



Deutsche Gesellschaft für Personalwesen e.V.

DGP INFORMATIONEN

47. Jahrgang, Heft 56, September 1999

Inhalt	Seite
Vorwort	1
Personalentwicklung (PE) in den neuen Bundesländern - Qualifizierungsreihe zum/zur Referent/in für Personalentwicklung <i>Joachim Rieke, Carmen Knorz, Kerstin Lange</i>	3
Potential-Assessment Center zur Bildung eines "Förderkreises Führungsnachwuchs" in einer Kreisverwaltung <i>Bernd Wolf</i>	17
Korruptionsvorsorge aus Fürsorge <i>Ina Voigt</i>	31
Konfliktberater/in bei der Polizei: Von der Stellenausschreibung zur Personalauswahl <i>Volker Klein-Moddenborg</i>	37
Personalauswahl mit dem BIS-r-DGP Test: Evaluation der prognostischen Validität <i>Martin Kersting</i>	43
Aus der (wechselvollen) Geschichte der Deutschen Gesellschaft für Personalwesen e.V. (DGP) <i>Klaus Althoff</i>	61

ISSN 0948-0579

Personalauswahl mit dem BIS-r-DGP Test: Evaluation der prognostischen Validität

M. Kersting

Die Qualität der Personalauswahl ist eine entscheidende Voraussetzung für den Erfolg einer Organisation. Dies gilt insbesondere in Zeiten, die durch Strukturwandel und technologische Innovationen einerseits und Personalabbau sowie Kosteneinsparungen andererseits geprägt sind. Durch die "Verschlankung" der Organisationen wachsen die Anforderungen, die an das verbleibende Personal gestellt werden – fehlerhafte Auswahlentscheidungen zeitigen unter diesen Umständen in organisatorischer, menschlicher und finanzieller Hinsicht fatale Konsequenzen.

Seit einiger Zeit kommt bei der *Deutsche Gesellschaft für Personalwesen (DGP)* ein modifiziertes mündliches und schriftliches Auswahlverfahren zum Einsatz. Dieses neue Verfahren wurde u.a. bei einer großen Finanz- und Steuerverwaltung eingesetzt, um unter einer großen Anzahl von Bewerber(inne)n diejenigen auszuwählen, die zum Fachhochschulstudium zur Diplom-Finanzwirtin/zum Diplom-Finanzwirt zugelassen werden sollen. Mittlerweile liegen die Ergebnisse der Zwischenprüfungen dieser Personen vor, so daß eine erste empirische Evaluation des neuen Auswahlverfahrens geleistet werden kann. Im folgenden wird das Vorgehen bei der Personalauswahl skizziert (Abschnitt 1). Bevor im Abschnitt 3 die im Auswahlverfahren getroffenen Prognosen den Ergebnissen der Zwischenprüfung gegenübergestellt werden, soll im zweiten Abschnitt eine Diskussion der Möglichkeiten und Grenzen der vorliegenden Evaluation vorangestellt werden.

1. Das Auswahlverfahren

Bei dem Auswahlverfahren wird im wesentlichen eine *zweistufige* Strategie verfolgt. Zunächst absolvieren die Bewerber(innen) einen ca. viereinhalbstündigen schriftlichen Leistungstest. Über die Zulassung zum Test entscheidet die einstellende Organisation. Im konkreten Fall hat die Finanz- und Steuerverwaltung diese Vorauswahl im wesentlichen aufgrund der Schulnoten in den Fächern "Deutsch" und "Mathematik" sowie aufgrund der "Abiturnote" entschieden. Nur die im Test zumindest teilweise erfolgreichen Bewerber(innen) werden zum zweiten, mündlichen Teil des Auswahlverfahrens eingeladen.

1.1 Der schriftliche Leistungstest

Als schriftliches Verfahren wurde der neue Leistungstest (Version 1.0) der DGP eingesetzt. Dieser sogenannte "BIS-r-DGP" Test wurde von Kersting und Beauducel mit dem Ziel konstruiert, im Rahmen der gegebenen testökonomischen Grenzen einer Abbildung wesentlicher Komponenten des Berliner Intelligenzstrukturmodells ("BIS") zu ermöglichen. Das Berliner Intelligenzstrukturmodell wurde von Prof. Dr. A.O. Jäger (1982, 1984) entwickelt, der von 1955-1968 Leiter der DGP war. Auf die Messung des im BIS-Modell vorgesehenen Einfallsreichtums wird im BIS-r-DGP Test verzichtet, ausgewertet wird lediglich die "Einfallsmenge". Der Test umfasst 38 Denkaufgaben. Bei dem Großteil dieser 38 Aufgaben lagen der Testentwicklung entsprechende Aufgaben des Berliner Intelligenzstruktur-Tests ("BIS-4 Test"; Jäger, Süß und Beauducel, 1997)

als Vorlage zugrunde. Die Aufgaben des BIS-r-DGP Tests sind aber nicht identisch mit den Aufgaben des BIS-4 Tests. Der BIS-r-DGP Test umfasst ausschließlich unveröffentlichtes, exklusiv für die Auftraggeber(innen) der DGP zur Verfügung gestelltes, Aufgabenmaterial. Neun der 38 Aufgaben des neuen Tests haben ihren Ursprung in vorherigen DGP-Tests, die sich in zahlreichen Bewährungskontrollen bewährt haben. Zwei Merkfähigkeitsaufgaben wurden vollständig neu konstruiert. Weitere Unterschiede zwischen dem BIS4 Test und dem BIS-r-DGP Test ergeben sich durch die Verwendung von Antwortbogen. Während die Testand(inn)en im BIS-4-Test ihre Lösungsversuche direkt in das Aufgabenheft eintragen, erfolgt das Markieren der Lösungen im BIS-r-DGP Test auf maschinenlesbaren Antwortbogen. Die Normen zum BIS-r-DGP Test basieren ursprünglich auf Daten von 3274 Personen, die im Zeitraum von August 1996 bis April 1997 im Rahmen ihrer Bewerbung für eine Ausbildung zum gehobenen Verwaltungsdienst den BIS-r-DGP Test (Version 1.0) bearbeitet haben. Diese Normen werden regelmäßig an zusätzlich gewonnenen aktuellen Daten von jährlich ungefähr 3000 - 4000 Personen überprüft und ggf. aktualisiert. Die aktuelle Datenbank zum BIS-r-DGP Test umfasst Testergebnisse von 9250 Personen. Mit den insgesamt 38 Aufgaben des BIS-r-DGP Tests werden Informationen zur Verarbeitungskapazität, Merkfähigkeit, Bearbeitungsgeschwindigkeit und Einfallsmenge sowie zu den sprachgebundenen, zahlengebundenen und figuralen intellektuellen Fähigkeiten der Bewerber(innen) erfasst. Eine ausführliche Darstellung des BIS-r-DGP-Tests findet sich bei Kersting (1995 b), Informationen zu einigen psychometrischen Testgütekriterien sind bei Kersting und Beauducel (1997) dokumentiert. Der BIS-r-DGP-Test gehört zu den bislang noch seltenen Leistungsprüfverfahren, bei denen von Beginn der Testkonstruktion an auch Daten von Testand(inn)en aus den neuen Bundesländern systematisch berücksichtigt wurden. Ein solches Vorgehen erscheint aufgrund der möglicherweise ostwestspezifisch unterschiedlichen Wirkungsweisen von Testmerkmalen (Kersting, 1996) sowie aufgrund der Auswirkungen von gruppenspezifisch unterschiedlichen Testdurchschnittsleistungen auf den Vorhersagefehler (Kersting, 1995 a) geboten.

Neben den intellektuellen Leistungen im engeren Sinne wurden die Kenntnisse der Bewerber(innen) in verschiedenen Wissensdomänen (Gemeinschaftskunde, Wirtschaft, Politik und Literatur) sowie in der Rechtschreibung überprüft. Außerdem kam eine Aufgabe zur Simulation einer Büroroutinetätigkeit zum Einsatz, bei deren Bearbeitung es auf Konzentration und Sorgfalt ankommt.

Alle Bewerber(innen) erhalten vor dem Verfahren eine 39 DIN-A-5 Seiten umfassende Broschüre mit allgemeinen Informationen über Personalauswahlverfahren sowie über einen beispielhaften Ablauf des schriftlichen und mündlichen Verfahrensteils. Der Großteil der Broschüre ist Beispielaufgaben zum schriftlichen Test vorbehalten und regt die Teilnehmer(innen) zur Übung an. Die Broschüre informiert abschließend darüber, welche Rechte Teilnehmer an Auswahlverfahren haben und wie die Bewerber(innen) ihre Rechte wahren können.

Die Testleistungen der Bewerber(innen) werden auf der weiter unten dargestellten "Empfehlungsgradskala" (siehe Tabelle 1) beurteilt, wobei die einzelnen

Leistungsdimensionen vor dem Hintergrund der in der Ausbildung und in der Praxis des Berufsbildes "Finanzwirt(in)" gestellten Anforderungen unterschiedlich gewichtet werden.

1.2 Der mündliche Teil

Bewerber(innen), die im schriftlichen Testverfahren eine bestimmte Mindestleistung erzielen, werden zu dem einige Wochen später stattfindenden mündlichen Teil des Verfahrens eingeladen. Dieser Verfahrensteil besteht aus insgesamt vier Stationen, nämlich zwei Gruppensituationen und zwei Vorstellungsgesprächen.

Zunächst sollen die Bewerber(innen) in Kleingruppen von normalerweise sechs (min. 4 bis max. 7) Personen zwei verschiedene Alltagsthemen miteinander diskutieren, beispielsweise die Einrichtung von Volksbegehren, die Frage der doppelten Staatsbürgerschaft, die Abschaffung des dreizehnten Schuljahres, die Erhebung eines Kinderlastenausgleichs bzw. eines Zukunftsbeitrags für Kinderlose usw. Das Thema der ersten Diskussionsrunde sollen die Gruppenmitglieder selbst bestimmen (Entscheidungsfindungsaufgabe), das weitere Thema gibt die Kommission vor.

Auch in der zweiten Station agieren die Bewerber(innen) in der Gruppe, die Zusammensetzung der Gruppenmitglieder wird aber gegenüber der ersten Station variiert. Während die Bewerber(innen) in der ersten Station lediglich miteinander diskutieren, sollen die Teilnehmer(innen) in der zweiten Station gemeinsam eine Gruppenaufgabe lösen, beispielsweise das konkrete Betreuungsprogramm für einen Schüler(innen)austausch organisieren oder konkrete Vorschläge für die Gewinnung von Sponsoren für eine AZUBI-Zeitung unterbreiten usw.

Die beiden Gruppenstationen werden von einer Kommission beobachtet und beurteilt. Die Kommission setzt sich aus Vertreter(inne)n der einstellenden Organisation einerseits sowie einer/einem Diplom-Psychologin/Psychologen der DGP andererseits zusammen. Die Kommissionsmitglieder beteiligen sich nicht an der Diskussion oder Aufgabenlösung.

Alle Kommissionsmitglieder wurden in Schulungen auf ihre Beobachtungs- und Beurteilungstätigkeit vorbereitet. Der Beobachtung und Beurteilung der Kommissionsmitglieder liegt ein standardisierter Beurteilungsbogen mit vorgegebenen Beurteilungskategorien zugrunde.

- Soziale Kompetenz/Teamfähigkeit
- Kommunikationsfähigkeit
- Sicherheit/Belastbarkeit
- Sachbezogenheit
- Einfallsreichtum
- Zielorientierte Initiative

Abbildung 1: Die sechs in den Gruppenstationen eingestuft "Verhaltensdimensionen"

Beobachtet werden solche Verhaltensdimensionen, die sich aufgrund von Anforderungsanalysen als berufsrelevant erwiesen haben. Bei ihrem Urteil sollen sich die Kommissionsmitglieder auf konkrete Verhaltensbeobachtungen stützen, Beobachtung und Beurteilung sollen voneinander getrennt werden.

In den beiden Gruppenstationen werden die in Abbildung 1 dargestellten sechs Verhaltensdimensionen beurteilt und auf einer Skala eingestuft. Den in der Abbildung genannten Fähigkeitsbegriffen wurden zwecks Operationalisierung beispielhaft eine Liste von Einzelbeobachtungen zugeordnet. Während und nach den beiden Gruppensituationen werden die Teilnehmer(innen) zunächst von jedem Kommissionsmitglied unabhängig beobachtet und durch den Eintrag eines Skalenwerts pro Verhaltensdimension beurteilt. Im Anschluß an jede Station haben die Kommissionsmitglieder dann Zeit, sich über ihre Beobachtungen zum Ablauf des Geschehens und zum Verhalten der Teilnehmer(innen) auszutauschen. Die Kommissionsmitglieder diskutieren ihre aus den Beobachtungen gezogenen Schlußfolgerungen (die Einzelbeurteilungen pro Verhaltensdimension) und legen für jede Verhaltensdimension ein gemeinsames Urteil fest. Dabei wird besonderes Augenmerk darauf gelegt, daß nur an konkreten Verhaltensweisen festgemachte Urteile abschließend berücksichtigt werden.

Im Anschluß an die beiden Gruppensituationen hat jede Bewerberin/jeder Bewerber zwei Vorstellungsgespräche zu absolvieren. Eines der beiden Vorstellungsgespräche wird mit den Vertreter(inne)n der einstellenden Organisation geführt, das andere Gespräch ist ein vier-Augen-Interview mit einer Diplom-Psychologin/einem Diplom-Psychologen der DGP.

Das zielgerichtete Gespräch der Mitarbeiter(innen) der DGP mit den Bewerber(inne)n dient insbesondere der Exploration, Beobachtung und Beurteilung der berufsbezogenen Motivation, die zusätzlich zu den in Abbildung 1 genannten Anforderungsdimensionen das siebente Eignungsmerkmal bildet.

Außer der berufsbezogenen Motivation können aber auch die Beobachtungen und Beurteilungen der anderen Verhaltensdimensionen (mit Einschränkungen bei der sozialen Kompetenz) im Vorstellungsgespräch noch einmal vertieft werden.

Mit dem Vorstellungsgespräch verbinden sich darüber hinaus noch weitere Ziele, die insbesondere in dem Gespräch der Bewerber(innen) mit den Vertreter(inne)n der einstellenden Organisation realisiert werden. So kann in diesem Gespräch in werbender oder korrigierender Absicht über die Ausbildung, den Beruf und die einstellende Organisation informiert werden oder es können Fragen der Bewerber(innen) beantwortet werden. Diese informationsvermittelnden Funktionen des Interviews dienen u.a. der sozialen Akzeptanz des Auswahlverfahrens (siehe Kersting, 1998; Schuler, 1990, 1993).

1.3 Bewertungsprozeß, Bewertungsskala

Sowohl die Ergebnisse des schriftlichen Leistungstests als auch die Beurteilungen des Verhaltens im mündlichen Teil des Auswahlverfahrens wer-

den auf der in Tabelle 1 dargestellten fünfstufigen/bzw. (bei Berücksichtigung der Zwischenstufen) neunstufigen Skala festgehalten.

Aus dem Urteil über die Testleistungen einerseits und dem Urteil über den mündlichen Teil des Auswahlverfahrens andererseits wird das *abschließende Verfahrensurteil (Gesamtempfehlungsgrad)* gebildet, wobei die schriftlichen Leistungen etwas höher gewichtet werden. Dieser abschließende *Gesamtempfehlungsgrad*, der ebenfalls den in Tabelle 1 dargestellten Abstufungen entspricht, stellt die Empfehlung der DGP hinsichtlich der Auswahlentscheidung dar.

2. Untersuchungsteilnehmer(innen), Prädiktoren und Kriterien

2.1 Untersuchungsteilnehmer(innen) und Datenausfälle

Für 135 Fachhochschulstudent(inn)en des Fachbereichs Steuerverwaltung lagen sowohl die Testergebnisse als auch die Ergebnisse der Anfang 1998 absolvierten Zwischenprüfung vor. Das Durchschnittsalter dieser 99 Anwärt(er)innen und 36 Anwärt(er) betrug zum Zeitpunkt der Untersuchung 20 Jahre und zwei Monate, wobei mit 57% das Alter der Mehrheit der Untersuchungsteilnehmer(innen) zwischen 17 und 19 Jahren lag. Für drei der 135 Untersuchungsteilnehmer(innen) konnte zum Zeitpunkt der Berichterstellung das Ergebnis des mündlichen Teil des Auswahlverfahrens nicht ermittelt werden, so daß die nachfolgenden Ergebnisse für den Gesamtempfehlungsgrad auf einer Stichprobe von 132 Personen basieren.

2.2 Prädiktoren: Ergebnisse des Auswahlverfahrens

Tabelle 1 zeigt die Verteilung der testbasierten Empfehlungsgrade sowie der Gesamtempfehlungsgrade in der vorliegenden Untersuchung.

DGP-Empfehlungsgrad	Testergebnis (N=135)		Gesamtergebnis (N=132)	
	Häufig- keit	in Prozent	Häufig- keit	in Prozent
„1“ („entspricht nicht den Anforderungen“)				
„1,5“				
„2“ („entspricht weitge. nicht d. Anforderungen“)				
„2,5“				
„3“ („entspricht nur teilweise den Anforderungen“)				
„3,5“	1	0,7 %	2	1,5 %
„4“ („entspricht weitgehend den Anforderungen“)	22	16,3 %	35	26,5 %
„4,5“	32	23,7 %	69	52,3 %
„5“ („entspricht den Anforderungen“)	80	59,3 %	26	19,7 %

2.3 Kriterium: Zwischenprüfung

Die Eignungsuntersuchungen wollen neben dem Berufserfolg auch den Ausbildungserfolg vorhersagen. Die Kund(inn)en der DGP sollen mit Hilfe der DGP u.a. davor bewahrt werden, Bewerber(innen) einzustellen, die den späteren Anforderungen der Ausbildung nicht genügen. Als ein Kriterium für die Treffsicherheit der DGP-Verfahren kann daher näherungsweise die Leistung in der Ausbildung gelten (zur eingeschränkten Eignung dieses Kriteriums siehe unten, Punkt 3.1.2). Für die hier analysierten Anwärterinnen und Anwärter standen zum Zeitpunkt der Analyse erst die Leistungen der Zwischenprüfung fest. Grundlage der vorliegenden Evaluation ist das *Gesamtergebnis der Zwischenprüfung*. Die Zwischenprüfungsergebnisse werden auf drei unterschiedlichen Niveaus analysiert. Zunächst wird das dichotomisierte Kriterium betrachtet und es wird darüber berichtet, ob die Anwärter(innen) die Zwischenprüfung bestanden haben oder nicht. In einem nächsten Schritt werden die Zwischenprüfungsergebnisse in den Notenstufen "sehr gut" bis "ungenügend" betrachtet. Abbildung 2 stellt die Umrechnung der Punkte in die Notenstufen da. Auf dem höchsten Präzisionsniveau erfolgt eine Analyse der Zwischenprüfungsergebnisse in ganzen Punkten, wobei die auf zwei Nachkommastellen genauen Ergebnisse der Zwischenprüfung jeweils zuvor auf einen glatten Punktwert auf oder abgerundet wurden.

von	0,00 bis	1,99 Punkte	=	ungenügend
von	2,00 bis	4,99 Punkte	=	mangelhaft
von	5,00 bis	7,99 Punkte	=	ausreichend
von	8,00 bis	10,99 Punkte	=	befriedigend
von	11,00 bis	13,49 Punkte	=	gut
von	13,50 bis	15,00 Punkte	=	sehr gut

Abb. 2: Information über die Bewertung der Prüfungsleistung

Die Ebene der gerundeten Punktzahl unterscheidet sich von der Ebene der Notenstufen. Während z.B. die Punktzahl "4,77" laut Abbildung 2 der Notenstufe "mangelhaft" entspricht, wird der Wert bei der Analyse der Punkte auf "5,0" aufgerundet. Weder die Punktzahl noch die Notenstufe ist ein hinreichender Indikator für das Bestehen der Zwischenprüfung. Neben der Punktzahl oder Note zählt auch die Anzahl an Fächern, die mit mangelhaften oder ungenügenden Leistungen absolviert wurden. Falls die Leistungen in mehr als zwei Fächern mit "mangelhaft" bewertet wurden, gilt die Prüfung auch dann als "nicht bestanden", falls der Punktwert/bzw. die Note insgesamt im Bereich "ausreichend" liegt. Dies war in der vorliegenden Untersuchungsgruppe bei zwei Personen der Fall.

Abbildung 2 zeigt die Verteilung der Ergebnisse der Zwischenprüfung für die unterschiedlichen Präzisionsniveaus.

Tab. 2: *Gesamtergebnis der Zwischenprüfung (N=135)*

Ergebnis			Note			Punkte		
Ergebnis	N	in %	Note	N	in %	Pkt.	N	in %
nicht bestanden	10	7,4 %	ungenügend			1		
			mangelhaft	8	5,9 %	2		
						3	3	2,2 %
						4	2	1,5 %
bestanden	125	92,6 %	ausreichend	41	30,4 %	5	7	5,2 %
						6	9	6,7 %
						7	16	11,9 %
			befriedigend	72	53,3 %	8	24	17,8 %
						9	25	18,5 %
						10	22	16,3 %
			gut	14	10,4 %	11	19	14,1 %
						12	6	4,4 %
			sehr gut			13	2	1,5 %
						14		
						15		

3. Möglichkeiten und Grenzen der vorliegenden Bewährungskontrolle

Bevor im Kapitel 4 die im Auswahlverfahren getroffenen Prognosen den Ergebnissen der Zwischenprüfung gegenübergestellt werden, soll auf einige Aspekte aufmerksam gemacht werden, die bei der Interpretation der Daten berücksichtigt werden müssen.

3.1 Unterschätzung der Aussagekraft des Auswahlverfahrens

3.1.1 Unterschätzung aufgrund von Streuungseinschränkungen

Die in Tabelle 1 dargestellte Verteilung der Empfehlungsgrade verdeutlicht ein grundsätzliches Problem von Bewährungskontrollen im Bereich der Personalauswahl. Obwohl es neun Kategorien von Empfehlungsgraden gibt, sind bis auf die Ausnahme einer Person (Test) bzw. zwei Personen (Gesamt) mit dem Empfehlungsgrad "3,5" nur die Kategorien "4", "4,5" und "5" besetzt. Dies ist verständlich, da es sich ja hier nicht um die Gesamtbewerber(innen)gruppe handelt, sondern um eine einseitige Auswahl aus dem Gesamtpool. Schließlich stellt die Organisation nur erfolgversprechende Kandidat(inn)en ein, und nur bei "5ern" und "4ern" besteht nach den

Erfahrungen der DGP eine relativ sichere Aussicht, die Prüfungen erfolgreich absolvieren zu können. Von der Einstellung von "1er" oder "2er" Kandidat(inn)en rät die DGP dringend ab. Da die einstellende Organisation diesen Empfehlungen weitgehend nachgekommen ist, beschränkt sich die Bewährungsprobe also überwiegend auf "sehr gute" Teilnehmer(innen). In der vorliegenden Stichprobe gab es nur zwei Personen mit dem Gesamtempfehlungsgrad "3.5" und niemanden mit einem Gesamtempfehlungsgrad von "1", "2" oder "3.0". In der Begrifflichkeit der Statistik spricht man in diesen Fällen von einer "abgeschnittenen" Verteilung und meint damit, daß ein Teil der Bewerber(innen) nicht in die Analyse mit eingeht: nämlich die in bezug auf die Bewerber(innen)gruppe am stärksten vertretene Gruppe von Personen mit einem Ergebnis von "3" und schlechter. Da diese Personen die Ausbildung erst gar nicht begonnen haben, kann die DGP auch nicht nachweisen, daß diese Personen zu Recht von der Ausbildung ausgeschlossen wurden. Dies ist ein Grund dafür, daß die berichteten Zusammenhänge zwischen den Ergebnissen der DGP-Eignungsuntersuchungen und den schriftlichen Leistungen in den Klausuren den tatsächlich bestehenden Zusammenhang deutlich *unterschätzen*. Dies gilt insbesondere falls, wie Althoff (1984) vermutet, bei den "schlechteren" Empfehlungsgraden ein engerer Zusammenhang besteht als bei den "besseren" Empfehlungsgraden (kurvilinearer Zusammenhang).

Einen Anhaltspunkt für die extreme Verteilungseinschränkung erhält man, wenn man die Verteilung der Testergebnisse der zugelassenen Bewerber(innen) mit den Ergebnissen in der nichtvorausgewählten Eichstichprobe von 3274 Bewerber(inne)n vergleicht. (Für nähere Angaben zur Eichstichprobe siehe Kersting und Beauducel, 1997.) Über die Hälfte (54%) der Bewerber(innen) in der Eichstichprobe erzielten im Test einen Empfehlungsgrad von "3" oder schlechter. Die Varianz der testbasierten Empfehlungsgrade betrug in der Eichstichprobe 2.047 gegenüber einer eingeschränkten Varianz in Höhe von .154 in der im vorliegenden Bericht analysierten Gruppe der zugelassenen Personen. Noch eine weitere Zahl mag das Ausmaß der Vorauswahl verdeutlichen. Die über alle intellektuellen Dimensionen gemittelten Leistungen variieren normalerweise auf einer Skala von "70" bis "130", wobei im Durchschnitt der Eichstichprobe ein Wert von 100,64 erzielt wurde. 49,1 % aller Bewerber(innen) der Eichstichprobe erzielten Testleistungen kleiner oder gleich einem Standardwert von 100. Dies dürfte auch für die Gesamtgruppe derjenigen Kandidat(inn)en gelten, die im Rahmen ihrer Bewerbung für die Verwaltung des BIS-r-DGP Leistungsprüfverfahren absolviert haben. Betrachtet man nun im Vergleich dazu diejenigen 135 Personen, die zur Ausbildung zugelassen wurden, so weist die schlechteste (!) Person dieser Gruppe mit einem Wert von 102 noch eine Leistung auf, die über dem Durchschnitt (!) der Eichstichprobe liegt! Der Mittelwert der zur Ausbildung zugelassenen Gruppe liegt bei 109,53. Die Streuung der Testwerte ist in der hier analysierten Gruppe ein vielfaches geringer als die Streuung der Testwerte im Normalfall der Auswahlentscheidung. Angesichts einer solchen extrem "abgeschnittenen" Verteilung würde es nicht gegen das Auswahlverfahren sprechen, falls sich keinerlei Zusammenhang zwischen den Ergebnissen des Auswahlverfahrens und den Ergebnissen der Zwischenprüfung zeigen würde. Selbst bei einem ausbleibenden Zusammenhang könnte man angesichts der

extrem ausgewählten Untersuchungsgruppe nicht ausschließen, daß für eine unausgelesene Gesamtgruppe die Leistungen im Auswahlverfahren deutlich mit den Leistungen in der Ausbildung assoziiert wären. Umgekehrt formuliert: Sollten sich trotz der extremen Verteilungseinschränkung noch statistisch bedeutsame Zusammenhänge zwischen dem Auswahlverfahren und den Ausbildungsleistungen zeigen, so wäre dies ein klarer Hinweis darauf, daß für die Gesamtgruppe substantielle Effekte vorliegen müssen und die aufgrund der Ergebnisse im Auswahlverfahren nicht zugelassenen Bewerber(innen) mit hoher Wahrscheinlichkeit in der Ausbildung versagt hätten.

Die tatsächliche Aussagekraft des Auswahlverfahrens kann mit der vorliegenden Untersuchung nicht demonstriert werden. Die eigentliche Leistung des Auswahlverfahrens besteht darin, innerhalb der großen Zahl mehrerer 100 zum Test zugelassenen Bewerber(innen) die aussichtsreichen Kandidat(inn)en auszuwählen. Da nur ein Teil der Bewerber(innen) überhaupt zum mündlichen Teil zugelassen wird, trägt der schriftliche Leistungstest die Hauptverantwortung für die Auswahlentscheidung. Daß professionelle psychologische Intelligenztestverfahren eine vergleichsweise exzellente Prognose beruflicher Leistungen ermöglichen, ist das eindeutige Ergebnis einer immensen Anzahl nationaler und internationaler wissenschaftlicher Studien aus über 85 Jahren entsprechender Forschung. Mittlerweile ist es mit Hilfe sogenannter Metaanalysen möglich, die Ergebnisse von Einzelstudien zur gleichen Fragestellung statistisch kontrolliert zusammenzufassen. Die auf den Daten mehrerer 100.000 Personen beruhenden Ergebnisse stellen den deutlichen Zusammenhang zwischen den Ergebnissen in Leistungstests und dem Ausbildungs- bzw. Berufserfolg unter Beweis (siehe z.B. Hunter und Hunter, 1984; Schmidt und Hunter, 1998. [Die zuletztgenannte Übersicht von Schmidt und Hunter liegt mittlerweile auch in deutscher Übersetzung vor, siehe Kleinmann & Strauß, 1998]). Dem Stand der Forschung zufolge muß davon ausgegangen werden, daß sich die Bewerber(innen), die in einem Intelligenztest deutlich schlecht abgeschnitten haben und daher nicht zur Ausbildung zugelassen wurden, mit einer hohen Wahrscheinlichkeit in der Ausbildung und / oder im Beruf nicht bewährt hätten. Dies läßt sich aber anhand des vorliegenden Datensatzes, bei dem die im Test schlechten Bewerber(innen) nicht berücksichtigt sind, nicht mehr aufzeigen. Da allein die Kosten für einen Ausbildungsplatz mit über DM 100.000 zu veranschlagen sind, haben sich die Kosten für das Auswahlverfahren bereits amortisiert, falls nur wenige Fälle zukünftiger Ausbildungsversager(innen) durch das Testverfahren im voraus erkannt werden können. Die tatsächlichen Investitionen, die durch das Auswahlverfahren abzusichern sind, eröffnen sich aber erst, wenn man die erfahrungsgemäß gegenüber den schlechtesten Mitarbeiter(inne)n zumindest zwei bis dreimal so große Produktivität der besten Mitarbeiter(innen) über den gesamten Zeitraum des Verbleibs in der Organisation vergleichend betrachtet. Aufgrund der Ergebnisse von Kosten-Nutzen-Analysen (z.B. Barthel, 1988), bei denen Auswahlverfahren unter Utilitätsaspekten betrachtet werden, ist davon auszugehen, daß durch den Verzicht auf wissenschaftlich fundierte Auswahlverfahren Fehlinvestitionen in Millionenhöhe verursacht werden. Die Ergebnisse betriebswirtschaftlicher Kostenrechnungen legen den Einsatz von professionellen Auswahlverfahren zwingend nahe.

3.1.2 Unterschätzung aufgrund der Kriteriumsdefizienz
 In der vorliegenden Untersuchung konnten lediglich die Zwischenprüfungsergebnisse, nicht aber die abschließenden Ausbildungsleistungen zur Verfahrensevaluation herangezogen werden. Da die Zwischenprüfungsergebnisse keine perfekte Vorhersage des Ausbildungserfolges erlauben, weist das gewählte Kriterium bereits als Indikator des Ausbildungserfolges Defizite auf. Deutlich wird die Kriteriumsdefizienz aber vor allem dann, wenn man sich vergewärtigt, daß mit dem Auswahlverfahren nicht nur der Ausbildungserfolg, sondern letztendlich auch der *Berufserfolg* vorhergesagt werden soll. Der Berufserfolg ist ein überaus komplexes und sicherlich auch uneinheitlich definiertes Kriterium. Unbestritten dürfte aber sein, daß gute Ausbildungsleistungen nicht mit dem Berufserfolg gleichzusetzen sind. Einige Testkomponenten, vor allem aber der mündliche Teil des Auswahlverfahrens würde in Hinblick auf die Vorhersage des *Berufserfolgs* maßgeschneidert. Die diesbezüglichen Vorhersagequalitäten des Verfahrens können in der vorliegenden Studie mit dem Kriterium "Zwischenprüfung" nicht evaluiert werden, der Ausbildungserfolg (und erst recht nicht die Ergebnisse in der Zwischenprüfung) stellt nicht das "eigentliche" (ultimate) Kriterium dar.

3.2 Interpretationsgrenzen: Begrenzte Aussagen zur Validität des mündlichen Verfahrens sowie von Schulnoten
 Als Konsequenz der mehrstufigen Strategie mit dem schriftlichen Test als erstem Schritt ergibt sich, daß die Aussagekraft weiterer Prädiktoren *nur noch in Kombination mit dem schriftlichen Leistungstest* betrachtet werden kann. Der Zusammenhang zwischen den Ergebnissen im mündlichen Teil des Auswahlverfahrens sowie den Schulnoten als Prädiktoren mit den Ergebnissen in der Zwischenprüfung als Kriterium kann nicht als Beleg für deren singuläre Validität gelten. Der statistische Zusammenhang zwischen Schulnoten und dem Erfolg in der Zwischenprüfung bedeutet lediglich, daß *innerhalb* der Gruppe der Anwärter(innen) mit guten und sehr guten Testleistungen die Schulnoten aussagekräftig sind. Dies sagt nichts darüber aus, ob die Schulnoten auch in der Gruppe aller Bewerber(innen) aussagekräftig wären und womöglich den Test ersetzen könnten. In der Gruppe derjenigen Bewerber(innen), die aufgrund ihrer schlechten Testergebnisse nicht zur Ausbildung zugelassen wurden, gab es Dutzende mit einem 1er Abiturdurchschnitt. 1997 wurden lediglich 28,3 % der Bewerber(innen) mit einem Schulnotendurchschnitt im Bereich 1.0 bis 1.9 zum Studium zugelassen. Falls diese Anwärter(innen) nun Erfolg in der Zwischenprüfung haben sollten, sagt dies nichts über den Erfolg oder Mißerfolg derjenigen 71,1% der Bewerber(innen) aus, die trotz ihrer hervorragenden Schulnoten im Test oder im mündlichen Teil des Auswahlverfahrens gescheitert sind.

Vergleichbar gilt für die Aussagekraft der im mündlichen Verfahrensteil gewonnenen Urteile, daß die vorliegende Studie nur eine Prüfung der Frage zuläßt, ob dieser mündliche Verfahrensteil eine sinnvolle *Ergänzung des schriftlichen Teils* darstellt oder nicht. Die nachfolgend dargestellten Zusammenhänge zwischen den Leistungen im mündlichen Teil des Auswahlverfahrens und den Leistungen in der Zwischenprüfung sind ein Indikator dafür, daß *innerhalb* einer nach Testergebnissen ausgewählten homogen leistungsstarken Gruppe die Verhaltensbeurteilungen sinnvoll sind. Daraus läßt sich nichts darüber ableiten,

ob diese Aussagekraft auch für eine nicht vorausgewählte leistungsheterogene Gruppe gelten würde. Keinesfalls ermöglicht die vorliegende Evaluation Spekulationen über die Aussagekraft von Verhaltensbeurteilungen oder Schulnoten bei einer Gruppe, die nicht aufgrund von Tests vorausgewählt wurde.

4. Gegenüberstellung der im Auswahlverfahren gewonnenen Empfehlungen und der Ergebnissen der Zwischenprüfung

4.1 Bestehen der Zwischenprüfung
 Tabelle 3 stellt die Verteilung der abschließenden DGP-Empfehlungsgrade (Gesamtempfehlungsgrad) der Verteilung des Kriteriums (dem Erfolg in der Zwischenprüfung) in Form einer Kreuztabelle deskriptiv gegenüber. Für jede "Zelle" der Tabelle sind drei unterschiedliche Angaben verzeichnet, die am Beispiel der ersten Zelle (DGP-Empfehlungsgrad *"teilweise anforderungsgerecht (3,5)"* und dem Ergebnis *"nicht bestanden"*) kurz erläutert werden sollen. Die Angabe in der ersten Reihe dieser Zelle bezieht sich auf die absolute Stichprobengröße von 1 Anwärter(in). Die erste Zahl ist eine absolute Nennung, die in Klammern aufgeführte Zahl der entsprechende prozentuale Wert. Also: *Nur eine der insgesamt 132 Teilnehmer(innen) hat einen DGP Empfehlungsgrad von "nur teilweise anforderungsgerecht" erzielt und die Zwischenprüfung "nicht bestanden". Das sind 0,8 % der Gesamtstichprobe von 132 Personen.* Die Zahl darunter, rechts neben dem Ergebnis der Zwischenprüfung, bezieht sich nur auf die Personen, die in der Zwischenprüfung "durchgefallen" sind: *11,1 % derjenigen, die in der Zwischenprüfung "durchgefallen" sind, erreichten im DGP-Eignungsverfahren einen Empfehlungsgrad von "3,5".* Die letzte, grau-hinterlegte Angabe pro Zelle, bezieht sich nur auf die Teilnehmer(innen) mit einem DGP-Gesamtempfehlungsgrad von "3.5": *50% derjenigen, die in dem Auswahlverfahren einen Gesamtempfehlungsgrad von "3,5" erzielten, haben die Zwischenprüfung nicht bestanden.*

N=132	DGP-Gesamtempfehlungsgrad			
	„3,5“ (nur teilweise anforderungsgerecht)	„4“ (weitgehend anforderungsgerecht)	„4,5“	„5“ (anforderungsgerecht)
Ergebnis der ↓ Zwischenprüfung ↓				
„nicht bestanden“	1 (0,8%)	3 (2,3 %)	5 (3,8%)	
	11,1 %	33,3 %	55,6 %	
	50 %	8,6 %	7,2 %	
„bestanden“	1 (0,8%)	32 (24,2 %)	64 (48,5 %)	26 (19,6 %)
	0,8 %	26,0 %	52,0 %	21,1 %
	50 %	91,4 %	92,8 %	100 %
	2 (100 %)	35 (100 %)	69 (100 %)	26 (100 %)

Tab. 3: Kreuztabelle

Um die Aussagekraft der Eignungsuntersuchungen der DGP zu bestimmen, ist diese zuletzt erläuterte Zahl in den grauhinterlegten Feldern am interessantesten. Die Organisation will erfahren, inwieweit sie sich darauf verlassen kann, daß ein(e) Bewerber(in), die/der von der DGP als *„den Anforderungen entsprechend“* klassifiziert wurde, auch tatsächlich den Anforderungen der Ausbildung genügt. Die Kreuztabelle ist ein deutlicher Beleg für die hohe Treffsicherheit der DGP-Empfehlungsgrade: Von den 26 Personen, die im Auswahlverfahren als *„den Anforderungen entsprechend“* (Gesamtempfehlungsgrad „5“) beurteilt werden, gab es **keine**, die den Anforderungen der Zwischenprüfung dann entgegen der Aussage des Auswahlverfahrens nicht genügten. Alle diese Personen bestanden die Zwischenprüfung. Der Gesamtempfehlungsgrad *„anforderungsgerecht“* der DGP impliziert also quasi eine Garantie für ein erfolgreiches Abschneiden bei der Zwischenprüfung. Eine solche Garantie gibt es natürlich für Personen mit *„leichten Einschränkungen“* nicht. Die Gesamtempfehlungsgrade *„4“* und *„4,5“* (*„entspricht weitgehend den Anforderungen“*) signalisieren entsprechend der DGP-Definition leichte Einschränkungen in bezug auf den zu erwartenden Ausbildungserfolg. In der vorliegenden Untersuchung entsprachen noch ca. 92% der Teilnehmer(innen) mit diesem Empfehlungsgrad den Anforderungen der Zwischenprüfung.

Ein hohes Risiko ergibt sich – mit der diesbezüglichen Definition der DGP vollständig konform – bei dem Empfehlungsgrad *„3,5“* – *„nur teilweise den Anforderungen entsprechend“*. Hier sagt die DGP, daß aufgrund der Ergebnisse im Auswahlverfahren Bedenken gegenüber einer Zulassung bestehen. Die so beurteilten Bewerber(innen) sind nicht durchgängig leistungsschwach (das sind die Bewerber(innen) mit den Empfehlungsgraden *„2“* und *„1“*), weisen aber neben Stärken auch Schwächen auf. Eine Prognose ist hier schwierig, es sind *„Risikokandidat(inn)en“*. Tatsächlich konnte nur eine der beiden so (mit dem Gesamtempfehlungsgrad *„3,5“*) beurteilten Personen die Zwischenprüfung bestehen. Die andere Person hat die Zwischenprüfung im übrigen nur äußerst knapp (mit 6,87 Punkten) bestanden. Die Organisation hätte sich die Investition in diese beiden Auszubildenden sparen können, wäre sie dem Gesamtempfehlungsgrad der DGP gefolgt. Da es insgesamt mehrere 100 Bewerber(innen) mit einem testbasierten Empfehlungsgrad bzw. mit einem Gesamtempfehlungsgrad von *„3,5“* und schlechter gab, ist davon auszugehen, daß durch das Auswahlverfahren Fehlinvestitionen in Millionenhöhe vermieden wurden. Kritisch anzumerken ist, daß die Differenzierung zwischen Anwärter(inne)n mit einem Gesamtempfehlungsgrad von *„4“* gegenüber solchen mit einem Gesamtempfehlungsgrad *„4,5“* nur hauchdünn gelingt.

N=132	DGP-Gesamtempfehlungsgrad			
Ergebnis der ↓ Zwischenprüfung ↓	„3,5“ (nur teilweise anforderungsgerecht)	„4“ (weitgehend anforderungsgerecht)	„4,5“	„5“ (anforderungsgerecht)
„gut“		2 (1,5 %)	6 (4,5 %)	5 (3,8 %)
		15,4 %	46,2 %	38,5 %
		5,7 %	8,7 %	19,2 %
„befriedigend“		14 (10,6 %)	39 (30,0 %)	18 (13,6 %)
		19,7 %	54,9 %	25,4 %
		40,0 %	56,5 %	69,2 %
„ausreichend“	1 (0,8 %)	17 (12,8 %)	20 (15,2 %)	3 (2,3 %)
	2,4 %	41,5 %	48,8 %	7,3 %
	50 %	48,6 %	29,0 %	11,5 %
„mangelhaft“	1 (0,8 %)	2 (1,5 %)	4 (3,0 %)	
	14,3 %	28,6 %	57,1 %	
	50 %	5,7 %	5,8 %	
Tab. 4: Kreuztabelle	2 (100 %)	35 (100 %)	69 (100 %)	26 (100 %)

4.2 Ergebnis der Zwischenprüfung in Notenstufen

Die Kreuztabelle 4 zeigt, daß ein besserer Gesamtempfehlungsgrad tendenziell mit einer besseren Note in der Zwischenprüfung einhergeht. Allerdings ist bereits der Gesamtempfehlungsgrad *„4,5“* ein Garant für gute Leistungen in der Zwischenprüfung, 84,7 % aller Personen, die die Zwischenprüfung mit der Note *„gut“* absolviert haben, können einen Gesamtempfehlungsgrad von *„4,5“* oder *„5“* aufweisen.

Allerdings erzielten noch 5,8 % aller Personen mit dem Gesamtempfehlungsgrad von *„4,5“* die Note *„mangelhaft“* – diese Note kommt unter den Personen mit dem besten Gesamtempfehlungsgrad, nämlich *„5“*, nicht vor. Umgekehrt gibt es keine einzige Person mit dem Gesamtempfehlungsgrad *„3,5“*, die in der Zwischenprüfung eine gute oder auch nur befriedigende Note erzielt hat. Es sei noch einmal daran erinnert, wieviele Hundert Bewerber(innen) im Test oder im gesamten Auswahlverfahren einen Empfehlungsgrad von *„3,5“* oder schlechter erzielt haben, dies waren mehr als die Hälfte aller Bewerber(innen). Ohne das neue, professionelle Auswahlverfahren wären wahrscheinlich etliche dieser Personen zugelassen worden und hätten wahrscheinlich so oder schlechter abgeschnitten wie die beiden hier analysierten Personen mit dem Gesamtempfehlungsgrad *„3,5“*, d.h. die Personen hätten vermutlich entweder die Zwischenprüfung so gerade eben mit *„ausreichend“* bestanden oder wären durchgefallen.

4.3 Korrelativer Zusammenhang

In einem weiteren Analyseschritt wurde der statistische Zusammenhang zwischen dem Testergebnis sowie dem Gesamtergebnis (jeweils Empfehlungsgrad) und dem gerundeten Punktwert der Zwischenprüfung berechnet. Die Analyse beschränkt sich auf die Testergebnisse derjenigen 132 Personen, für die sowohl die Testergebnisse als auch die Gesamtempfehlungsgrade vorliegen. Als Maß für den Zusammenhang zwischen den Indikatoren wird der Korrelationskoeffizient berichtet, der von -1.0 (vollständig negativer Zusammenhang) bis +1.0 (vollständig positiver Zusammenhang) variieren kann. Ein Koeffizient von 0,0 zeigt an, daß kein Zusammenhang zwischen den Indikatoren besteht.

In der vorliegenden Untersuchung betrug die Korrelation zwischen der Testleistung in den Empfehlungsgradstufen und dem Erfolg in der Zwischenprüfung (in gerundeten Punktwerten) $r=.31^{**}$ und konnte mit einer Irrtumswahrscheinlichkeit unter 1 % zufallskritisch abgesichert werden. Der Gesamtempfehlungsgrad war zu $r=.36^{**}$ mit dem in der Zwischenprüfung erzielten Gesamtpunktwert assoziiert. Dies verdeutlicht, daß die Berücksichtigung des mündlichen Teils des Auswahlverfahrens die Aussagekraft nominell steigert. Betrachtet man hingegen nur den Empfehlungsgrad, der allein aufgrund des mündlichen Teils des Auswahlverfahrens vergeben wurde, so war dieser "Empfehlungsgrad Verhalten" lediglich zu $r=.25^{*}$ mit dem Punktwert der Zwischenprüfung korreliert. Der mündliche Teil des Auswahlverfahrens allein stellt also selbst dann keine Alternative zur Kombination aus einer testbasierten und verhaltensbasierten Empfehlung dar, wenn die Kommissionsmitglieder eine aufgrund der Testergebnisse vorausgewählte leistungsstarke Bewerber(innen)gruppe beurteilen. Als Ergänzung zum Test erlangt der Verhaltensteil aber eine eigenständige Bedeutung bei der Vorhersage der Zwischenprüfungsergebnisse. Dies konnte durch eine multiple Regression gezeigt werden. Zum Ende des letzten Schritts, nach Aufnahme der Empfehlungsgrade für den schriftlichen und mündlichen Teil in die Regressionsgleichung, betrug die multiple Korrelation $R=.38$ ($F=11.01$, $p<.001$). Im ersten Schritt wurde die Variable "Empfehlungsgrad Test" aufgenommen, im zweiten Schritt der "Empfehlungsgrad Verhalten". Durch den zusätzlichen Prädiktor stieg das R von .31 auf .38 an ($F_{\text{change}} = 8,49$, $p<.001$). Diese Ergebnisse stehen unter dem Vorbehalt einer ausstehenden Kreuzvalidierung.

Vergleichbares gilt für die Schulnoten, die ebenfalls deutlich mit dem Ergebnis der Zwischenprüfung assoziiert waren. Dies bedeutet, daß *innerhalb einer nach Tests ausgewählten leistungsstarken Gruppe* die Schulnoten eine eigenständige, zusätzliche Informationsquelle für den zukünftigen Ausbildungserfolg enthalten. Den Analysen zufolge ist es sinnvoll, die Schulnote (insbesondere die Note im Fach "Mathematik") zumindest dann als zusätzliches Entscheidungskriterium heranzuziehen, wenn die Wahl zwischen zwei oder mehreren testleistungsstarken Bewerber(inne)n getroffen werden muß, die sich aufgrund der Ergebnisse im schriftlichen und mündlichen Verfahrensteil nicht unterscheiden. Auf diese Art und Weise könnte der sicherlich auch in Schulnoten enthaltene prognostisch (für den Ausbildungserfolg prognostisch) verwertbare Anteil über die ja bereits nach Schulnoten durchgeführte Vorauswahl hinaus genutzt werden (siehe z.B. die

Ergebnisse der Metaanalyse von Baron-Boldt, Schuler und Funke, 1988). Die Angemessenheit eines solchen Vorgehens wäre allerdings von Zeit zu Zeit zu überprüfen, da sich in der zitierten Studie von Baron-Boldt et al. (1988) gezeigt hatte, daß Schulnoten aus jüngerer Zeit weniger aussagekräftig sind als Schulnoten, die vor längerer Zeit abgegeben wurden.

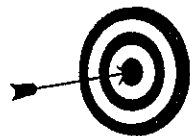
Wie in Abschnitt 3.1.1 dargestellt, wird die tatsächliche Aussagekraft des Auswahlverfahrens durch die berichteten Korrelationskoeffizienten deutlich unterschätzt, da die im Test schlechten Bewerber(innen) nicht zugelassen wurden und die Streuung der Testwerte in der analysierten Gruppe daher künstlich eingeschränkt ist. Da für alle testbasierten Werte die ursprüngliche Varianz aus der Eichstichprobe bekannt ist, läßt sich für die hier analysierte Gruppe mit einer eingeschränkten Varianz eine Korrektur des Validitätskoeffizienten entsprechend den empirisch vorliegenden Varianzverhältnissen vornehmen (siehe Formel 13.29 [correlation corrected for restriction of range, case 1] Guilford, 1950, S. 349). Die entsprechende korrigierte Validität der Testempfehlungsgrade beträgt $r=.76$.

Bezieht man auch die drei Personen mit ein, für die zwar die Testergebnisse, nicht aber die Ergebnisse des mündlichen Verfahrensteils vorliegen, so betrug bei diesen 135 Personen die prognostische Validität des testbasierten Empfehlungsgrades $r=.29$ bzw. $r=.74$ nach Korrektur (siehe Tabelle 5).

Im letzten Analyseschritt wurden die Korrelationen zwischen den einzelnen Testskalen und dem Zwischenprüfungsergebnis (gerundete Punktwerte) ermittelt, die Ergebnisse sind in Tabelle 5 verzeichnet und beziehen sich auf die Gesamtgruppe im Umfang von 135 Personen.

Insbesondere die Dimensionen "Verarbeitungskapazität" und "numerisch" erwiesen sich als prognostisch bedeutsam. Innerhalb der geprüften Wissensdomänen war das Wirtschaftswissen der nominell beste Prädiktor. Diesen genannten Skalen wurden bereits als Ergebnis eines Expert(inn)enratings bei der Auswahlentscheidung ein besonders hohes Gewicht zuerkannt. Die Ergebnisse bestätigen die Ansicht der Expert(inn)en und ermöglichen es, die Aussagekraft des Tests in Zukunft noch gezielter auf die hier in Frage stehende Ausbildung zuzuschneiden. Leistungsdimensionen, die in keinem bedeutsamen Zusammenhang mit dem Ausbildungserfolg stehen, können aufgrund des bimodalen Testkonzepts gleichwohl zur Vorhersage beitragen. So gehen in den Wert für die Skala Verarbeitungskapazität beispielsweise auch die entsprechenden Aufgaben mit figuralem Material mit ein.

Tab. 5:
Ergebnis der Zwischenprüfung: (Gesamtergebnis in
gerundeten Punktwerten) Vorhersage durch...
(N = 135)



O P E R A T I O N E N	Verarbeitungskapazität	.38**	(.54)
	Einfallsmenge	.02	
	Merkfähigkeit	.00	
	Bearbeitungsgeschwindigkeit	-.04	
I N H A L T E	verbal	.07	(.12)
	numerisch	.33**	(.44)
	figural	-.01	
F E R T I G K E I T E N	Arbeitseffizienz (Postaufgabe)	.16	(.19)
	Rechtschreibung (Diktat)	.11	(.13)
W I S S E N	Kenntnisse (Mittelwert)	.26*	(.32)
	Gemeinschaftskunde	.22	(.28)
	Literatur	.06	
	Wirtschaft	.29**	(.30)
	Politik	.19	(.24)
Σ	Allgemeine Intelligenz	.18	(.34)
	Testgesamtpunktwert (anforderungsspezifisch gewichtet)	.37**	(.63)
	Empfehlungsgrad Test	.29**	(.74)

(in Klammern: corrected for restriction of range) ↑

5. Zusammenfassung

Die vorliegende Untersuchung bestätigt die Treffsicherheit und den Wert der DGP-Eignungsuntersuchungen durch die folgenden fünf Hauptbefunde:

- Von den 26 Personen, die im Auswahlverfahren als "den Anforderungen entsprechend" (Gesamtempfehlungsgrad "5") beurteilt werden, gab es **keine**, die den Anforderungen der Zwischenprüfung entgegen der Aussage des Auswahlverfahrens nicht genügte.
- 84,7 % aller Personen, die die Zwischenprüfung mit der Note "gut" absolviert haben, hatten im Auswahlverfahren einen Gesamtempfehlungsgrad von "4,5" oder "5" erzielt.
- Es ist wenig erfolgversprechend, Bewerber(innen), die von der DGP nicht empfohlen werden konnten, zur Ausbildung zuzulassen. Es gab keine einzige Person mit dem relativ "schlechten" Gesamtempfehlungsgrad "3,5", die in der Zwischenprüfung eine gute oder auch nur befriedigende Note erzielt hat.
- Von den beiden Personen mit dem schlechtesten Gesamtempfehlungsgrad der vorliegenden Untersuchungsgruppe ("3,5") hat eine Person die Zwischenprüfung nur äußerst knapp (mit 6,87 Punkten) bestanden, die andere Person ist "durchgefallen". Da es mehrere hundert Bewerber(innen) mit einem Gesamtempfehlungsgrad von "3,5" oder schlechter gab, ist davon auszugehen, daß durch das neue Auswahlverfahren Fehlinvestitionen in Millionenhöhe vermieden werden konnten.
- Trotz der extremen Vorauswahl nach Testleistungen und der damit verbundenen Streuungseinschränkung betrug die zufallskritisch gesicherte Korrelation zwischen dem Gesamtempfehlungsgrad und dem Erfolg in der Zwischenprüfung (in gerundeten Punktwerten) $r = .36^{**}$. Dieser Zusammenhang muß angesichts der extremen Streuungseinschränkung als überaus zufriedenstellend und substantiell eingestuft werden, der für Streuungseinschränkungen korrigierte Vorhersagekoeffizient für die testbasierten Empfehlungsgrade betrug $r = .76^{**}$, $n = 132$, siehe Seite 43.

Eine Bewährungsprobe wie die hier vorgelegte darf kein singuläres Ereignis sein, sondern muß als ein Baustein einer fortlaufenden Qualitätssicherung und Optimierung der Personalauswahl angesehen werden. Zum Zeitpunkt der Berichtserstellung gingen bei der DGP die Ergebnisse der Zwischenprüfung des folgenden Studienjahrgangs ein, so daß die hier vorliegende Studie in Kürze fortgesetzt und ergänzt werden kann. Die Ergebnisse der Evaluation erlauben Rückschlüsse über die Anforderungen im Studium. Besondere prognostische Bedeutung in bezug auf den Erfolg in der Zwischenprüfung kam in der vorliegenden Untersuchung den Tests zur Verarbeitungskapazität, zum zahlengebundenen Denken sowie der Überprüfung der Wirtschaftskennntnisse zu. Als Ergebnis der vorliegenden Studie kann diesen Dimensionen in Zukunft bei der Auswahlentscheidung ein noch höheres Gewicht zuerkannt werden. Dadurch dürfte sich die Qualität des Auswahlverfahrens noch weiter steigern lassen.

6. Literatur

- Althoff, K. (1984). Zur prognostischen Validität von Intelligenz- und Leistungstests im Rahmen der Eignungsdiagnostik. *Psychologie und Praxis - Zeitschrift für Arbeits- und Organisationspsychologie*, 28, 144-148.
- Baron-Boldt, J., Schuler, H. & Funke, U. (1988). Prädiktive Validität von Schulabschlußnoten: Eine Metaanalyse. *Zeitschrift für pädagogische Psychologie*, 2, 79-90.
- Barthel, E. (1988). Nutzen eignungsdiagnostischer Verfahren bei der Bewerberauswahl. Frankfurt /M.: Lang.
- Guilford, J.P. (1950). *Fundamental statistics in psychology and education*. New York: McGrawHill.
- Hunter, J.E. & Hunter, R.F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, 96, 72-98.
- 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.
- Jäger, A.O., Süß, H.-M. & Beauducel, A. (1997). *Berliner Intelligenzstruktur-Test. Form 4. Handanweisung*. Göttingen: Hogrefe.
- Kersting, M. (1995a). Der Einsatz "westdeutscher" Tests zur Personalauswahl in den neuen Bundesländern und die Fairneßfrage. Auswirkungen der Testleistungsdisparität zwischen Ost und West auf die Auswahlentscheidung. *Report Psychologie*, 20, 32-41.
- Kersting, M. (1995b). Entwicklung eines neuen Leistungstests zur Personalauswahl für Bewerber(innen) für die Laufbahn des gehobenen Dienstes. *DGP-Informationen*. Hannover: Deutsche Gesellschaft für Personalwesen, 54, 3-20.
- Kersting, M. (1996). Ost-West-Leistungsunterschiede in Berufseignungstests in Abhängigkeit von der kulturspezifischen Wirkung einiger Aufgabenmerkmale. *Zeitschrift für Arbeits- und Organisationspsychologie*, 40, 106-117.
- Kersting, M. (1998). Differentielle Aspekte der sozialen Akzeptanz von Intelligenztests und Problemlöseszenarien als Personalauswahlverfahren. *Zeitschrift für Arbeits- und Organisationspsychologie*, 42, 61-75.
- Kersting, M. & Beauducel, A. (1997). Der neue DGP-Leistungstest auf der Basis des Berliner Intelligenzstrukturmodells: Informationen zu ausgewählten Testgütekriterien und zur Normierung. *DGP-Informationen*. Hannover: Deutsche Gesellschaft für Personalwesen, 55, 93-103.
- Kleinmann, M. & Strauß, B. (Hrsg.) (1998). *Potentialfeststellung und Personalentwicklung*. Göttingen: Verlag für Angewandte Psychologie.
- Schmidt, F.L. & Hunter, J.E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, 124, 262-274.
- Schuler, H. (1990). Personalauswahl aus der Sicht der Bewerber: Zum Erleben eignungsdiagnostischer Situationen. *Zeitschrift für Arbeits- und Organisationspsychologie*, 34, 184-191.
- Schuler, H. (1993). Social validity of selection situations: a concept and some empirical results. In H. Schuler, J.L. Farr & M. Smith (Eds.), *Personnel selection and assessment. Individual and organizational perspectives* (pp. 1126). Hillsdale, NJ: Erlbaum



Deutsche Gesellschaft für Personalwesen e.V.

DGP INFORMATIONEN

47. Jahrgang, Heft 56, September 1999

Inhalt	Seite
Vorwort	1
Personalentwicklung (PE) in den neuen Bundesländern - Qualifizierungsreihe zum/zur Referent/in für Personalentwicklung <i>Joachim Rieke, Carmen Knorz, Kerstin Lange</i>	3
Potential-Assessment Center zur Bildung eines "Förderkreises Führungsnachwuchs" in einer Kreisverwaltung <i>Bernd Wolf</i>	17
Korruptionsvorsorge aus Fürsorge <i>Ina Voigt</i>	31
Konfliktberater/in bei der Polizei: Von der Stellenausschreibung zur Personalauswahl <i>Volker Klein-Moddenborg</i>	37
Personalauswahl mit dem BIS-r-DGP Test: Evaluation der prognostischen Validität <i>Martin Kersting</i>	43
Aus der (wechselvollen) Geschichte der Deutschen Gesellschaft für Personalwesen e.V. (DGP) <i>Klaus Althoff</i>	61

ISSN 0948-0579