalb der Zugangsberechtigten in beiden Teilen Deutschlands aus, muß damit in den neuen Bundesländern auch der Anteil potentiell erfolgreicher Personen kleiner sein (Basisrate). Die Verwaltungen in den neuen Bundesländern müssen, um die gleiche Prozentzahl an erfolgreichen Bewerber(inne)n wie die Altbundesländer zu erreichen, einen höheren Prozentsatz aller Zugangsberechtigten -und somit tendenziell auch die unteren Ränge der Normalverteilung- ins Auswahlverfahren nehmen. Dies kann sich dann in den niedrigeren Testerwartungswerten niederschlagen.

Vor dem Hintergrund der möglicherweise unterschiedlichen Selbstselektion, der Binnenwanderungsprozesse in Deutschland und der Unterschiede in der prozentualen Anzahl von Zulassungsberechtigten, erscheint es sehr unwahrscheinlich, daß es sich bei den untersuchten Stichproben um repräsentative Ost-West-Gruppen handelt.
Der Stichproben-Bias würde auch erklären, warum Forscher(inn)en, die u.a. die intellektuelle Leistungsfähigkeit von weniger selektierten Stichproben

[7]für entsprechende Hinweise und Quellenangaben an das *DGP*-Büro Leipzig wäre der Autor dankbar.

(Schüler(inn)en, Student(inn)en) mit vergleichbaren Instrumenten untersuchten, (fast) keinen Ost-West-Leistungsunterschied feststellen konnten. (Z.B. Baumert, Köller und Lehrke, 1992; Strohschneider, 1992, 1994).
Die hier dargestellte These des "Stichproben-Bias" ließe sich einfach überprüfen, indem der *DGP*-Test mit Schüler(innen) der Abschlußklassen ost- und westdeutscher Gymnasien durchgeführt würde. Da es sich dabei um nichtselektierte Stichproben handelt, dürfte sich dieser These zufolge dann kein Ost-West-Leistungsunterschied zeigen.

5. Zusammenfassung, Konsequenzen und Ausblick

Die Befunde lassen sich wie folgt zusammenfassen:

✓ Die nachgewiesenen durchschnittlichen Testleistungsunterschiede zwischen den Bewerber(inne)n aus den neuen und alten Ländern sind quantitativ als geringfügig einzustufen. In Ergänzung hierzu muß erwähnt werden, daß der kulturelle Gruppenleistungsunterschied sich allerdings als zeitlich stabiler erweist als erwartet. Eine Auswertung der Ergebnisse von 3457 (1241 aus den neuen Bundesländern) Vortestkandidat(inn)en (gehobener Dienst) der Saison 1993/1994 ergab, daß im Westen 46,4 % und im Osten 31,3 % den für die Zulassung zum Haupttest von den Ostverwaltungen gesetzten kritischen Punktwert von 80 erreichten.

✓ Die Annahme der Invarianz der Faktorenstruktur über die beiden Gruppen hinweg erscheint plausibel.

✓ Die Westbewerber(innen) in der Stichprobe bearbeiteten in gleicher Zeit durchschnittlich geringfügig mehr Items der untersuchten Aufgaben als ihr Gegenüber aus dem Osten. Dieser Effekt konnte die Leistungsunterschiede zwischen den beiden Gruppen aber nicht vollständig aufklären, da die Westdeutschen im Durchschnitt auch eine günstigere Trefferquote pro bearbeitetem Item aufwiesen.

✓ Die Frage, ob die einzelnen Items in den beiden Gruppen unterschiedlich wirken, muß weiterhin offen bleiben. Die beiden Untersuchungen zu dieser Fragestellung (Itemanalysen der Sprachtests, Variation des Diktats) zeitigten zwar einige diese Richtung weisende Befunde, diese Ergebnisse bedürfen aber noch der Kreuzvalidierung.

✓ Die untersuchten Bewerber(innen) aus den neuen Ländern haben von einer kulturellen Anpassung der Kenntnisfragen an die jüngste gemeinsame Vergangenheit profitiert. Dies spricht dafür, daß die durchschnittlich höheren Erwartungswerte bei den westdeutschen Bewerber(inne)n zum Teil auf eine größere Vertrautheit mit dem Gegenstand der Kenntnisfragen zurückzuführen waren. Auch nach der Modifikation erwiesen sich die durchschnittlichen gemeinschaftskundlichen Kenntnisse der Westdeutschen aber im deutsch-deutschen Vergleich als fundierter.

- ✓ Zahlreiche empirische Untersuchungen an kulturell sehr unterschiedlichen Gruppen begründen die Hoffnung, daß der Test in beiden Teilen Deutschlands für eine gleich gute Vorhersage des Ausbildungserfolgs sorgt.

- ✓ Im Gruppenvergleich durchschnittlich niedrigere Testerwartungswerte führen bei der Personalauswahl prinzipiell dazu, daß auch bei gleicher Vorhersagekraft der Prädiktoren potentiell in der Ausbildung erfolgreiche Bewerber(innen) der leistungsschwächeren Gruppe vergleichsweise häufiger zu Unrecht von der Ausbildung/dem Beruf ausgeschlossen werden. Mit Modellrechnungen konnte aber gezeigt werden, daß im Vergleich von jeweils 100 West- und Ost-Bewerber(inne)n nur zwei Ostdeutsche mehr als Westdeutsche von diesem Fehler betroffen sind.

- ✓ Die Literaturhinweise dafür, daß es Ost-West-Unterschiede in nicht-kognitiven Variablen gibt, verdichten sich. Einige dieser Variablen könnten das Bewerbungsverhalten der ehemaligen DDR-Bürger(innen) beeinflussen. Insbesondere die "Selbstselektion" der Bewerber(innen) in den neuen Ländern könnte davon negativ betroffen sein. Neben der Selbstselektion können auch die deutschen Binnenwanderungsprozesse und die im Vergleich zum Westen in den neuen Bundesländern prozentual kleinere Anzahl an Personen, die die Einstellungsvoraussetzungen erfüllen, mit dafür verantwortlich sein, daß es sich bei den untersuchten Stichproben nicht um einen repräsentativen Ost-West-Vergleich, sondern um den Vergleich spezifischer Gruppen aus Ost und West handelt. Die thematisierten Leistungsunterschiede würden dann im wesentlichen Unterschiede in der Stichprobenzusammensetzung widerspiegeln.

Welche Konsequenzen lassen sich aus diesen Befunden ableiten? Eine Standardreaktion auf den Bericht von Gruppenleistungsunterschieden ist die Forderung nach gruppenspezifischen "Korrekturen" der Leistungsergebnisse. Diese Forderungen wurden aber zumeist in solchen Fällen formuliert, in denen die Leistungsunterschiede zwischen den Gruppen wesentlich ausgeprägter sind als im vorliegenden Fall (z.B. Hartigan et al., 1989, S.283). Unter dem Postulat, daß es keine Gruppenleistungsunterschiede zwischen "Ost" und "West" geben *darf*, kann man die Ergebnisse einer Person auf die Mittelwerte und Streuungen *ihrer* Kulturgruppe beziehen und somit z.B. Normen für Angehörige der fünf neuen Bundesländer einführen. Man kann auch fordern, bei der Selektion in den neuen Bundesländern ein weniger strenges Auswahlkriterium zugrundezulegen oder ein Bonus/Malus-System einzuführen usw.
Bei jeder "Korrekturmaßnahmen" muß man sich zunächst fragen, ob eine im Vergleich zu Westdeutschland nur äußerst geringfügig höhere Betroffenenquote in Ostdeutschland eine solche Maßnahmen überhaupt rechtfertigen kann. Ob eine Korrekturmaßnahmen gerechtfertigt ist, wenn es zahlreiche Anhaltspunkte dafür gibt, daß -nicht die Ostdeutschen, sondern:- die spezielle Gruppe der Bewerber(innen) aus den neuen Bundesländern aus Gründen einer weniger geglückten Selbstselektion, aus Gründen der ungleichen Verteilung der Zulassungsberechtigungen und aus Gründen der Binnenwanderungsprozesse tatsächlich geringfügig leistungsschwächer ist?

Jegliche Form einer "gruppenspezifischen Testauswertung" wirft ausufernde Zuordnungsprobleme auf. Die in Abschnitt 3 des Artikels beschriebenen Effekte sind ja nicht auf die Testteilnehmer(innen) aus den neuen Bundesländern begrenzt, sondern betreffen alle Gruppen, die gegenüber einer anderen Gruppe im Durchschnitt geringere Testerwartungswerte aufweisen. Mit der Einführung von Normen für die neuen Bundesbürger(innen) würde den -gleich begründeten- Forderungen zahlreicher anderer "Gruppen" nach spezifischen Normen das Tor geöffnet. Normen für Frauen gegenüber Männern, für Fach- und Abendschüler(innen) gegenüber Gymnasiast(inn)en, für die Landbevölkerung gegenüber der Stadtbevölkerung, für Sozialschwache gegenüber Wohlhabenden usw.. Auch Kombinationen sind denkbar, z.B. die Forderung nach speziellen Tests, Auswertungen und/oder Selektionskriterien für sozialschwache Frauen, die in ländlichen Regionen der neuen Bundesländer aufgewachsen sind und dort ihre "Berufsausbildung mit Abitur" erworben haben - falls gerade diese Gruppe im Durchschnitt schlechte Testleistungen erzielen sollte. Die Einführung von Differenzierungen sollte äußerst sparsam vorgenommen werden. In der Eignungsdiagnostik hat man sich bislang überwiegend auf unterschiedliche Tests und Anforderungswerte für verschiedene Berufsbilder und auf die Einführung von Altersnormen beschränkt. Beides ist unter dem Gesichtspunkt gerechtfertigt, daß die Intelligenzentwicklung über die Lebensjahre und die Anforderungen bestimmter Berufstätigkeiten in ihren wesentlichen Komponenten hinreichend erforscht sind. Dies dürfte für mehr oder minder beliebig gebildete Gruppen nicht gelten.
Solange die Annahme gleich guter Vorhersagbarkeit des Ausbildungserfolges in Ost- und Westdeutschland aufrechterhalten wird und solange es das Ziel des Auswahlverfahrens ist, unter den ausgewählten Personen eine maximale Anzahl an Erfolgreichen zu finden, solange ist eine "mildere" Selektion in den neuen Bundesländern kontraindiziert. Unter diesen Prämissen können aus den Befunden folgende Konsequenzen für eine verantwortungsvolle Personalauswahl in Ostdeutschland gezogen werden:

☞Tests:

Der Einfluß der Zeitbegrenzung der Tests und der Einfluß der Sprache sowie der abgebildeten Kenntnisdomänen ist nicht nur für die Gesamtgruppe, sondern auch für die Untergruppe der Ost- und Westdeutschen zu kontrollieren. Die *DGP* führt zur Zeit weitere Itemanalysen durch, bei der auch diese Thematik berücksichtigt wird. Darüber hinaus ist beabsichtigt, bei einigen Gruppen in Zukunft zusätzlich zur Testbatterie der *DGP* zu Forschungszwecken auch einen Test zum Berliner Intelligenzstrukturmodell nach Jäger (1982, 1984) zu applizieren. Dies ermöglicht es, die differenzierte Wirkung unterschiedlicher Tests zum (teilweise) gleichen Konstrukt in den Kulturgruppen zu untersuchen. Schließlich wird die *DGP* bei den anstehenden Testrevisionen und -neuentwicklungen den Aspekt der Gefahr einer kulturellen Testverzerrung von Anfang an berücksichtigen. Die Frage nach einer möglicherweise differentiellen oder singulären Kriteriumsvalidität des *DGP*-Tests ist nur aufgrund der Ergebnisse zukünftiger Ost-West-Bewährungsuntersuchungen zu beantworten.

☞Bewerbungen

Die Verwaltungen in den neuen Bundesländern sollten durch verständliche und ausführliche Informationen den Prozeß der Selbstselektion der Bewerber(innen) in den neuen Bundesländern positiv beeinflussen. Angestrebt werden sollte auch eine sachgerechte "Image"werbung, so daß die Verwaltungen für potentiell geeignete Personen attraktiver werden. Besonders den Bewerber(inne)n aus den neuen Bundesländern muß die Möglichkeit eingeräumt werden, sich mit Hilfe der Broschüre *"Bewerberinnen und Bewerber fragen - die DGP antwortet"* auf das Testverfahren vorzubereiten.

Potentiell erfolgreiche Bewerber(innen) für die Verwaltungen der neuen Bundesländer sollten den Effekt, daß sie im Vergleich zu Bewerber(inne)n aus dem Westen etwas stärker von der Gefahr bedroht sind, im Auswahlverfahren "unterschätzt" zu werden, kompensieren, indem sie sich häufiger bewerben.

Es wäre aufschlußreich, solche Informationen über die Bewerber(innen) aus den neuen und den alten Bundesländern zu erheben und im Ost-West-Vergleich auszuwerten, die über die bloße Leistungsmessung und über die Angabe demographischer Merkmale hinausgehen. Gedacht werden sollte dabei an verschiedene nicht-kognitive Merkmale, wie z.B. Bewerbungsverhalten, Einstufung der Attraktivität der ausgeschriebenen Stelle und der Verwaltung aus Sicht der Bewerber(innen), (Test- und Existenz-)Angst, Selbstwirksamkeitserwartungen usw.). Pragmatisch -wenngleich mit Auswertungsnachteilen verbunden- wäre hier eine anonymisierte Befragung während der Eignungsuntersuchungen. Verwaltungen, die an einer solchen Beschreibung ihrer Bewerber(innen) und an Beschreibungen der Verwaltung durch die Bewerber(innen) interessiert sind, möchten sich bitte an die *DGP* wenden, damit man über ein gemeinsames Vorgehen in der Erhebung nachdenken kann.

☞Personalauswahl

Die *DGP* geht vorläufig davon aus, daß die *DGP*-Tests in beiden Teilen Deutschlands eine gleich gute Vorhersage des Ausbildungs- und Berufserfolgs ermöglichen. Verwaltungen aus den neuen Bundesländern müssen daher an den Selektionsmaßstäben der Altbundesländer festhalten, wenn sie erreichen wollen, daß der Anteil an Erfolgreichen unter den den Ausgewählten ebenso hoch ist wie im Westen. Die ungünstigere Basisrate muß unter dieser Zielvorgabe zu einer geringeren Selektionsrate führen. Dies bedeutet, daß die Verwaltungen in den neuen Bundesländern für eine bestimmte Anzahl an z.B. Ausbildungsplätzen angesichts der ungünstigeren Basisrate im Durchschnitt mehr Personen testen müssen als Westverwaltungen in vergleichbaren Situation.

♦

5. Literatur

Apel, H. (1992). Integrative Bildungsmobilität in den alten und neuen Bundesländern. In Jugendwerk der Deutschen Shell (Hrsg.), Jugend '92 (Bd. 2, Im Spiegel der Wissenschaften, S. 353-370). Opladen: Leske + Budrich.

Bartussek, D. (1982). Modelle der Testfairness und Selektionsfairness (Trierer Psychologische Berichte). Trier: Universität Trier.

Baumert, J., Köller, O. & Lehrke, M. (1992). Entwicklung von Kontrollüberzeugungen unter unterschiedlichen Schul- und Systembedingungen. In L. Montada (Hrsg.), Bericht über den 38. Kongreß der Deutschen Gesellschaft für Psychologie in Trier 1992 (Bd. 1, S. 628). Göttingen: Hogrefe.

Becker, P. (1991). Beim freien Text die alten Fehler. Der Tagesspiegel, 13933.

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.

Boehnke, K. (1993). Lehrerinnen und Lehrer als Wertmultiplikatoren im veränderten Bildungssystem der neuen Bundesländer: Probleme und Perspektiven. Pädagogik und Schulalltag, 48, 94-105.

Borg, I., Braun, M. & Häder, M. (1993). Are East Europeans ready for a market economy? Some answers from work values in East und West Germany (Fourth International Facet Theory Conference). Prague, Czech Republic: Facet Theory Association, 38-48.

Bortz, J. (1989). Lehrbuch der Statistik. Berlin: Springer.

Butcher, J.N. (1982). Cross-cultural research methods in clinical psychology. In P.C.Kendall & J.N. Butcher (Hrsg.), Handbook of research methods in clinical psychology (S. 273-308). New York: J. Wiley.

Dorsch, F. (1982). Psychologisches Wörterbuch. (10. Auflage) Bern: Huber.

Eberle, G. (1980). Analysen zur Speed-Power-Problematik beim Mannheimer Intelligenztest für Kinder und Jugendliche (MIT-KJ). Frankfurt /M.: Lang.

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.

Fay, E. (1991). Die Entwicklung der Studienplatz- und der Bewerberzahlen sowie der Zulassungsgrenzen in den medizinischen Studiengängen. In G.Trost (Hrsg.), Test für medizinische Studiengänge (TMS):Studien zur Evaluation. 15. Arbeitsbericht (S. 148-170). Bonn: Institut für Test- und Begabungsforschung.

Frese, M., Heinbokel-Erbe, M., Grefe, J., Rybowiak, V. und Weike, A. (1994). "Mir ist es lieber, wenn ich genau gesagt bekomme, was ich tun muß": Probleme der Akzeptanz von Verantwortung und Handlungsspielraum in Ost und West. Zeitschrift für Arbeits- und Organisationspsychologie, 38, 22-33.

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

Guldin, A. (1991). Vorschlag eines Ordnungskonzeptes zur Testfairness. In H. Schuler & U. Funke (Hrsg.), Eignungsdiagnostik in Forschung und Praxis (S. 216-220). Stuttgart: Verlag für Angewandte Psychologie.

Hänsgen, K.D., Kasielke, E., Schmidt, L.R. & Schwenkmezger, P. (1992). Ostdeutsche und Westdeutsche im Vergleich: Emotionalität und Objektive Persönlichkeitsvariablen. Zeitschrift für Klinische Psychologie, Psychopathologie und Psychotherapie, 40, 346-363.

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 standardizes 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.
Hesse, J. & Schrader, H.L.(1988). Erlebnisse aus 1001 Bewerbungen. Frankfurt/M.:Eichborn.
Hille, B. (1993). Lebenssituation und Lebensperspektiven Jugendlicher im vereinten Deutschland. Aus Politik und Zeitgeschichte, 24, 14-20.
Hilliard, Asa G. (1984). IQ Testing as the Emperor's New Clothes. A Critique of Jensen's Bias in Mental Testing. In C.R.Reynolds & R.T.Brown (Hrsg.), Perspectives on Bias in Mental Testing (S. 139-169). New York: Plenum Press.
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. (1976). Critical Analysis of the Statistical and Ethical Implications of Various Definitions of Test Bias. Psychological Bulletin, 83, 1053-1071.
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.
Hunter, J.E., Schmidt, F.L. & Rauschenberger, J.M. (1977). Fairness of psychological tests: Implications of four definitions for selection utility and minority hiring. Journal of Applied Psychology, 62, 245-260.
Hunter, J.E., Schmidt, F.L. & Rauschenberger, J. (1984). Methodological, Statistical, and Ethical Issues in the Study of Bias in Psychological Tests. In C.R.Reynolds & R.T.Brown (Hrsg.), Perspectives on Bias in Mental Testing (S. 41-99). New York: Plenum Press.
Iseler, A. (1970). Leistungsgeschwindigkeit und Leistungsgüte. Weinheim: Beltz.
Jäger, A.O. (1970). Personalauslese. In K.Gottschaldt, Ph. Lersch. F.Sander & H.Thomae (Hrsg.), Handbuch der Psychologie (Bd.9, Betriebspsychologie, S.613-667). Göttingen: Hogrefe.
Jäger, A.O. (1982). Mehrmodale Klassifikation von Intelligenzleistungen. Experimentell kontrollierte Weiterentwicklung eines deskriptiven Intelligenzstrukturmodells. Diagnostica, 28, 195-226.
Jäger, A.O. (1984). Intelligenzstrukturforschung: Konkurrierende Modelle, neue Entwicklungen, Perspektiven. Psychologische Rundschau, 35, 21-35.
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. (1993). Personalauswahl in den neuen Bundesländern: Tests ohne Grenzen oder Grenzen der Tests? Eine vergleichende Analyse der durchschschnittlichen Testergebnisse in Ost und West (Teil I) (*DGP* Informationen). Hannover: Deutsche Gesellschaft für Personalwesen, 52, 62-87.
Klieme, E. & Stumpf, H. (1991). DIF: A Computer Programm for the Analysis of Differential Item Performance. Educational and Psychological Measurement, 51, 669-671.
Kong, S.L. (1989). Understanding and coping with individual differences. In R.J.Samuda, S.L.Kong, J.Cummins, J.Pascual-Leone & J.Lewis (Hrsg.), Assessment and placement of minority students (S. 69-93). Göttingen: Hogrefe.
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.
Macharzina, K. & Wolf, J. (1994). Materialismus und Postmaterialismus in den neuen Bundesländern. Zeitschrift für Arbeits- und Organisationspsychologie, 38, 13-21.
Maier, G.W., Rappensperger, G., Rosenstiel, L. & Zwarg, I. (1994). Berufliche Ziele und Werthaltungen des Führungsnachwuchses in den alten und neuen Bundesländern. Zeitschrift für Arbeits- und Organisationspsychologie, 38, 4-12.
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.
Petersen, N.S. & Novick, M.R. (1976). An Evaluation of Some Models for Culture-fair Selection. Journal of Educational Measurement, 13, 3-29.
Pollmer, K. & Hurrelmann, K. (1992). Neue Chancen oder neue Risiken für Jugendliche in Ostdeutschland? - Eine vergleichende Studie zur Streßbelastung sächsischer und nordrhein-westfälischer Schülerinnen und Schüler. Zeitschrift für Sozialisationsforschung, 12, 2-29.
Poortinga Y.H. (1983). Psychometric approaches to intergroup comparison: The problem of equivalence. In S.H. Irvine & J.W. Berry (Hrsg.), Human assessment and cultural factors (S. 237-257). New York: Plenum Press.
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 Press.
Samuda, R.J. (1989). Psychometric factors in the appraisal of intelligence. In R.J.Samuda, S.L.Kong, J.Cummins, J.Pascual-Leone & J.Lewis (Hrsg.), Assessment and placement of minority students (S. 25-40). Göttingen: Hogrefe.
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.
Schramm, F. (1992). Beschäftigungsunsicherheit. Wie sich die Risiken des Arbeitsmarktes auf die Beschäftigten auswirken - Empirische Analysen in Ost und West. Berlin: Edition Sigma.
Schuler, H. & Funke, U. (1989). Berufseignungsdiagnostik. In E.Roth (Hrsg.), Organisationspsychologie. Enzyklopädie der Psychologie. Themenbereich D, Praxisgebiete. Serie III 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.
Stepper, S. & Strack, F. (1993). Stereotype Beurteilung von Ost- und Westdeutschen: Der Einfluß der Stimmung. Zeitschrift für Sozialpsychologie, 24, 218-225.
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. (1992). Vom doofen Wessi und schlauen Ossi - und umgekehrt: Intelligenzunterschiede und strategische Unterschiede beim Umgang mit einem komplexen Problem (Unveröffentlichter Vortrag auf der 34. Tagung experimentell arbeitender Psychologen in Osnabrück).
Strohschneider, S. (1994). Strategien beim Umgang mit einem komplexen Problem: Ein deutsch-deutscher Vergleich. Zeitschrift für Arbeits- und Organisationspsychologie, 38, 34-40.
Tabachnick, B.G. & Fidell, L.S.(1989). Using multivariate statistics. New York:Harper & Row.
Todt, E. (1992). Interesse männlich - Interesse weiblich. In Jugendwerk der Deutschen Shell (Hrsg.), Jugend '92 (Bd. 2, Im Spiegel der Wissenschaften, S. 301-317). Opladen: Leske + Budrich.
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.
Vernon, P.E. (1979). Intelligence: Heredity and Environment. San Francisco: Freeman.

Vijver, F.J.R. van de & 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.
Weber, W. (1986). Entwicklung und Aufgabenanalyse des Diktats "Auf einem Amt" (*DGP* Informationen), 46, 106-122.
Weinert, A.B. (1987). Lehrbuch der Organisationspsychologie. München: Psychologische Verlags Union.
Wiggins, J.S. (1973). Personality and Prediction: Principles of Personality Assessment. Reading, MA: Addison-Wesley.
Wilpert, B. & Maimer, H. (1993). Culture or society? A research report on work-related values in the two Germanies. Social-Science Information, 32, 259-278.
Wottawa, H. (in Druck). Veränderungen und Veränderbarkeit berufsrelevanter Eigenschaften im Ost-West Vergleich.
Wottawa, H. & Amelang, M. (1980). Einige Probleme der "Testfairness" und ihre Implikationen für Hochschulzulassungsverfahren. Diagnostica, 26, 199-221.