

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

Martin Kersting

1. Einführung

Daß es leichter ist, einen Atomkern zu spalten als ein Vorurteil - diese Albert Einstein zugeschriebene Erkenntnis wird in zahlreichen Auseinandersetzungen mit Themen der deutsch-deutschen Wiedervereinigung leider allzuoft bestätigt. Zu diesen Themen gehört auch die Personalauswahl in den neuen Ländern. Einige "Besserwessis" und -mindestens ebenso schlimm- "Ossi-Beschützer-Wessis" haben sich ihre Meinung zu den Qualitäten der Bewerber(innen) aus den neuen Ländern und zu den richtigen Methoden ihrer Auswahl längst schon aufgrund einer Kombination aus Gelegenheitseindrücken und Wunschvorstellungen gebildet -ebenso wie einige "Ossis" wissen, was sie davon zu halten haben.

Dieser Hintergrund erschwert einerseits die nüchterne Auseinandersetzung mit den Möglichkeiten und Grenzen einer verantwortungsvollen Personalauswahl in den neuen Ländern. Andererseits wächst aber gerade mit den Vorurteilen auch die Notwendigkeit, anstelle von Ideologie Wissenschaftlichkeit zu etablieren. Kann man, so lautet die Frage, die in den alten Ländern entwickelten und bewährten Verfahren so ohne weiteres auf die neuen Länder übertragen oder führt dies nicht möglicherweise zu einer systematischen "Diskriminierung" der Bewerber(innen) aus den neuen Ländern?

Der vorliegende Bericht liefert erste Bausteine zur Beantwortung dieser Frage, indem die Daten von über 1000 Bewerber(innen) aus den neuen und den alten Ländern auf Gruppenunterschiede hin analysiert werden und indem ein Interpretationsraum für die gefundenen Ergebnisse ausgelotet wird. Der hier abgedruckte erste Teil des Berichts endet mit der Formulierung von Hypothesen, mit denen die Gruppenunterschiede erklärt werden **könnten**. Die Ergebnisse der Hypothesenprüfungen werden im zweiten Teil dieses Berichtes dargestellt, der im nächsten Heft der "DGP-Informationen" veröffentlicht wird.

2. Vergleichende Analyse der Leistungsdaten von Bewerber(innen) aus den neuen und aus den alten Bundesländern

2.1 Stichprobe und Meßinstrumente

Seit 1990 setzt die *Deutsche Gesellschaft für Personalwesen (DGP)* ihre psychologischen Untersuchungsverfahren zur Prognose des Ausbildungserfolges auch in den neuen Ländern ein. Grundlage der folgenden vergleichenden Analysen sind die Ergebnisse des sogenannten "Vortests" von insgesamt 1377 Bewerber(innen) für die Ausbildung zum gehobenen nicht-

technischen Verwaltungsdienst aus der Bewerbungssaison 1991/1992.¹ 696 Bewerber(innen) mit einem Durchschnittsalter von 20,6 Jahren hatten sich in den drei ostdeutschen Bundesländern Brandenburg (270), Sachsen-Anhalt (248) und Mecklenburg-Vorpommern (178) beworben. 681 Personen waren Kandidat(inn)en der Alt-Bundesländer (Land Niedersachsen (288), Stadt Hannover (155) und Schleswig-Holstein (238)) mit einem Durchschnittsalter von 21,8 Jahren. Sowohl in der Ost- als auch in der Weststichprobe waren Frauen mit etwa 56 % der jeweiligen Gesamtgruppe etwas häufiger vertreten als Männer. Ein Teil der hier genannten demographischen Angaben sowie die Häufigkeitsverteilung für den Schulabschluß der Bewerberinnen und Bewerber ist in Tabelle 1 wiedergegeben.

Tab. 1: Demographische Daten zur Vortest-Stichprobe

Vortest, gehobener Dienst, Saison '91 / '92; N = 1377						
	OST			WEST		
	Frauen	Männer	Gesamt	Frauen	Männer	Gesamt
EOS / Abitur	317	189	506	246	149	395
Fach(hoch)schule	28	28	56	132	154	286
Beruf mit Abitur	43	79	122			
Sonstige	2	10	12			
Gesamt	390	306	696	378	303	681
Altersdurchschnitt	19,5	22	20,6	20,6	23,3	21,8

Der Vortest stellt eine Art "screening" dar, welches eine erste Vorauswahl der Kandidat(inn)en gestattet. In ca. drei Stunden werden mit zwei standardisierten Tests zum sprachlichen Denken, mit drei Tests zum rechnerischen Denken, mit einem Test zum Arbeitsverhalten und mit je einer schriftlichen Überprüfung der Rechtschreibung, der Grundrechen- und der gegenwarts-kundlichen Kenntnisse die Grundfertigkeiten und Grundfähigkeiten der Bewerberinnen und Bewerber überprüft.

Bei allen schriftlichen Tests der *DGP* wird eine Leistungsskala von 70 (schlechtester Wert) bis 130 (bester Wert) zugrundegelegt, mit einer sogenannten Standardabweichung von 10. Im Vortest werden die Leistungen der Bewerber(innen) zusätzlich in jedem der insgesamt neun Tests noch in Form von Punkten bewertet. Der Gesamtpunktwert bildet dann die Grundlage für die Zulassung zum Haupttest.

¹Die Datenbasis wurde von verschiedenen Mitarbeiter(innen) der *DGP* in Hannover zusammengetragen, codiert und abgelocht. Hervorzuheben ist vor allem die Arbeit von Herrn *Michael Linke*, der durch die umsichtige Gestaltung der Ablochpläne und die überaus gewissenhafte und sorgfältige Dateneingabe die Voraussetzungen für einen soliden empirischen Vergleich geschaffen hat.

548 der in die Analyse des Vortests einbezogenen 1377 Bewerber(innen) wurden zum sogenannten Haupttest zugelassen, wo sie erneut eine -diesmal umfangreichere- Reihe von schriftlichen Tests zu bewältigen hatten. Die entsprechenden demographischen Angaben zu dieser Gruppe sind in Tabelle 2 verzeichnet.

Tab. 2: Demographische Daten zur Haupttest-Stichprobe

Haupttest, gehobener Dienst, Saison '91 / '92; N=548						
	OST			WEST		
	Frauen	Männer	Gesamt	Frauen	Männer	Gesamt
EOS / Abitur	114	91	205	96	82	178
Fach(hoch)schule	7	6	13	47	58	105
Beruf mit Abitur	13	29	42			
Sonstige	1	4	5			
Gesamt	135	130	265	143	140	283
Altersdurchschnitt	19,6	22	20,8	20,8	22,9	21,8

Die im Haupttest eingesetzte Testbatterie ist in den neuen und in den alten Ländern unterschiedlich umfangreich. Nur ein Drittel der zusätzlichen 18 schriftlichen Aufgaben, die die Bewerber(innen) in den alten Ländern während des Haupttests erfüllen, werden auch in den neuen Ländern eingesetzt. Aus Kostengründen verzichten die Verwaltungen der neuen Länder auf die in den Altländern (mit Ausnahme von Hessen) übliche Diagnose der Merkfähigkeit (zwei Tests), der intellektuellen Leistungen mit figuralem Material (ein Test) und der Kenntnisse in verschiedenen Wissensdomänen (fünf Tests). Außerdem bleibt die Diagnose des Arbeitsverhaltens auf zwei (in den alten Bundesländern: vier) zusätzliche Indikatoren beschränkt, und es werden im Gegensatz zu den zwei zusätzlichen Tests in den alten Ländern hier keine weiteren Tests zur Erfassung der Grundkenntnisse aufgenommen. Die im folgenden dargestellte Analyse bleibt auf die Tests begrenzt, die in Ost und West eingesetzt werden. Die abschließende Leistungsbeurteilung geschieht auf der Grundlage aller Tests, einschließlich der neun Aufgaben aus dem Vortest. Sie basiert somit in den alten Ländern auf 27 und in den neuen Ländern auf 15 Tests. Während im Vortest die Leistungen aus den einzelnen Aufgaben einfach zu einem Gesamtwert aufaddiert werden -ein sogenanntes "Kompensationsmodell" bei dem ein niedriger Punktwert in einem Test durch einen hohen Punktwert in einem anderen Test ausgeglichen werden kann- werden beim abschließenden Leistungsurteil des Haupttests die aus objektiven Tests gewonnenen Daten durch die Profilinterpretation der Psycholog(inn)en zu einem sogenannten "Empfehlungsgrad" weiterverarbeitet. Im Gegensatz zum "Kompensationsmodell" des Vortests erfolgt diese Weiterverarbeitung hier nach einem "multiplen cutoff-Modell", demzufolge ein vorher definierter Minimalwert für bestimmte Bereiche erforderlich ist. Die abschließenden Empfehlungsgrade variieren von "1" bis "5" mit den Zwischenstufen "plus"

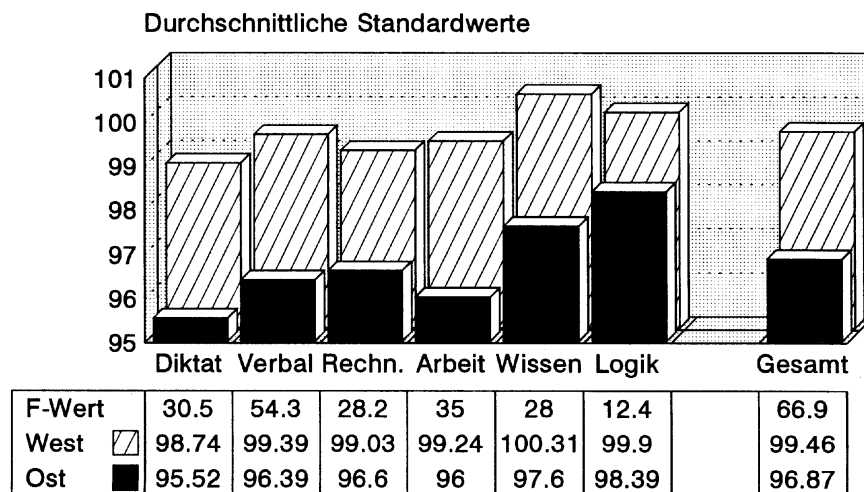
und "minus". Der Wert "1" steht dabei für den schlechtesten Empfehlungsgrad mit der Bedeutung "entspricht nicht den Anforderungen", der Wert "5" repräsentiert den besten Empfehlungsgrad und bedeutet "entspricht den Anforderungen". Neben den Testleistungen beurteilen die Diplom-Psycholog(inn)en der DGP im Haupttest die Ausprägungen verschiedener Verhaltensdimensionen. Diese Verhaltensdaten aus dem Rund- und Einzelgespräch des Haupttest finden in dem vorliegenden Bericht keine Berücksichtigung.

2.2 Analyse der Vortest-Ergebnisse

Im Balkendiagramm der Graphik Nr. 1 sind die durchschnittlichen Leistungen für die Oststichprobe (schwarz) und für die West-Stichprobe (weiß-gestreift) über die einzelnen Testdimensionen hinweg verzeichnet.

Abgetragen sind die Werte fürs Diktat, für zwei Tests zum sprachlichen Denkvermögen ("Verbal"), zur praktischen Rechenfertigkeit ("Rechn."), zur Konzentration und Sorgfalt des Arbeitsverhaltens ("Arbeit"), zu gemeinschaftskundlichen Kenntnissen ("Wissen") und zu zwei Denkaufgaben zur logisch-analytischen Befähigung ("Logik") sowie ein aggregiertes Maß ("Gesamt").

**Graphik Nr. 1: Ergebnisse Vortest gehobener Dienst
Signifikante Leistungsunterschiede Ost-West**



N: 1377 (696 Ost und 681 West)

Alle dargestellten Unterschiede sind statistisch hochsignifikant, d.h. sie sind mit großer Wahrscheinlichkeit nicht durch Zufallseffekte zu erklären. Die deutlichsten Unterschiede zeigen sich im gedanklichen Umgang mit verbalem Aufgabenmaterial, beim Diktat und im Arbeitsverhalten, die kleinste Differenz beim logischen Denkvermögen. Im Durchschnitt unterscheiden sich die beiden Gruppen um 2,6 Einheiten auf der von 70-130 variierenden Leistungsskala. Dies entspricht gerade einmal einer Viertel Standardabweichung, einer Maßeinheit für die Bedeutsamkeit von Variationen innerhalb einer Verteilung. Der Gruppenunterschied ist als gering einzustufen. Dies gilt insbesondere auch deshalb, weil er fast in der Größenordnung des sogenannten "Standardmeßfehlers" liegt, der als wichtiges Kennmaß für Tests die Diskrepanz zwischen aktuellen und wahren Werten beschreibt.

Qui numerare incipit, errare incipit - wer anfängt zu zählen, fängt an zu irren. Im vorliegenden Fall sollte dieser alte Merksatz daran erinnern, daß die Frage nach der statistischen Signifikanz, die im wesentlichen von der Stichprobengröße abhängt, gegenüber der Frage nach der **Bedeutung** der Ergebnisse zweitrangig ist. Loehlin (1992) hat gerade mit Blick auf die Untersuchung von Intelligenzunterschieden zwischen Gruppen gefordert, auf die Frage einzugehen, was die Daten bedeuten- und was sie **nicht bedeuten**. Anstelle der Signifikanz rückt dann die sogenannte Effektgröße in den Fokus dieser quantitativen Perspektive.

Den ermittelten Daten zufolge schneiden die Bewerber(innen) aus den neuen Ländern als Gruppe gegenüber der Gruppe der Bewerber(innen) aus den alten Bundesländern durchschnittlich etwas schlechter ab. Allerdings besagen diese Daten nicht, daß allein die Kulturzugehörigkeit (Ost oder West) die gefundenen Gruppenunterschiede erklärt. Tatsächlich werden lediglich 4,6 % der Varianz durch den Faktor dieser "Gruppenzugehörigkeit" aufgeklärt. Diese Zahl kann man annähernd wie folgt veranschaulichen: Geht man einmal für ein Beispiel davon aus, daß sich der Einfluß des Faktors Zugehörigkeit zur Ost- oder zur Weststichprobe nicht -wie in Wirklichkeit- in geringem Ausmaß auf alle Personen verteilt, sondern als "Alles-oder-Nichts" Variable wenige Personen im vollen Ausmaß betrifft, dann könnte man bei einer Gruppe von 100 Personen (50 West und 50 Ost) für ungefähr fünf Personen das geringfügig schlechtere Abschneiden im Test dadurch erklären, daß diese Personen aus den neuen Ländern stammen (und man muß dann natürlich erklären, wieso dies die Testleistungen beeinflussen kann, siehe unten, Punkt 3.) Bei den übrigen 45 Personen aus den neuen Ländern muß das geringfügig schlechtere Abschneiden durch andere Faktoren erklärt werden.

Mit den durchschnittlichen Leistungsunterschieden ist natürlich -zumindest theoretisch- in beiden Gruppen eine unterschiedliche Zulassungsquote zum Haupttest verbunden. Theoretisch heißt, falls die Verwaltungen sich bei der Zulassung an die Empfehlungen der *DGP* halten. Diesen Empfehlungen zufolge sollten Bewerber(innen) im Vortest idealerweise mehr als 90 Punkte, im Gesamttest (in Schleswig-Holstein nach dem dort geltenden anderen Punktsystem: 126) mindestens aber 82 Punkte (in Schleswig-Holstein: 111) erzielen, um mit ausreichenden Erfolgsaussichten zum Haupttest zugelassen zu werden. Geht man einmal von diesen Empfehlungen aus, so wären in der

untersuchten Stichprobe 58 % aller "Wessis", aber 71,1 % aller "Ossis" im Vortest "durchgefallen". Bewerber(innen) aus den neuen Ländern fallen demzufolge also 1,2 mal häufiger durch den Vortest als Bewerber(innen) aus den alten Ländern.

2.3 Analyse der Haupttest-Ergebnisse

Bevor die Haupttest-Ergebnisse analysiert werden können, müssen sie um einen Effekt "bereinigt" werden, der nichts mit den Unterschieden zwischen den Bewerber(innen)-Gruppen zu tun hat, sondern mit Ost-West Unterschieden bezüglich der Zulassungspraxis zum Haupttest. Im vorherigen Abschnitt wurden bereits die Empfehlungen angesprochen, die die DGP den Verwaltungen in Bezug auf die Zulassung zum Haupttest ausspricht. Die Verwaltungen in den alten Bundesländern halten sich zum großen Teil an diese Empfehlungen. 49,5 % aller Personen aus den Altbundesländern, die zum Haupttest zugelassen wurden, hatten mindestens 90 Punkte im Vortest erzielt (bzw. 125 Schleswig-Holsteinische Vortestpunkte²). In der Gruppe der zum Haupttest zugelassenen Bewerber(innen) für die neuen Länder hatten hingegen nur 39,6 % ein vergleichbar gutes Vortestergebnis erzielt. In den neuen Ländern werden also -wahrscheinlich als Reaktion auf die insgesamt schlechteren Vortestergebnisse- in größerem Umfang auch "schlechtere" Kandidat(inn)en noch zum Haupttest zugelassen. Der Vortestpunktwert steht aber in einem engen Zusammenhang zum Haupttestergebnis³, eine Zulassung von schwachen Kandidatinnen führt also mit hoher Wahrscheinlichkeit zu schwachen Haupttestergebnissen. Ohne eine Berücksichtigung dieser Thematik läuft man Gefahr, solche Effekte der Zulassungspraxis als Ost-West-Unterschiede zu mißinterpretieren. Interessiert man sich für tatsächliche Leistungsunterschiede, muß man diesen Effekt der unterschiedlichen Punktzahl im Vortest kontrollieren. Dies kann man entweder mittels bestimmter statistischer Prozeduren⁴ oder indem man einfach alle Personen mit einem Vortestpunktwert unter 90 (125) Vortestpunkten aus der Analyse ausschließt. Da dieses Vorgehen -nur Personen mit mindestens 90 Punkten im Vortest miteinander zu vergleichen- leichter nachzuvollziehen ist als die statistische "Manipulation", wird für den Bericht dieses Ausschlußverfahren gewählt. Der damit verbundene Nachteil einer reduzierten Stichprobengröße -nur 245 (105 Ost und 140 West) der 548 zum Haupttest zugelassene Personen konnten ein solch' "gutes" Vortestergebnis aufweisen- ist vertretbar, da die berichteten Ergebnisse sich bei dem zusätzlich durchgeführten statistischen Vorgehen auch für die Gesamtgruppe von 548 Personen nachweisen ließen.

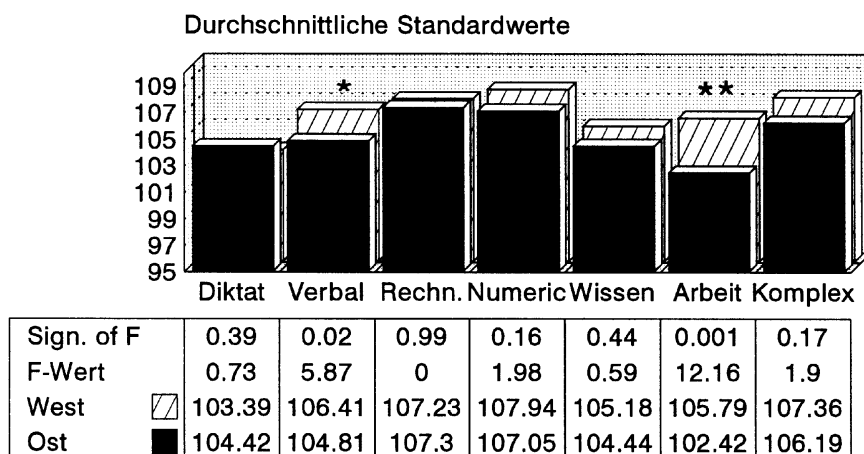
² Schleswig-Holstein hat eine andere Vortest-Praxis als die übrigen der hier analysierten Auftraggebern, so daß die Punktzahl hier kein direkt vergleichbarer Index für die Zulassungspraxis ist.

³Die schriftlichen Aufgaben des Vortests finden bei der Beurteilung des Haupttests eine erneute Berücksichtigung.

⁴ Hierbei handelt es sich um eine "Kovarianzanalyse" bei der die Effekte der Vortestpunktwerte auf die Haupttestergebnisse dadurch ausgeschlossen werden, daß die Mittelwerte der Haupttestergebnisse sozusagen an die Werte adjustiert werden, die sie annehmen würden, wenn alle Personen den gleichen Vortestpunktwert hätten.

Graphik 2 zeigt die durchschnittlichen Werte der beiden Gruppen im Diktat, im Bereich des intellektuellen Umgangs mit sprachlichem ("Verbal", vier Tests) und numerischem Material ("Numeric", 5 Tests), in den Grundrechenfertigkeiten ("Rechn.", ein Test) in der Gemeinschaftskunde ("Wissen", ein Test), im Arbeitsverhalten ("Arbeit", drei Tests) und in der Verarbeitungskomplexität ("Komplex", je ein Test aus den schon aufgeführten Bereichen "Verbal" und "Numeric").

Graphik Nr.2: Ergebnisse Haupttest gehobener Dienst
Reduzierte Stichprobe: Nur Personen mit min. 90 Vortestpunkten
Leistungsunterschiede



N: 245 (105 Ost und 140 West)

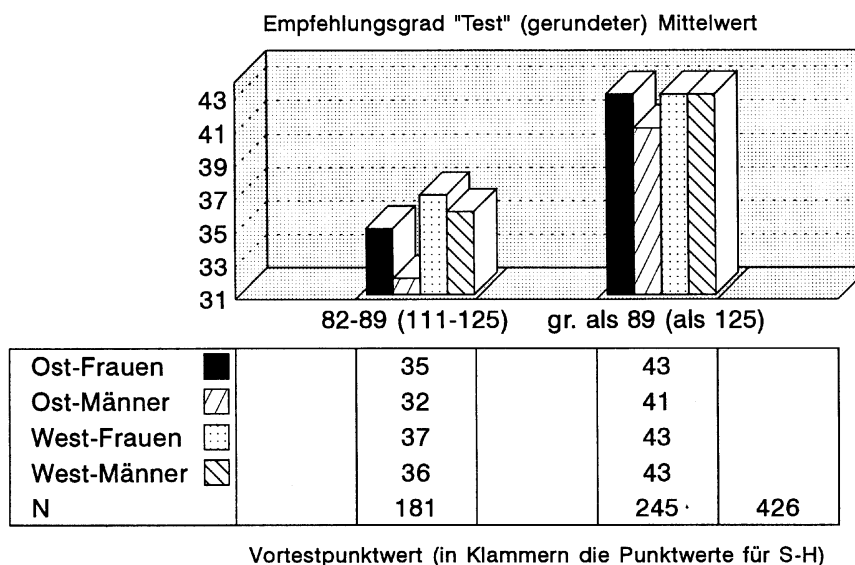
Signifikanz: * < 0,05; ** < 0,01

In dieser durch den Vortest selektierten Gruppe der Haupttest-Teilnehmer(innen) zeigt sich auf der Gesamtebene eines Aggregats über alle Tests nur noch ein minimaler (wenngleich nach wie vor statistisch signifikanter) Leistungsunterschied zwischen den Bewerber(innen) aus den alten und aus den neuen Ländern, der sich allerdings auf der Ebene der Testleistungsbeurteilung durch die Diplom-Psycholog(inn)en der *DGP* nicht mehr findet. Auf der Ebene der sieben einzelnen Leistungsbereiche übertreffen die Bewerber(innen) aus den alten Ländern ihre Vergleichspartner aus den neuen Ländern nur noch in zwei Domänen. In allen übrigen Bereichen gibt es keine statistisch bedeutsamen Leistungsunterschiede. Bei den beiden Domänen handelt es sich zum einen um die Effizienz und Sorgfalt im Arbeitsverhalten und zum anderen um den intellektuellen Umgang mit verbalem Material (die Signifikanz der Unterschiede ist in den Graphiken durch Sternchen angezeigt). Um genau zu sein, ist es vor allem *eine* Aufgabe aus dem Bereich "verbal", bei dem die

Bewerber(innen) aus den neuen Ländern um 2,8 Skaleneinheiten schlechter abschneiden, nämlich die "Textanalyse". Bei diesem Test kommt es auf eine sachgerechte Interpretation und Anwendung von komplexen sprachlichen Informationen an. Dabei liegt auch dieser Gruppenunterschied von der Größenordnung her im Bereich des Standardmeßfehlers.

Bei den Haupttestergebnissen muß außerdem noch darauf hingewiesen werden, daß bei einigen Variablen der nachgewiesene Effekt der Kulturzugehörigkeit (Ost oder West) von einer Interaktion zwischen dem Faktor Geschlecht (weiblich / männlich) und dem Faktor Kulturzugehörigkeit (Ost oder West) überlagert wird. Dies sieht man in der Graphik Nr. 3, die diesmal jene 426 Personen berücksichtigt, welche im Vortest mindestens 82 (111) Punkte erreicht haben. Abgetragen ist der von den Diplom-Psycholog(inn)en der *DGP* vergebene Empfehlungsgrad für die Leistungen im schriftlichen Test, und zwar einmal für die Personen mit mehr als 89 (125) Vortestpunkten und einmal für die Personen mit 82-89 (111-125) Vortestpunkten.

Graphik Nr. 3: Testergebnis in Abhängigkeit von den Vortestkategorien
Reduzierte Stichprobe: Nur Personen mit min. 82 Punkten im Vortest



Während die Ost-Frauen hier zumindest mit den West-Männern (fast) gleichauf liegen, fällt lediglich die Gruppe der Ost-Männer leistungsmäßig ab. Dieser Befund soll hier mitgeteilt werden, ohne auf das eigenständige Thema der im Bereich der intellektuellen Fähigkeiten oft gefundenen und dokumentierten geschlechtsspezifischen Leistungsunterschiede näher eingehen zu wollen. Bei den geschlechtsspezifischen Leistungsunterschieden geht es nicht um die absolute Überlegenheit von Frauen oder Männern, sondern um unterschiedliche Begabungsschwerpunkte, für die es verschieden akzentuierte Erklärungsansätze gibt. (Genannt seien hier nur die physiologischen (z.B. Ankney, 1992), pädagogischen oder psychologischen (z.B. motivationale Variablen, Hoefert & Schmidt, 1980) Erklärungsmuster). Die unterschiedlichen Begabungsschwerpunkte von Frauen und Männern sind -im Gegensatz zu ihren Ursachen- weitgehend bekannt. Der gemeinsame Trend zahlreicher unabhängiger Untersuchungen zeigt, daß Bewerberinnen offenbar im Arbeitsverhalten und bei Aufgaben mit verbalem Inhalt die besseren Leistungen zeigen, Bewerber dagegen eher bei numerischen oder räumlich-visuellen Aufgabentypen. Die Vertrautheit mit den geschlechtsspezifischen Leistungsunterschieden erlaubt es, die diesbezügliche Testfairness durch die **Zusammenstellung** der Aufgaben zu gewährleisten. In den Testbatterien der *DGP* sind die genannten Aufgabentypen im ausgeglichenen Verhältnis enthalten. Damit wird erreicht, daß auf der Ebene des Gesamtempfehlungsgrades die Chancengleichheit von Frauen und Männern bewahrt wird, ohne daß die Stärken und Schwächen der einzelnen Personen hinter der globalen Bewertung zurücktreten.

Durch die Analyse der hier vorgestellten Daten hat sich allerdings gezeigt, daß die Chancengleichheit zwischen den Geschlechtern zwar insgesamt gewahrt ist, daß im Vortest aber Aufgaben aus den Begabungsschwerpunktbereichen der Männer etwas überproportional vertreten waren. Dadurch wurde der Zugang zum Haupttest für die Frauen minimal schwerer als für die Männer. Im Haupttest selbst gab es dann mehr Aufgaben aus der "weiblichen Leistungsdomäne". Die *DGP* hat auf diesen Befund umgehend reagiert und wird ab der nächsten Saison im Vortest eine neue, geschlechterspezifisch ausgewogenere Testbatterie einsetzen.

Ähnlich wie bei den Vortestpunktwerten spiegelt der berichtete Interaktionseffekt im Haupttest zu einem Teil also eine für Frauen und Männer geringfügig unterschiedliche Zulassungshandhabung wieder, die nicht inhaltlich interpretiert werden sollte. Wenn die "Zulassungshürde" etwas mehr auf die Begabungsschwerpunkte der Männer zugeschnitten ist, folgt daraus, daß vergleichsweise schwächere Männer noch zum Haupttest zugelassen werden. Der "Vorsprung" der Männer wird dann bei den Aufgaben des Haupttest, die die Begabungsschwerpunkte der Frauen proportional mehr berücksichtigen, von den Frauen "ein-" oder gar "aufgeholt".

Der andere, erklärungsbedürftige Teil betrifft die Tatsache, daß der Faktor Geschlecht (weiblich, männlich) mit dem Faktor Kultur (Ost, West) **interagiert**. Die Ursachen dafür, daß der Leistungsvergleich zwischen Frauen und Männern **innerhalb** der Gruppen "Ost" und "West" in den neuen Ländern für die Frauen

(sowohl im Vortest als auch im Haupttest) geringfügig besser ausfällt als in den alten Ländern, sind vielleicht in den unterschiedlichen Sozialisationsbedingungen für die Geschlechter in der BRD und in der ehemaligen DDR zu sehen. Möglicherweise hat die moderne Politik der Frauenförderung in der ehemaligen DDR (*"Die Förderung der Frau, besonders in der beruflichen Qualifizierung, ist eine gesellschaftliche und staatliche Aufgabe"*, hieß es im Artikel 20 der DDR-Verfassung) Früchte getragen. Geschlechtsspezifische Leistungsunterschiede können über nicht-kognitive Variablen vermittelt sein (siehe z.B. Kong, 1989, p.79). Für den hier vorliegenden Befund bieten sich z.B. motivationale Erklärungsmuster an. So berichten Pollmer und Hurrelmann (1992), daß Mädchen in Sachsen Schulleistungen wichtiger nehmen als ihre Geschlechtsgefährtinnen in Nordrhein-Westfalen, während sich bei den Jungen keine Unterschiede zwischen den beiden Bundesländern zeigen. Über eine besondere Wirkung des Leistungsmotivs bei Mädchen in den Schulen der ehemaligen DDR berichten auch Witruk, Schnabel und Mussiol (1992).

2.4 Hinweise auf Ergebnisse anderer Forschungsgruppen

Die Analyse von Gemeinsamkeiten und Unterschieden der Menschen in den beiden -nun wiedervereinten- früheren "Teilen" Deutschlands gehört zu den aktuellen Schwerpunkten der deutschen psychologischen Forschung. Entsprechende Buchveröffentlichungen oder Sonderhefte von psychologischen Zeitschriften zu diesem Themenkomplex befinden sich zur Zeit aber noch in der Vorbereitungsphase, ebenso wie bei vielen Einzelarbeiten die Veröffentlichung noch aussteht. In Bezug auf deutsch-deutsche Vergleiche ist man deshalb im Moment zum Großteil auf Kongreßberichte, noch nicht veröffentlichte Manuskripte und auf mündliche Mitteilungen von Kolleg(inn)en angewiesen. Der nachstehend aufgeführte Bericht über empirische Ergebnisse im Kontext von Leistungsvergleichen kann daher nur eine grobe Vorinformation darstellen, die keinerlei Anspruch auf Vollständigkeit erhebt.

Wottawa (in Vorbereitung) und Stratemann (1992) berichten über den Einsatz des WILDE-Intelligenztests (Jäger & Althoff, 1984) bei Ost- und Westdeutschen Führungskräften. Signifikante Unterschiede -und zwar zu Ungunsten der Ost-Gruppe- ergaben sich nur im Bereich des "rechnerischen Denkens".

Beim Einsatz eines Tests des Berliner Intelligenzstrukturmodells (Jäger, 1982) bei Student(inn)en und Azubis aus den neuen und alten Bundesländern konnte Strohschneider (1992a und 1992b) auf Skalenebene für die West-Student(inn)en eine Überlegenheit im "Einfallsreichtum" und für die West-Azubis eine Überlegenheit in der "Verarbeitungskapazität" sowie im "figural-bildhaften Denken" ausmachen. Auf der Zellebene des Modells schnitten die "Wessis" in beiden Stichproben (Student(inn)en und Azubis) bezüglich der Verarbeitungskapazität verbalen Materials und bezüglich des Einfallsreichtums bei figural-bildhaften Material vergleichsweise besser ab.

Auch die freiwillige Annahmestelle der Bundeswehr berichtete in einer persönlichen Mitteilung über ein -im Ost-West Vergleich- besseres Abschneiden der Bewerber aus den alten Bundesländern in einigen kognitiven Leistungstests, welches sich allerdings mit zunehmender Testdauer verringert.

Demgegenüber fanden Baumert, Köller und Lehrke (1992), daß in ihrer Stichprobe von ca. 3000 Siebtkläßlern und 2500 Zehnklässlern die Schüler(inn)en aus den neuen Ländern bei Aufgaben zu Figurenanalogien ca. eine Viertel Standardabweichung besser abschnitten als die Schüler(innen) aus den alten Ländern.

Ein von Wissenschaftler(inn)en aus Bremen, Rostock und Berlin durchgeführter empirischer Vergleich der Rechtschreibleistungen von insgesamt knapp 5000 Erst- bis Viertkläßlern ergab ein differenziertes Bild: während die Fehlerzahl der West-Schüler(inn)en im Diktat höher lag als bei den Kindern aus dem Osten, stimmten die Fehlerquoten bei freien Texten fast überein (Becker, 1991).

Schmuck (1992) konnte in seinen Studien zur "strategischen Flexibilität" nachweisen, daß die ostdeutschen Student(inn)en und Lehrlinge sich häufig flexibler verhielten als die westdeutschen.

Bei einer Untersuchung der "Problemlösefähigkeit" erwies sich bei Strohschneider (1992a und 1992b), daß die Ostdeutschen einen größeren Zeitbedarf bei der initialen Problemexploration haben und daß die Westdeutschen bei der Dosierung ihrer Systemeingriffe eher zögerlicher oder vorsichtiger sind.

3. Mögliche Interpretationen der Befunde

Die Frage nach Gruppenunterschieden in der intellektuellen Leistungsfähigkeit steht nach wie vor im Spannungsfeld weltanschaulicher und wissenschaftlicher Auseinandersetzungen. Bei der Suche nach möglichen Interpretationen für die Leistungsunterschiede zwischen den Bewerber(innen) aus den alten und aus den neuen Ländern kann man sich auf die zahlreichen Arbeiten zur kulturvergleichenden Psychologie stützen. Kulturunterschiede und -gemeinschaften sind ein klassischer und extensiv bearbeiteter Gegenstand psychologischer Forschung. Berichte über Unterschiede zwischen Schwarzen und Weißen, mexikanischen Amerikaner(inne)n und Anglo-Amerikaner(inne)n in den USA, Ausführungen über die Besonderheiten von Eskimos in Kanada und Maori-Kindern in Neuseeland füllen Bibliotheken, eigenständige psychologische Fachzeitschriften (z.B. "*Journal of Intercultural Relation*") und Aktenordner mit Gerichtsurteilen. Dabei steht die gruppenvergleichende Psychologie nach wie vor im Brennpunkt einer zum Teil berechtigten und zum Teil erbitterten und emotional geführten Kritik.

Aufgabe der ost-westdeutschen Vergleichsforschung wird es sein, die für diese Fragestellung übertragbaren Komponenten der Kulturvergleichenden Psychologie herauszufinden, zu adaptieren und weiterzuentwickeln sowie die für die vorliegende Fragestellung unbrauchbaren Ansätze zurückzuweisen. Unbrauchbar für einen deutsch-deutschen Vergleich sind sicherlich alle Erklärungsmuster, die mit physiologischen oder genetischen Erklärungsansätzen operieren (siehe hierzu etwa Samuda, 1989b; Matarazzo, 1992). Ein grundsätzliches Defizit der kulturvergleichenden Forschung betrifft auch den innerdeutschen Vergleich: Die Voraussetzung einer Gruppeneinteilung ist an und für sich eine inhaltliche Definition der Merkmale der Gruppen "Ost" und "West". Es müßten also zunächst einmal die psychologischen, typographischen und demographischen Dimensionen benannt sein, die Mitglieder der "Ost"- bzw. der "West"gruppe auszeichnen sollen (z.B. eine höhere Aus-

prägung der Variablen "Extraversion" bei den "Wessis" oder ähnliches). Die Ausprägung der Dimensionen müßten dann je Gruppe gemessen werden, um schließlich aufzuzeigen, wie die entsprechende inhaltliche Dimension das interessierende Merkmal (z.B. kognitive Leistungsfähigkeit) beeinflusst (siehe hierzu etwa Helms, 1992). Hätte man eine solche inhaltliche Definition von "Ossis" und "Wessis", könnte es sein, daß in Berlin aufgewachsene "Ossis" den Gruppenmerkmalen der Ostdeutschen weniger entsprechen als manche Wessis usw.

Für die oben berichteten geringfügigen Leistungsunterschiede zwischen ostdeutschen und westdeutschen Bewerber(inne)n können verschiedene theoretische Erklärungsmuster angeboten werden. Ziel des vorliegenden ersten Teils des Berichts ist es, verschiedene Erklärungshypothesen zu entwickeln und zu diskutieren. Diese Erklärungshypothesen lassen sich auf verschiedenen Interpretationsebenen ansiedeln, die man auseinanderhalten sollte (auch wenn die einzelnen Aspekte der Ebenen natürlich miteinander verbunden sind). Die Ebenen betreffen (1) die Psychometrie, (2) die Stichprobengewinnung, (3) den diagnostischen Prozeß, und (4) die beteiligten Personen. Einige der im folgenden aufgeführten Hypothesen, werden zur Zeit überprüft. Von den Ergebnissen dieser Überprüfungen handelt der zweite Teil dieses Berichts im nächsten Heft.

3.1 Psychometrische Ebene

Auf der psychometrischen Ebene sind alle Argumente angesiedelt, die die gefundenen Unterschiede auf möglicherweise unterschiedliche psychometrische Eigenschaften der eingesetzten Tests zurückführen (siehe "*psycho*)metric equivalence", z.B. bei Butcher 1982, Poortinga, 1983). Psychometrische Kennwerte wie z.B. die Aufgabenschwierigkeit oder die Reliabilität sind -ebenso wie Testnormen- keine feststehenden Eigenschaften eines Tests, sondern abhängig von der jeweiligen Zielgruppe. Wenn z.B. die Ostdeutschen aus irgendwelchen Gründen (siehe unten) in der vorgegebenen Zeit insgesamt weniger Fragen des Tests überhaupt bearbeiten würden als die Westdeutschen, so würde sich dies nicht nur auf die Streuung der Werte, sondern unter Umständen auch auf die sogenannte Reliabilität, die Genauigkeit und Zuverlässigkeit der Messung auswirken und somit letztendlich auf die Aussagefähigkeit der Testergebnisse (siehe z.B. Samuda, 1989 a).

Aus psychometrischer Sicht ist z.B. der deutsch-deutsche Vergleich der Testempfehlungsgrades zu problematisieren. Diese Einstufung erfolgt nämlich -wie oben bereits ausgeführt- im Westen auf der Basis von 27 Tests, im Osten aber auf der Basis von nur 15 Tests.

In Bezug auf die Zuordnung von einzelnen Tests zu Anforderungskomplexen (z.B. "Intellektuelle Leistungen mit sprachlichem Material") konnte mit einer sogenannten Faktorenanalyse allerdings ein Nachweis der weitgehenden Vergleichbarkeit der Ost- und Westtestdaten erbracht werden.

3.2 Ebene der Stichprobengewinnung

Der Titel dieses Berichts sollte ursprünglich "(k)lein" Ost-Westvergleich lauten und zum Nachdenken über die Vergleichbarkeit der Stichproben anregen. Sofern man mit Stichproben arbeitet, muß man sich fragen, **welche** Grundgesamtheit durch die Stichprobe überhaupt repräsentiert wird. Will man die Ergebnisse des Leistungsvergleichs als allgemeine Ost-West-Unterschiede interpretieren, müßte gewährleistet sein, daß die beiden Gruppen nicht nur in Bezug auf das spezifische Merkmal einer Bewerbung für die Ausbildung zum gehobenen Dienst, sondern in Bezug auf alle Merkmale (*globale Repräsentativität*) Ost- und Westdeutscher vergleichbar sind. (Siehe "*Sampling equivalence*", z.B. bei Butcher 1982.)

Falls sich die beiden Gruppen neben dem Merkmal der "Kulturzugehörigkeit" (Ost versus West) noch in weiteren Merkmalen unterscheiden, so könnten natürlich auch diese Unterschiede in den weiteren Merkmalen für die durchschnittlichen Leistungsunterschiede zwischen den Gruppen verantwortlich sein. Man muß sich also fragen, ob sich in den alten und in den neuen Ländern nicht jeweils ein anderer Personenkreis für die Bewerbung zu einer Ausbildung in der Verwaltung entscheidet.

Tatsächlich sprechen vor allem zwei Gründe dafür, daß sich die Bewerber(innen) aus Ost- und Westdeutschland noch in anderen Merkmalen unterscheiden als nur in ihrer Zugehörigkeit zu den Bundesländern, nämlich (1) die Binnenwanderungsbewegungen in Deutschland und (2) die Selbstselektion.

Binnenwanderungsbewegungen in Deutschland

Politische, soziale und ökonomische Veränderungen ziehen Wanderungsbewegungen nach sich. Neben den historischen Flüchtlingsbewegungen vor dem Mauerbau, die in ihrer Ausrichtung nicht rein zufällig gewesen sein dürften, gilt es vor allem die aktuellen Abwanderungen in Rechnung zu stellen. Nach einer Studie der Bundesforschungsanstalt für Landeskunde und Raumordnung (Maretzke und Müller, 1992) sind 1991 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. Dabei unterscheidet sich die Altersstruktur der Wanderungsströme zwischen den alten und neuen Ländern deutlich. Die Abwanderung aus den neuen in die alten Länder konzentriert sich auf die jüngeren Personen unter 25 Jahren.

Während sich die Abwanderung (aus der Perspektive der neuen Länder) also auf die erwerbsfähigen und möglicherweise leistungsstarken Personen konzentriert, spekuliert Wottawa (in Vorbereitung) darüber, daß sich die Zuwanderung im wesentlichen auf zwei Personenkreise von Westdeutschen beschränkt, die nur bedingt geeignet sind, die Aufbauarbeit zu leisten. Er nennt hier zum einen leistungsfähige und aufstiegsorientierte Nachwuchskräfte, die die neuen Länder als Karrieresprungbrett benutzen, und zum anderen jene Mitarbeiter(innen), die von ihren Firmen oder vorgesetzten Behörden zu einem Wechsel bewegt werden, da es für sie im Westen keine Entwicklungsmöglichkeiten mehr gibt.

Nimmt man einmal an, daß der Ostgruppe in unserem Leistungsvergleich tatsächlich ein Teil der Spitzenleute auf Grund von Abwanderungsbewegungen fehlt, so würde dies allein eventuell schon ausreichen, um den Effekt der Leistungsunterschiede im Gruppenvergleich zu erklären.

Selbstselektion

Eine wesentliche Bedingung für die Personalauslese (und somit für die Ergebnisse in einem Leistungsvergleich von Verfahren zur Personalauswahl) ist die sogenannte "Basisrate", der Anteil der tatsächlich Qualifizierten an der Gesamtzahl der Bewerber(innen) (Jäger, 1970). Es gibt Hinweise dafür, daß diese Grundquote in den neuen Ländern ungünstiger ausfällt. Die empirisch gefundenen Leistungsunterschiede würden diesem Erklärungsansatz zufolge, diese unterschiedlichen Basisraten korrekt widerspiegeln.

Als Ursache dafür, daß der Anteil der geeigneten Bewerber(innen) an der Gesamtzahl der Bewerber(innen) in den neuen Ländern ungünstiger ausfallen könnte, nennt Stratemann (1992) die schlechteren Voraussetzungen für eine erfolgreiche "Selbstselektion". Bei der Selbstselektion geht es nach Weinert (1987) darum, daß Individuen Informationen und Eindrücke über Berufe, Arbeitsplätze und Organisationen sammeln, um sich die für sie "passenden" Verhältnisse, Bedingungen und Gegebenheiten ihrer zukünftigen Arbeit herauszusuchen. Dieser Prozeß der Selbstselektion ist dem organisationalen Selektionsprozeß vorgeschaltet. Zum Eignungstest erscheinen nur die Bewerber(innen), die den Verwaltungsberuf zumindest in Erwägung ziehen. Damit die Selbstselektion funktioniert, müssen aber mindestens zwei Voraussetzungen gegeben sein. Erstens müssen interpretierbare Informationen verfügbar sein, und zweitens muß sich das Individuum in einer gesicherten Position befinden, damit es überhaupt wählen kann. Beide Voraussetzungen sind in den neuen Ländern möglicherweise in geringerem Maße gegeben als in den alten Ländern. So bestehen in den neuen Ländern vielleicht weniger Vorstellungen und Kenntnisse darüber, was mit der in der Stellenanzeige ausgeschriebenen Ausbildung zur "Diplom-Verwaltungswirtin" oder zur "Inspektor-anwärterin" gemeint ist. Dies kann zur Folge haben, daß sich in größerem Ausmaß Personen bewerben, die überhaupt nicht zu diesem Berufsbild passen -und dementsprechend im Auswahlverfahren schlecht abschneiden. Das größere Ausmaß an Arbeitslosigkeit und/oder die möglicherweise größere Angst der Ostdeutschen, den Existenzkampf in der neuen "offenen Wettbewerbsgesellschaft" zu bestehen, kann in den neuen Ländern außerdem zu einem mehr oder minder wahllosen interessen- und begabungsunabhängigen Bewerbungsverhalten führen -mit den gleichen mißlichen Konsequenzen für die Basisrate in den neuen Ländern. Schließlich bleibt zu prüfen, ob die Verwaltungen für die Bundesbürger in den beiden zusammenwachsenden Hälften Deutschlands gleich attraktiv sind. Vielleicht verbindet man in den neuen Ländern mit dem Wort "Verwaltung" noch die negativen Erfahrungen mit dem Staatsapparat der DDR. Nach unattraktiven Ausbildungsplätzen halten aber auch nur "unattraktive" Bewerber(innen) Ausschau. Würde diese These zutreffen, müßten die Verwaltungen in den neuen Ländern mehr Imagepflege betreiben.

Die hier genannten, auf der Ebene der Stichprobengewinnung angesiedelten, Hypothesen erklären einen Teil der Inkonsistenz der in Punkt 2.4. berichteten Ergebnisse anderer Forschungsgruppen. Es sind vor allem Untersuchungen mit **selektierten** Stichproben (z.B. Bundeswehrranwärter), bei denen sich - möglicherweise in Folge der mißglückten Selektion- eine Überlegenheit der Gruppe aus den alten Bundesländern zeigt, während bei Untersuchungen mit unselektierten Stichproben (z.B. mit Schüler(innen)) solche Effekte nicht zu beobachten sind.

Die Verwaltungen in den neuen Ländern sind hier aufgerufen, den Prozeß einer erfolgreichen Selbstselektion durch eine verständliche und ausführliche Anzeigenformulierung zu unterstützen und den Bewerber(innen) Gelegenheit zu weiteren Informationen (z.B. am Telefon oder an einem Informationsabend) einzuräumen.

Weiterhin empfiehlt es sich, bei der Auswahl der Bewerber(innen) auf die Ausgangslage einer möglicherweise ungünstigeren Basisrate mit einer **stufenweisen** Selektion zu reagieren. Das für die Verwaltungen und Bewerber(innen) aufwendige Verfahren des Haupttests (einschließlich der Rund- und Einzelgespräche) sollte erst **nach** einer Verbesserung der Grundquote durch den vorgeschalteten Vortest erfolgen. Dabei muß man sich dann natürlich an die berechtigten Empfehlungen der *DGP* bezüglich der Zulassung zum Haupttest halten.

3.3 Ebene des diagnostischen Prozesses

Der testdiagnostische Prozeß, die Erhebung von Informationen mit dem Ziel, Erklärungen oder Entscheidungshilfen zu gewinnen, gliedert sich nach Michel und Conrad (1982) in vier Phasen: die Provokation, die Registrierung, die Auswertung und die Interpretation. Grundsätzlich können sich auf allen Ebenen Minoritätenprobleme manifestieren, die weitgehend "mechanisch" ablaufende Registrierung und Auswertung der testdiagnostischen Informationen bei der *DGP* lassen eine kulturabhängige Verzerrung bei diesen beiden Phasen allerdings unwahrscheinlich werden. (Anders verhält es sich natürlich bei der hier nicht thematisierten Erfassung von Verhaltensdaten). Die folgenden Hypothesen zur Erklärung der kulturabhängigen Leistungsunterschiede, bleiben daher auf die Phasen der Provokation und der Interpretation der Ergebnisse beschränkt.

Provokation

Die Provokation von Informationen, die Konfrontierung der Bewerber(innen) mit den Testaufgaben, geht von standardisierten Untersuchungsbedingungen aus. Während eine solche Standardisierung für die äußeren Untersuchungsbedingungen leicht gewährleistet werden kann, lassen sich die **innerpsychischen** Bedingungen der Teilnehmer(innen) natürlich nicht standardisieren. Hierzu gehören zunächst die Effekte der persönlichen Einstellungen und Erwartungen der Person sowie die Effekte nicht-kognitiver Merkmale, die im nächsten Punkt erörtert werden. So wird sich z.B. eine unterschiedliche Testmotivation natürlich auch auf die Ergebnisse nieder-

schlagen (siehe unten).

Hierzu gehören aber auch die Effekte der allgemeinen Vertrautheit mit Untersuchungssituationen, mit schriftlichen Leistungstests und ihren Besonderheiten (z.B. "Zeitdruck") und die unterschiedliche Vertrautheit mit Antwortprozeduren ("multiple-choice" Antwortformat, "Ratewahrscheinlichkeiten" usw.).

Unter der Gültigkeit von zwei Prämissen können die gefundenen deutsch-deutschen Leistungsunterschiede auf eine Verzerrung bei der Provokation der Informationen zurückgeführt werden: Man muß erstens annehmen, daß die Bewerber(innen) in den neuen Länder weniger Testerfahrung mitbringen als ihre Vergleichsgruppe in den Altbundesländern, und zweitens, daß sich die Testerfahrung positiv auf das Leistungsergebnis auswirkt. Obwohl auch in der ehemaligen DDR psychologische Leistungstests (und zwar zum Teil (modifizierte) "West"-Tests) eingesetzt wurden (siehe z.B. Wolfram, Neumann & Wieczorek, 1989), erscheint die erste Annahme einer vergleichsweise geringeren Verbreitung von Tests plausibel. Mit der zweiten Annahme der positiven Auswirkung von Testerfahrung auf die Leistung begibt man sich in die Diskussion um die Trainierbarkeit und Übbarkeit von Testleistungen. Vermeidet man dabei einmal die grundsätzliche Diskussion über die Zeit- und Bedingungsvarianz von Eignungs- und Bewährungsmerkmalen und deren Beziehungen zueinander (siehe auch Althoff, 1973) sowie die Kontroverse zwischen Eigenschaftstheoretikern und Situationisten und behandelt das Problem eher pragmatisch, so muß man vor allem versuchen, den **Effekt** der Testerfahrung und den Effekt des Trainings abzuschätzen. Als Resultat einer Meta-Analyse über 38 Studien zu diesem Thema berichten Kulik, Bangert-Drowns und Kulik (1984) eine durchschnittliche Effektgröße des Trainings von 0.33. Dies würde für die *DGP* Skalen mit einem Mittelwert von 100 und einer Standardabweichung von 10 bedeuten, daß sich eine durchschnittliche Leistung von 100 durch ein Testtraining auf einen Mittelwert von 103 verbessern ließe. Der erwartete Effekt des Testtrainings liegt also exakt in der Größenordnung der gefundenen Gruppenunterschiede. Dabei hängt die Effektgröße natürlich von der Trainingsdauer und der Zugeschrittenheit des Trainings auf einen speziellen Test ab.

Ein anderer, in der Literatur immer wieder berichteter Befund, besagt, daß sich bei einem gemeinsamen Training aller Personen die Rangreihe der Personen bei einer zweiten Testung im Anschluß an das Training so gut wie nicht verändert (siehe z.B. Fay, 1992) und daß es einen Sättigungseffekt in der Wirksamkeit des Trainings gibt.

Als Fazit wird daher immer wieder gefordert, die Lernmöglichkeiten der Testteilnehmer(innen) bei der Testung zu berücksichtigen (z.B. Brambring, 1982, Samuda, 1989 c). Die *DGP* hat dieser Forderung durch die seit August 1990 (und im August 1991 verbesserte) Bewerber(innen)broschüre Rechnung getragen. Diese Informationsbroschüre wird allen Bewerber(innen) mit der Einladung zum Test zugesandt. Zurück zur Ausgangshypothese: Weiter oben wurde formuliert, daß sich die Leistungsunterschiede möglicherweise auf die unterschiedliche Testerfahrung der beiden Gruppen zurückführen lassen. Träfe dies zu, so müßte man erwarten, daß die Bewerber(innen) in den neuen Ländern von der Trainingsmöglichkeit durch die Broschüre in größerem Aus-

maß profitieren, als die Bewerber(innen) aus den alten Ländern, die in Bezug auf die Testerfahrung schon "gesättigter" sind. Tatsächlich zeigt ein Vergleich der hier berichteten Daten mit Daten aus einem Ost-West-Vergleich aus der Saison 1990/1991, daß sich die Leistungen der Bewerber(innen) für die neuen Länder -möglicherweise in Folge der Broschüre- um durchschnittlich 1,94 Skaleneinheiten verbessert haben, während die Bewerber(innen) der Altbundesländer nur 0,96 Einheiten zulegen konnten.

Der hier berichtete Erklärungsansatz rechtfertigt die Hoffnung, daß sich die gefundenen Unterschiede mit zunehmender Testpraxis in den neuen Ländern und mit zunehmender Nutzung der *DGP*-Vorbereitungsbroschüre vollständig verlieren.

Im Rahmen der Provokation von Informationen muß auch die möglicherweise in Ost und West unterschiedlich ausgeprägte Vertrautheit mit den im Test geforderten **Antwortprozeduren** thematisiert werden. Bei der Bearbeitung von multiple-choice Aufgaben unter Zeitdruck empfiehlt es sich, zunächst alle Fragen recht zügig zu bearbeiten und erst später -sofern noch Zeit vorhanden ist- alle Lösungen noch einmal gründlich zu studieren. Auf keinen Fall sollte man sich bei einzelnen Fragen, die einem besonders schwerfallen, "festbeißen". Stattdessen sollte man die Lösung dieser Frage zunächst offen lassen und zur nächsten Frage übergehen. Obwohl diese Hinweise im Rahmen der Testeinführung bei den Eignungsuntersuchungen der *DGP* den Bewerber(innen) mitgeteilt werden, läßt sich natürlich nicht ausschließen, daß die Bearbeitungsstrategien -etwa als Konsequenz der geringeren Testerfahrung- in den neuen Ländern weniger effizient und angemessen sind als in den alten Ländern. Dies könnte bedeuten, daß das durchschnittliche Leistungsdefizit der ostdeutschen Bewerber(innen) allein dadurch zu erklären ist, daß diese Bewerber(innen) insgesamt in der vorgegebenen Zeit weniger Fragen bearbeiten -und dementsprechend auch bei identischer Fehlerquote insgesamt weniger Pluspunkte sammeln können. Diese These läßt sich nur durch aufwendige Analysetechniken, sogenannte Itemanalysen, verifizieren oder falsifizieren. Diese Analysen werden zur Zeit für einige ausgesuchte Aufgaben durchgeführt, über die Ergebnisse wird im nächsten Heft berichtet.

Im Zusammenhang mit der möglicherweise unterschiedlichen Testerfahrung der Bewerber(innen) in Ost und West und den sich daraus ergebenden Konsequenzen, müssen die Verwaltungen in den neuen Ländern aber auch die Praxis der sogenannten "verkürzten" Testbatterien überdenken. Gerade wenn die Gewöhnung an die Testsituation eine Rolle spielt, sollte man den Bewerber(innen) durch eine inhaltlich und zeitlich ausgedehntere Testung die Chance einräumen, ihr tatsächliches Leistungsniveau zu zeigen. Solange aus den hier und den weiter unten angeführten Gründen berechnete oder emotional begründete Zweifel an der Gültigkeit von testbasierten Aussagen in den neuen Bundesländern bestehen, ist eine -gemessen an der bewährten Testanzahl in den Altbundesländern (mit Ausnahme von Hessen)- **Ver-ringerung** der Testquantität kontraindiziert. Dieser Punkt läßt sich durch die vorliegenden Daten noch untermauern. Nimmt man einmal an, daß sich die mangelnde Testerfahrung in Form von Aufregung usw. negativ auf die Testleistungen auswirkt, so müßte dieser Effekt zu Beginn der Testung am ausge-

an Aktualität wieder gewinnen sollte. Für Intelligenztests im engeren Sinne muß der Nachweis der Kontentvalidität ausbleiben, da es keine Verständigung über die Grundgesamtheit intelligenter Fähigkeiten gibt (Jäger, 1986, S. 275). Der Aspekt der möglicherweise beim Einsatz in den neuen Ländern eingeschränkten Kontentvalidität der *DGP*-Tests bezieht sich daher insbesondere auf das Diktat und den Wissenstest.

Die Kontentvalidität dieser beiden Kenntnistests wäre dann eingeschränkt (*"item-verzerrt"*, siehe z.B. Poortinga, 1983), wenn die Tests überwiegend westdeutsche Kenntnisse repräsentieren und ostdeutsche Kenntnisse unterrepräsentieren. Dieser topic betrifft, allgemeiner gefaßt, die möglicherweise grundsätzlich unterschiedliche Wirkung der Itemstimuli in Ost- und Westdeutschland (ist in dieser weiteren Auffassung aber nicht mehr der Kontentvaliditäts-Problematik zuzuordnen). Eine solche Perspektive würde in einer extremen Auslegung davon ausgehen, daß sich in der ehemaligen DDR und der BRD zwei unterschiedliche deutsche Sprachen herausgebildet haben und daß der Einsatz der *DGP*-Test für die in der DDR sozialisierten Personen eine Sprachbarriere beinhaltet (siehe *"linguistic equivalence"* z.B. bei Helms, 1992).

Kommen wir zurück zur Kontentvalidität des Diktats und der 'Gemeinschaftskundlichen Wissensfragen', so hat die *DGP* hier umgehend auf die ange-deutete Gefahr reagiert. Der Test zur Gegenwartskunde wurde in Zusammenarbeit mit Lehrer(inne)n und Kolleg(inne)n aus den neuen Ländern neu konstruiert und soll nun einen gemeinsamen Wissensfundus umfassen. Auch bei der zur Zeit anstehenden Neukonstruktion des Wissenstests zum Kenntnisbereich 'Geschichte', wird die Problematik der in den beiden ehemaligen "Hälften" Deutschlands unterschiedlichen Schwerpunkte in den Geschichtskenntnissen von Anfang an berücksichtigt. Die in Ost und West möglicherweise unterschiedliche Ausprägung der Kenntnisse zu spezifischen Geschichtsfragen werden zu diesem Zweck vorab erhoben und bei der Neukonstruktion des Gesamttests explizit berücksichtigt. Damit wird der Repräsentationsfrage Rechnung getragen, ohne auf die Verwendung kulturspezifischer Inhalte zu verzichten. Eine Verwendung von tendenziell kulturfreierem Testmaterial empfiehlt sich nämlich nicht unbedingt, da sich dadurch der Aufgabencharakter grundsätzlich verändern kann (Jensen & McGurk, 1987). Bei den Rechtschreibkenntnissen soll zunächst überprüft werden, ob der hier angedachte Effekt der Kulturspezifität des Diktattextes tatsächlich existiert. Zu diesem Zweck hat die *DGP* eine neue Diktataufgabe auf der Grundlage von Schulbüchern der ehemaligen DDR konstruiert und dieses Diktat in den neuen und in den alten Bundesländern zusätzlich eingesetzt. Würden die ostdeutschen Bewerber(innen) wirklich durch die Verwendung von Westtexten benachteiligt, müßten sie beim "heimischen" Text des "Ostdiktates" vergleichsweise besser abschneiden. Über die Ergebnisse dieser Überprüfung wird im nächsten Heft berichtet.

Kriteriumsvalidität

Als Kriteriumsvalidität wird der Grad der Genauigkeit bezeichnet, mit dem von den Ergebnissen eines Tests auf ein Kriteriumsverhalten (z.B. das Ausbildungsverhalten) geschlossen werden kann. Tests zur Personalauswahl müssen in einem nachweisbaren Zusammenhang mit der Ausbildung oder dem Berufserfolg stehen. Dieser Zusammenhang begründet die Vorhersagekraft der Testergebnisse. Für die **Selektions-Fairness** kommt es darauf an, für verschiedene Gruppen eine gleiche Vorhersagbarkeit des Kriteriums zu erzielen⁵ (siehe "*selection test fairness*" z.B. bei Hunter, Schmidt und Rauschenberger, 1977). Für den Einsatz von "West"-Tests in den neuen Ländern muß daher zunächst der explizite Nachweis der Kriteriumsvalidität auch für diese neue Testpopulation gefordert werden. Erste Analysen über den Zusammenhang zwischen den *DGP*-Empfehlungsgraden und dem Abschneiden in der Ausbildung auch in den neuen Ländern lassen sich in dem Artikel "*Zur Treffsicherheit der DGP-Prognosen in den neuen Ländern*" im vorliegenden Heft der *DGP*-Informationen nachlesen. Bewährungsproben dieser Art können den Nachweis erbringen, daß die *DGP* Tests auch in den neuen Ländern eine treffsichere Vorhersage des Ausbildungserfolgs erlauben. Kulturverzerrte Kriteriumsvaliditäten lassen sich aber nur dort nachweisen oder mit Sicherheit ausschließen, wo die Azubis aus den neuen Ländern und den alten Ländern gemeinsam unterrichtet und bewertet werden, da sich sonst eine Anpassung der Bewertung an das jeweilige Gruppenniveau der Azubis nicht ausschließen läßt.

Die Behauptung, daß ein Test eine singuläre (nur für eine Gruppe) oder differentielle (für verschiedene Gruppe unterschiedlich hohe) Kriteriumsvalidität aufweist, ist allerdings sehr unplausibel und konnte selbst bei deutlich disperateren Kulturgruppen bislang nicht bestätigt werden (Walsh & Betz, 1985; Schmidt, Berner & Hunter, 1973; Hunter, Schmidt & Hunter, 1979; Hunter & Hunter, 1984).

Konstruktvalidität

Bei der Konstruktvalidität geht es darum, inwieweit man von den Ergebnissen eines Tests auf den Ausprägungsgrad eines theoretischen (nicht gefundenen, sondern *erfundenen* (Jäger 1986)) Konstrukts schließen kann. Bei den *DGP*-Tests repräsentieren beispielsweise bestimmte Vorstellungen über ausbildungsrelevante Facetten der Intelligenz dieses Konstrukt.

Findet man nun in zwei kulturell unterschiedlichen Gruppen unterschiedliche Testergebnisse, so kann man dies auch so interpretieren, daß sich die Konstrukte in den beiden Gruppen unterscheiden. "Intelligenz" ist ein kulturell definiertes Konstrukt (Samuda, 1989a), das kulturelle und soziale Umfeld beeinflusst in Form der Aufgaben- und Zielgestaltung und in Form von Feedbacks, Kontrollen und Evaluationen die Entwicklung und das Zusammenwirken kognitiver Fähigkeiten (Simmons, 1985).

⁵Gefordert sind identische Regressionslinien für unterschiedliche Populationen. Für diese Definition der Testfairness von Cleary und Hilton und für andere Definitionen der Testfairness siehe Hunter, Schmidt und Rauschenberger, 1977.

an Aktualität wieder gewinnen sollte. Für Intelligenztests im engeren Sinne muß der Nachweis der Kontentvalidität ausbleiben, da es keine Verständigung über die Grundgesamtheit intelligenter Fähigkeiten gibt (Jäger, 1986, S. 275). Der Aspekt der möglicherweise beim Einsatz in den neuen Ländern eingeschränkten Kontentvalidität der *DGP* -Tests bezieht sich daher insbesondere auf das Diktat und den Wissenstest.

Die Kontentvalidität dieser beiden Kenntnistests wäre dann eingeschränkt (*"item-verzerrt"*, siehe z.B. Poortinga, 1983), wenn die Tests überwiegend westdeutsche Kenntnisse repräsentieren und ostdeutsche Kenntnisse unterrepräsentieren. Dieser topic betrifft, allgemeiner gefaßt, die möglicherweise grundsätzlich unterschiedliche Wirkung der Itemstimuli in Ost- und Westdeutschland (ist in dieser weiteren Auffassung aber nicht mehr der Kontentvaliditäts-Problematik zuzuordnen). Eine solche Perspektive würde in einer extremen Auslegung davon ausgehen, daß sich in der ehemaligen DDR und der BRD zwei unterschiedliche deutsche Sprachen herausgebildet haben und daß der Einsatz der *DGP*- Test für die in der DDR sozialisierten Personen eine Sprachbarriere beinhaltet (siehe *"linguistic equivalence"* z.B. bei Helms, 1992).

Kommen wir zurück zur Kontentvalidität des Diktats und der 'Gemeinschaftskundlichen Wissensfragen', so hat die *DGP* hier umgehend auf die ange deutete Gefahr reagiert. Der Test zur Gegenwartskunde wurde in Zusammenarbeit mit Lehrer(inne)n und Kolleg(inne)n aus den neuen Ländern neu konstruiert und soll nun einen gemeinsamen Wissensfundus umfassen. Auch bei der zur Zeit anstehenden Neukonstruktion des Wissenstests zum Kenntnisbereich 'Geschichte', wird die Problematik der in den beiden ehemaligen "Hälften" Deutschlands unterschiedlichen Schwerpunkte in den Geschichtskenntnissen von Anfang an berücksichtigt. Die in Ost und West möglicherweise unterschiedliche Ausprägung der Kenntnisse zu spezifischen Geschichtsfragen werden zu diesem Zweck vorab erhoben und bei der Neukonstruktion des Gesamttests explizit berücksichtigt. Damit wird der Repräsentationsfrage Rechnung getragen, ohne auf die Verwendung kulturspezifischer Inhalte zu verzichten. Eine Verwendung von tendenziell kulturfreierem Testmaterial empfiehlt sich nämlich nicht unbedingt, da sich dadurch der Aufgabencharakter grundsätzlich verändern kann (Jensen & McGurk, 1987). Bei den Rechtschreibkenntnissen soll zunächst überprüft werden, ob der hier angedachte Effekt der Kulturspezifität des Diktattextes tatsächlich existiert. Zu diesem Zweck hat die *DGP* eine neue Diktataufgabe auf der Grundlage von Schulbüchern der ehemaligen DDR konstruiert und dieses Diktat in den neuen und in den alten Bundesländern zusätzlich eingesetzt. Würden die ostdeutschen Bewerber(innen) wirklich durch die Verwendung von Westtexten benachteiligt, müßten sie beim "heimischen" Text des "Ostdiktates" vergleichsweise besser abschneiden. Über die Ergebnisse dieser Überprüfung wird im nächsten Heft berichtet.

Kriteriumsvalidität

Als Kriteriumsvalidität wird der Grad der Genauigkeit bezeichnet, mit dem von den Ergebnissen eines Tests auf ein Kriteriumsverhalten (z.B. das Ausbildungsverhalten) geschlossen werden kann. Tests zur Personalauswahl müssen in einem nachweisbaren Zusammenhang mit der Ausbildung oder dem Berufserfolg stehen. Dieser Zusammenhang begründet die Vorhersagekraft der Testergebnisse. Für die **Selektions-Fairness** kommt es darauf an, für verschiedene Gruppen eine gleiche Vorhersagbarkeit des Kriteriums zu erzielen⁵ (siehe "*selection test fairness*" z.B. bei Hunter, Schmidt und Rauschenberger, 1977). Für den Einsatz von "West"-Tests in den neuen Ländern muß daher zunächst der explizite Nachweis der Kriteriumsvalidität auch für diese neue Testpopulation gefordert werden. Erste Analysen über den Zusammenhang zwischen den *DGP*-Empfehlungsgraden und dem Abschneiden in der Ausbildung auch in den neuen Ländern lassen sich in dem Artikel "*Zur Treffsicherheit der DGP-Prognosen in den neuen Ländern*" im vorliegenden Heft der *DGP*-Informationen nachlesen. Bewährungsproben dieser Art können den Nachweis erbringen, daß die *DGP* Tests auch in den neuen Ländern eine treffsichere Vorhersage des Ausbildungserfolgs erlauben. Kulturverzerrte Kriteriumsvaliditäten lassen sich aber nur dort nachweisen oder mit Sicherheit ausschließen, wo die Azubis aus den neuen Ländern und den alten Ländern gemeinsam unterrichtet und bewertet werden, da sich sonst eine Anpassung der Bewertung an das jeweilige Gruppenniveau der Azubis nicht ausschließen läßt.

Die Behauptung, daß ein Test eine singuläre (nur für eine Gruppe) oder differentielle (für verschiedene Gruppe unterschiedlich hohe) Kriteriumsvalidität aufweist, ist allerdings sehr unplausibel und konnte selbst bei deutlich disperateren Kulturgruppen bislang nicht bestätigt werden (Walsh & Betz, 1985; Schmidt, Berner & Hunter, 1973; Hunter, Schmidt & Hunter, 1979; Hunter & Hunter, 1984).

Konstruktvalidität

Bei der Konstruktvalidität geht es darum, inwieweit man von den Ergebnissen eines Tests auf den Ausprägungsgrad eines theoretischen (nicht gefundenen, sondern *erfundenen* (Jäger 1986)) Konstrukts schließen kann. Bei den *DGP*-Tests repräsentieren beispielsweise bestimmte Vorstellungen über ausbildungsrelevante Facetten der Intelligenz dieses Konstrukt.

Findet man nun in zwei kulturell unterschiedlichen Gruppen unterschiedliche Testergebnisse, so kann man dies auch so interpretieren, daß sich die Konstrukte in den beiden Gruppen unterscheiden. "Intelligenz" ist ein kulturell definiertes Konstrukt (Samuda, 1989a), das kulturelle und soziale Umfeld beeinflusst in Form der Aufgaben- und Zielgestaltung und in Form von Feedbacks, Kontrollen und Evaluationen die Entwicklung und das Zusammenwirken kognitiver Fähigkeiten (Simmons, 1985).

⁵Gefordert sind identische Regressionslinien für unterschiedliche Populationen. Für diese Definition der Testfairness von Cleary und Hilton und für andere Definitionen der Testfairness siehe Hunter, Schmidt und Rauschenberger, 1977.

Auf der Ebene der Konstruktvalidität könnte man nun argumentieren, daß die deutsch-deutschen Leistungsunterschiede unterschiedliche "Intelligenz"-ausprägungen widerspiegeln und daß man mit "West"-Tests lediglich die Ausprägung einer westdeutschen Intelligenz messen würden, so daß Bewerber(innen) aus den neuen Ländern zwangsläufig schlechter abschneiden müßten. Man müßte also behaupten, daß -z.B. in Abhängigkeit unterschiedlicher Wertvorstellungen- ein Verhalten, das in der ehemaligen BRD als intelligentes Verhalten galt, in der ehemaligen DDR ein -im Extremfall- ein dummes Verhalten darstellte und vice versa (siehe den Argumentationsgang zur Erklärung von Rassenunterschieden bei Helms, 1992).

Eine solche Auffassung widerspricht allen Befunden zur Universalität der Annahmen zur Intelligenzstruktur (siehe z.B. Kleine & Jäger, 1987; Kleine & Jäger, 1989; Rosas, 1990). Auch die vorliegenden Daten lassen eine solche Interpretation kaum zu. So konnte zum einen eine weitgehende Invarianz der Faktorenstruktur in beiden Gruppen nachgewiesen werden, zum anderen sind die gefundenen Leistungsunterschiede gerade im Bereich der "Verarbeitungskapazität" am **geringsten**. Die zur "Verarbeitungskapazität" zählenden Aufgaben zum induktiven und deduktiven Denken sind in Intelligenzmodellen aber von zentraler Bedeutung. Die den Leistungen bei diesen Aufgaben zugrundeliegenden Fähigkeiten können als Voraussetzung für den Erwerb und die Anwendung von Wissen angesehen werden (Süß, Kersting und Oberauer, 1991). Als Befürworter(in) der These unterschiedlicher Konstruktvaliditäten des *DGP* Tests in Ost und West hätte man daher gerade in diesem Bereich die deutlichsten Unterschiede erwarten müssen.

3.4 Ebene der Person

Für den diagnostischen Laien mag es am naheliegendsten sein, Unterschiede in den Leistungen auf Unterschiede in den Personen zurückzuführen. Hypothesen, die auf dieser Interpretationsebene anzusiedeln sind, bilden den letzte Abschnitt der sukzessiven Auslotung des Erklärungsraumes. Dabei unterscheiden Vijver und Poortinga (1992) kulturelle Verzerrungen aufgrund von Merkmalen der getesteten Person, der testenden Person und aufgrund von Interaktionen zwischen beiden.

Eine verzerrende Tendenz auf Seiten der testdurchführenden Personen könnte sich beispielsweise beim abschließenden Leistungsurteil über das Abschneiden im Haupttests zeigen, wenn die aus objektiv Tests gewonnenen Daten durch die Profilinterpretation der Psycholog(inn)en zu einem sogenannten "Empfehlungsgrad" weiterverarbeitet werden. Ein solcher Effekt ließ sich bei einem Vergleich dieser von Menschen getroffenen Urteile mit einem nachträglich berechnetem statistischen Urteil aber nicht aufweisen. Die *DGP*-Psycholog(innen) urteilen stattdessen in den neuen und in den alten Ländern annähernd gleich "streng" oder "milde".

Bei möglichen Unterschieden innerhalb der getesteten Personen muß man neben den direkten kognitiven Variablen auch die nicht-kognitiven Variablen wie z.B. Leistungsmotivation, Selbstwirksamkeit und Testängstlichkeit

berücksichtigen. Unterschiedlich ausgeprägte Leistungen können durch unterschiedliche Ausprägungen in solchen nicht-kognitiven Variablen bedingt sein. Im Zusammenhang hiermit berichtet beispielsweise Wottawa (in Vorbereitung), daß sich ost- und westdeutsche Führungskräfte bezüglich der Selbstdarstellung ihrer intellektuellen Effizienz (niedrigere Werte in den neuen Bundesländern) unterscheiden. Pollmer und Hurrelmann (1992) zufolge ist das schulische Belastungserleben bei den Schüler(inne)n in Sachsen größer als in Nordrhein-Westfalen. Sie weisen allerdings darauf hin, daß sich dieses Belastungserleben nicht in Form verstärkter psychophysischer Beschwerden und/oder in Form von überhöhten delinquenten Verhaltensformen äußert. Es wäre zu prüfen, ob und wie sich dieses Belastungserleben auf die Leistungsfähigkeit auswirkt. Oettingen und Baltes (1992) fanden bei Westberliner Schüler(inne)n vergleichsweise höhere Selbstwirksamkeitsurteile als bei Ostberliner Schüler(inne)n.

Kognitive Unterschiede innerhalb der getesteten Person darf man nicht nur in der Größe der Ausprägung bestimmter kognitiver Merkmale suchen, denkbar sind auch interindividuell unterschiedliche "kognitive Stile". Als "kognitiven Stil" bezeichnet man die Art und Weise, in der Individuen mit Informationen umgehen, z.B. feldunabhängig versus feldabhängig, analytisch versus global, verdrängend versus sensitivierend, akzentuierend versus nivellierend usw. (siehe z.B. Kong, 1989; Baron, 1985). Die Leistungsunterschiede zwischen den ost- und den westdeutschen Bewerber(inne)n könnten natürlich auch die Folge unterschiedlicher kognitiver Stile in den neuen und in den alten Bundesländern sein.

4. Zusammenfassung und Ausblick

In dem vorliegenden Bericht wurden die durchschnittlichen Testleistungen von Bewerber(innen) für Verwaltungen in den neuen Ländern mit denen von Bewerber(innen) für Verwaltungen in den alten Bundesländern verglichen. Die dabei nachgewiesenen Leistungsunterschiede zuungunsten der Ostdeutschen waren zwar statistisch signifikant, sind quantitativ aber als gering einzustufen. Im Hauptteil des Artikel wurden unterschiedliche Interpretationsmöglichkeiten der Befunde vorgestellt und diskutiert. Die Weite des aufgezeigten Interpretationsraums zeigt, daß es auf schwierige Fragen wie die Frage nach den Ursachen für kulturspezifische Leistungsunterschiede keine einfachen Antworten gibt und daß es nicht ausreicht, die Lösungen von Bewerber(innen) einfach auszuzählen. In Bezug auf eine testgestützte Personalauswahl in den neuen Ländern läßt sich daraus ableiten, daß die Probleme weniger in der Testanwendung als vielmehr in der sachgerechten **Interpretation** der Testergebnisse liegen. Vor diesem Hintergrund ist vor allem vor der sich ausweitenden Praxis der Anwendung psychologischer Tests durch Laien zu warnen. Samuda (1989, p.25) zitiert in diesem Zusammenhang Goldman, der 1961 bereits formuliert hat, daß für jeden Betrag in Höhe von 500 \$, der für Tests ausgegeben wird, ein Betrag in Höhe von 15.000 \$ für Fachpersonal zur Anweisung und Interpretation der Tests ausgegeben werden sollte.

Die quantitative Geringfügigkeit der beobachtbaren Unterschiede und die sich in ersten Analysen abzeichnende Treffsicherheit der *DGP* Prognosen auch in den neuen Ländern lassen den gesamtdeutschen Einsatz der *DGP*-Tests unter drei Voraussetzungen gerechtfertigt erscheinen. Erstens muß die Überwachung der Eignungsuntersuchungen auf kulturbedingte Testverzerrungen hin andauern, zweitens müssen aus der Analyse Konsequenzen für die Veränderung bestehender Tests, für die Testneuentwicklung und für die Rekrutierung der Bewerber(innen) gezogen werden und drittens muß in Bezug auf das Kriterium des Ausbildungs- und Berufserfolgs der Nachweis gleicher Test-Vorhersageleistungen in beiden Gruppen (Ost und West) erbracht werden.

Ad 1:

Die *DGP* wird den Einsatz der Tests in den neuen Bundesländern weiterhin wissenschaftlich begleiten und kontrollieren. Zur Zeit werden unter anderem die folgenden Fragen überprüft:

- * Bearbeiten die Bewerber(innen) in den neuen Ländern in der vorgegebenen Zeit ebenso viele Testfragen wie die Bewerber(innen) in den Altbundesländern?
- * Sind die Gruppenunterschiede auf Aufgabenebene auf ganz bestimmte - kritische- Fragen zurückzuführen?
- * Gibt es Beurteilereffekte in dem Sinne, daß westdeutsche Psycholog(inn)en ostdeutsche Bewerber(inn)en anders beurteilen als westdeutsche Bewerber(inn)en- und umgekehrt?
- * Erzielen die Bewerber(innen) aus den neuen Ländern bessere Leistungen, wenn man das Testmaterial als Quellen ehemaliger DDR-Literatur zusammenstellt?

Über die Ergebnisse der Untersuchungen wird im nächsten Heft berichtet.

Ad 2:

In den Fällen, in denen es plausible Vermutungen über die Ursachen der Gruppenleistungsunterschiede gab, hat die *DGP* bereits Tests geändert. Solche Änderungen betrafen etwa die gemeinschaftskundlichen Wissensfragen. Änderungen wurden aber nicht nur unter dem Gesichtspunkt der Leistungen vorgenommen, sondern berücksichtigten auch die Frage der Akzeptanz der Tests durch die Bewerber(innen) aus den neuen Ländern. Wenn etwa bei der alltagsorientierten "Postaufgabe" bislang nur Städtenamen aus den alten Ländern zuzuordnen waren, so sind nach der Änderung jetzt auch ostdeutsche Städtenamen vertreten.

Andererseits sollten Tests nicht vorschnell verändert werden. Gerade für die Anwendung von Tests in neuen Populationen empfiehlt es sich bewährter Testverfahren einzusetzen, bei denen man über entsprechend umfangreiche Erfahrungen verfügt und für die verlässliche Kennwerte vorliegen (Lonner & Sundberg, 1985).

Die Verwaltungen sind aufgerufen, die Grundquote geeigneter Bewerber(innen) in den neuen Ländern durch Maßnahmen zur Information, Werbung und Imagepflege zu verbessern.

Ad 3:

Der Nachweis der prädiktiven Validität kann zur Zeit nur begrenzt geführt werden, da die *DGP*-getesteten Bewerber(innen) in den neuen Bundesländern sich noch überwiegend in der Ausbildung befinden. Was momentan möglich ist -Bewährungskontrolle mit den Zwischenzeugnissen als Kriterium- hat die *DGP* geleistet. Die hier im Heft an anderer Stelle berichteten Ergebnisse sind ermutigend. Es gibt zur Zeit keinen Grund anzunehmen, daß die Tests in den neuen Ländern weniger aussagekräftig sind als in den Altbundesländern.

Die Frage nach der Angemessenheit der westlichen Testverfahren in den östlichen Bundesländern, läßt sich aber nicht durch die Registrierung von Leistungsunterschieden oder Leistungsgleichheiten der beiden Gruppen beantworten, sondern muß sich letztendlich an diesem Kriterium der **Vorhersageleistung** orientieren. Selbst wenn ein Test nachweisliche Verzerrungen zuungunsten einer spezifischen Gruppe mit sich bringt, bedeutet dies nicht, daß der Test untauglich ist. Die Frage muß stets lauten, tauglich oder untauglich **wofür?** Vijver und Poortinga (1992) bringen in diesem Zusammenhang das Beispiel eines kognitiven Fähigkeitstests mit einem hohen Anteil an sprachlichen Aufgabenkomponenten, der zum einen bei Muttersprachler(inne)n und zum anderen bei zugezogenen Ausländer(inne)n eingesetzt wird. Wenn man aufgrund der Testergebnisse Rückschlüsse über die allgemeine intellektuelle Leistungsfähigkeit ziehen will, muß man den Einsatz dieses sprachgebundenen Tests ablehnen. Will man aber eine Prognose über die Erfolgsaussichten beim Besuch solcher Schulen erstellen, in denen in der Sprache des Tests unterrichtet wird, so kann dergleiche Test eine gute Wahl darstellen. Die Entscheidung darüber, welche Fragestellungen man mit Hilfe der Testergebnisse überhaupt beantwortet haben möchte, ist unabhängig von der Analyse der Testergebnisse.

5. Literatur

- Althoff, K. (1973). Untersuchungen zum Einfluß von Berufsausbildungen und Berufstätigkeiten auf die Ergebnisse einer Eignungstestbatterie. *Diagnostica*, 19, 143-180.
- American Psychological Association. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Ankney, C.D. (1992). Sex differences in relative brain size: the mismeasure of woman, too? *Intelligence*, 16, 329-336.
- Baron, J. (1985). What kinds of intelligence components are fundamental? In S.F. Chipman, J.W. Segal & R. Glaser (Hrsg.), *Thinking and learning skills* (Bd. 2, Research and open questions, S. 365-390). Hillsdale, NJ: Erlbaum.
- 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). Göttingen: Hogrefe.
- Becker, P. (1991). Beim freien Text die alten Fehler. *Der Tagesspiegel*, 13933.
- Brambring, M. (1982). Spezielle Eignungsdiagnostik. In K.J. Groffmann & L. Michel (Hrsg.), *Enzyklopädie der Psychologie. Themenbereich B, Methodologie und Methoden. Serie II, Psychologische Diagnostik. Band 1, Grundlagen psychologischer Diagnostik* (Bd. 1, S. 414-481). Göttingen: Hogrefe.