„Beurteilung eines Personalauswahlverfahrens unter besonderer Berücksichtigung der prognostischen Validität“

Dissertation
zur Erlangung des Doktorgrades
der Philosophischen Fakultät
der Christian-Albrechts-Universität
zu Kiel
vorgelegt von
Isabel Petersen

Kiel
2004

Erstgutachter: Prof. Dr. Günter Köhnken
Zweitgutachter: Prof. Dr. Thomas Bliesener

Tag der mündlichen Prüfung:01.12.2004
Durch Prodekan, Prof. Dr. Norbert Nübler
zum Druck genehmigt am: 18.10.2005

Vorwort

Als ich mich 1998 entschloss, meinen geplanten Erziehungsurlaub für eine weitergehende Qualifizierung zu nutzen, ahnte ich nicht, welche Folgen dies für mich haben würde. Damals standen das Interesse am Thema und die wahrgenommene Relevanz der vorliegenden Untersuchung für die praktische Eignungsdiagnostik im Vordergrund. Der Alltag mit meinen Kindern hat mich jedoch schnell eingeholt und ich habe wiederholt daran gezweifelt, ob ich die Arbeit wirklich erfolgreich werde beenden können. Insofern gebührt der erste und ganz besondere Dank meinem Sohn Leander Arthur, geb. 1999, und meiner Tochter Greta Maleen, geb. 2002. Beide haben nicht immer verstehen können, warum „Mama“ so viele Stunden nicht mit ihnen spielen konnte, es aber mit der ihnen größtmöglichen Rücksicht ertragen müssen. Weiter danke ich meinem Mann, Herrn Dr. Rainer Petersen, der mich bei der Vorbereitung des Themas durch seine damalige berufliche Tätigkeit hilfreich unterstützte und mir immer wieder den Rücken stärkte, mich motivierte und niemals den Glauben an den Abschluss verlor.

Ganz besonders herzlich möchte ich mich bei Herrn Prof. Dr. Köhnken bedanken. Trotz zweimaliger Unterbrechung der Arbeit durch die Geburten meiner Kinder war Herr Prof. Köhnken als Betreuer und Erstgutachter stets für mich erreichbar, unterstützend und fördernd. Seine aufmunternden Worte und seine fachliche Unterstützung machten den Abschluss der Arbeit erst möglich. Herzlicher Dank gilt ferner Herrn Prof. Dr. Bliesener für seine Unterstützung als Zweitgutachter.

Nicht zuletzt möchte ich all jenen meinen aufrichtigen Dank aussprechen, die mich bei der Durchführung der Untersuchung und der Datenerhebung halfen. Hier sind insbesondere die Mitarbeiter der Polizei Hamburg zu erwähnen, die das Zustandekommen der Arbeit hilfreich begleiteten und mir im Rahmen einer Hospitation erlaubten, die Grundlage für die Arbeit zu erheben.

Inhaltsverzeichnis

1. Einleitung

Nach Ludwig (1996) gehört der Beruf des Polizeibeamten[1] zu den sehr belastenden Berufen. Dies ist insbesondere auf die Vielfalt der Aufgaben und daraus resultierender Tätigkeiten im Rahmen des Polizeidienstes zurückzuführen. Zum einen ist an dieser Stelle insbesondere die Gefahrenabwehr durch allgemeine Präsenz in der Bevölkerung, Schutzmaßnahmen, etwa für gefährdete Objekte und Bürger, oder Eingriffsmaßnahmen bei der Durchsetzung von Straf- oder Haftbefehlen zu erwähnen. Weiter werden im Rahmen der Verbrechensbekämpfung von den Beamten unterschiedliche Aufgaben bei der Fahndung nach oder Vernehmung von Straftätern erwartet. Darüber hinaus sind diverse unterstützende Funktionen für andere Behörden (wie z.B. Ordnungsbehörde oder Jugendamt) Bestandteil der täglichen Arbeit. Sogar die tagtägliche Auseinandersetzung mit dem Bürger kann bei diesen auf Unverständnis stoßen und zu Frustrationen und teilweise auch Aggressivität führen. Dabei wünschen sich die meisten Bürger einen freundlichen Polizeibeamten, der korrekt und freundlich hilft, wann immer er benötigt wird.

Ein Polizeibeamter sollte grundsätzlich im Umgang mit anderen Menschen ein angemessenes Handeln zeigen, welches ohne ausreichende "Soziale Kompetenz" kaum erreichbar ist (Penning, 1996). Hierzu zählen die Fähigkeiten, sich auf Empfindungen anderer einzulassen und in Anspannungs- und Konfliktsituationen die gegebenen Fertigkeiten auch effektiv einsetzen zu können. Weiter bedarf es ausreichender "Persönlicher Kompetenz" für die Bewältigung von besonderen Anforderungssituationen wie schwierigen Mitarbeitergesprächen, Stresssituationen, Überbringen unangenehmer Nachrichten etc.. Von den Beamten wird erwartet, dass sie über undogmatische Vorstellungen, angemessene Bewertungen von Negativsituationen, ein differenziertes Selbstwertkonzept und eine hohe Frustrationstoleranz verfügen.
Die genannten Aufgaben machen deutlich, wie verantwortungsvoll und vielfältig die alltägliche Arbeit sich gestaltet und wie schwierig es sein muss, den vielseitigen Grundanforderungen des
Berufes Polizeibeamter zu entsprechen. Ein Polizist muss überaus lernfähig sein, da er sich immer wieder auf neue Situationen einstellen können muss. Er muss Techniken und Kenntnisse

[1] In der vorliegenden Arbeit wurde zur Vereinfachung die männliche Form verwendet. Unabhängig hiervon sind selbstverständlich Personen beiderlei Geschlechts gemeint.

aus unterschiedlichen Wissensgebieten beherrschen (Kriminologie, Einsatzlehre, Sport, Psychologie, Recht etc.), um vor Ort beim Einsatz die richtige Entscheidung treffen zu können. Er sollte sich auf neue Techniken einstellen und mit ihnen umgehen können (Stichwort Internetkriminalität). In schwierigen Situationen darf er sich nicht provozieren lassen, sondern sollte in der Lage sein, deeskalierend und besonnen zu reagieren. Und dies, obwohl insbesondere in Großstädten, der regelmäßige Umgang mit sozialen Problem- und Randgruppen zermürbend ist und manchmal kaum leistbar scheint. Den Auswirkungen von Schichtdienst und traumatischen Erlebnissen sollte er stressresistent begegnen können. Er muss sportlich sein und darüber hinaus in der Lage, sich schnell unterschiedlichen Gegebenheiten anzupassen. Er wird mit Situation konfrontiert, auf deren Bewältigung er nicht vorbereitet ist. So gehört bspw. das Überbringen von Todesnachrichten zu den belastendsten Tätigkeiten, die Polizeibeamte ausüben müssen, die massiven Stress erzeugen und zu erheblichen Verhaltensunsicherheiten führen (Huber, 1996).

Wo finden sich Menschen, die diesen Anforderungen gerecht werden können und die in der Regel bereit sind, für häufig wenig Anerkennung in der Gesellschaft und keine außergewöhnliche Bezahlung zu arbeiten? Die Antwort ist recht einfach, denn die hohen Anzahlen von Bewerbungen sprechen dafür, dass sich an den deutschen Schulen, trotz einer Vielzahl interessanter Alternativen auf dem Berufsmarkt, immer noch ausreichend junge Menschen für den Beruf des Polizeibeamten interessieren. Es herrschte in den letzten Jahren kein Mangel an Bewerbern und es konnten kaum alle Interessierten zu einem Eignungsfeststellungsverfahren eingeladen werden, da dies die organisatorischen Möglichkeiten nicht zuließen. Doch werden immer die „richtigen" jungen Leute zum Testverfahren eingeladen und erhalten nur jene eine Ablehnung, die im Vergleich zu anderen weniger oder keine Chancen auf eine erfolgreiche Berufsausübung haben? Wie kann oder besser wie sollte ein Eignungsfeststellungsverfahren aussehen, welches dieser großen Anzahl von Anforderungen gerecht werden kann? Gelingt es mit dem bestehenden Verfahren jene Bewerber auszuwählen, die diesen Anforderungen gerecht werden und letztendlich die „guten Polizeibeamten" werden, die wir Bürger uns immer wünschen?

Bei der Polizei werden im Rahmen der Auswahl geeigneter Bewerber für den Polizeivollzugsdienst regelmäßig Entscheidungen über Annahme und Ablehnung getroffen. Inwieweit die Leistungsmöglichkeiten der Bewerber mit diesen vielfältigen Anforderungen des speziellen Berufes jedoch übereinstimmen ist vermutlich kaum einmal untersucht worden.

Dabei ist bekannt, dass je besser die tätigkeitsrelevanten Merkmale eines Arbeitsplatzes mit den fähigkeitsrelevanten Merkmalen einer Person zusammenpassen, desto höher ist letztlich auch der Nutzen für die Organisation und somit schließlich für den Bürger zu dessen Wohle und Schutz der Beamte eingesetzt wird. Es gilt somit, die beiden Seiten „Arbeitsplatz“ und „Person“ in Übereinstimmung zu bringen und diejenigen Merkmale, die zum erfolgreichen Ausüben der Tätigkeit führen, valide durch entsprechende Auswahlverfahren zu erfassen.

Heutzutage steht eine erhebliche Zahl unterschiedlicher Verfahren zur Auswahl von Mitarbeitern zur Verfügung, die jedoch in ihrer Qualität erheblich variieren. Mit Hilfe testtheoretischer Gütekriterien und Methoden der Evaluation ist es möglich, brauchbare Verfahren der Personalauswahl von solchen zu unterscheiden, für die lediglich eine Behauptung oder eigene Intuition spricht. Es war im Rahmen der Untersuchung nicht möglich und auch nicht primäres Ziel, den vielfältigen Aufgaben der Tätigkeit gerecht zu werden. Zweck dieser Arbeit war es, die prognostische und inkrementelle Validität des Auswahlverfahrens für die Einstellung in den gehobenen Polizeivollzugsdienst, d.h. für ein Studium an der Fachhochschule für Öffentliche Verwaltung – Fachbereich Polizei – zu erschließen. Anhand der Höhe der prognostischen Validität lässt sich z.B. eine Aussage darüber machen, inwieweit anhand der verwendeten Prädiktoren (Bestandteile des Auswahlverfahrens) eine Vorhersage über das erfolgreiche Verhalten im Beruf (Kriterium) möglich ist. Da sich eine Beurteilung eines Auswahlverfahrens ebenso an Kosten-Nutzen Aspekten orientieren muss, sollte sichergestellt werden, dass es keine anderen Verfahren gibt, die mit geringerem Aufwand an Zeit und Kosten eine gleich hohe Güte in der Vorhersage erzielen können. In diesem Zusammenhang sollte auch die kontrovers diskutierte Bedeutung von Schulnoten berücksichtigt werden, da in verschiedenen Bundesländern Noten im Rahmen der Vorauswahl für die Einstellung in den Polizeivollzugsdienst eingesetzt werden. Weiter sollten auch Kosten und Nutzen Aspekte des Auswahlverfahrens beleuchtet und die Frage beantwortet werden, ob das vorliegende Verfahren dazu beitragen kann, neben deutschen, auch vermehrt ausländische Bewerber für den Polizeivollzugsdienst zu gewinnen.

2. Theoretischer und empirischer Hintergrund

2.1 Psychologische Personalauswahl

2.1.1 Grundlagen und Kontext der psychologischen Personalauswahl

Der erste Abschnitt des zweiten Kapitels widmet sich dem geschichtlichen Abriss der Entwicklung in dem Themengebiet der psychologischen Personalauswahl. Anschließend werden die Begriffe der Berufseignungsdiagnostik und der Eignung kurz anhand von Definitionen einiger Autoren zu erläutern versucht. Bei der Frage, inwieweit Personen und Arbeitsplatz miteinander effektiv und ökonomisch miteinander verbunden werden können, werden die Begriffe der Selektion und Platzierung von geeigneten Personen angesprochen. Anhand einer Darstellung von Rösler (1992) folgen schließlich beispielhaft wissenschaftliche Grundsätze zur Entwicklung von Verfahren zur Selektion geeigneter Personen. Es werden Begriffe wie Arbeitsplatzanalyse, Prädiktoren, Kriterien, Validitätsanalyse und Entscheidungsbildung angesprochen.

In der heutigen Zeit werden tagtäglich Personalentscheidungen getroffen. Die Relevanz berufseignungsdiagnostischer Entscheidungen ist von nahezu herausragender Bedeutung. Sie legen nicht nur den weiteren beruflichen Weg für einen potentiellen Bewerber um einen Arbeitsplatz fest, sondern bestimmen ebenso den Handlungsspielraum des betroffenen Unternehmens. Jede Organisation kann nur so gut sein, wie ihre ausgewählten Mitarbeiter dies zu lassen. Letztendlich sind sie es, welche die Produkte herstellen, die Organisation und Planungen übernehmen etc.. Der Kernpunkt jeglicher Personalentscheidung liegt darin, anhand effizienter Auswahlverfahren die individuellen Fähigkeiten und Fertigkeiten möglichst objektiv zu erfassen und im weiteren durch eine möglichst optimale Passung zwischen den Merkmalen einzelner Mitarbeiter und den Anforderungen des jeweiligen Arbeitsplatzes den größtmöglichen Nutzen für die Organisation auf der einen und die Mitarbeiter auf der anderen Seite zu erzielen.

Seit mehr als 100 Jahren beschäftigt sich die Psychologie mit der Frage nach der Messung des individuellen Verhaltens und dessen Umsetzung in objektivierbare und somit quantifizierbare Merkmale. Der Gedanke, mittels systematischer und kontrollierter Methoden Menschen aufgrund ihrer Fähigkeiten auszuwählen, reicht, basierend auf unterschiedliche Quellen, auf bis vor mehr als 3000 Jahre zurück. Nach DuBois (1970) wurden bereits zu dieser Zeit in

China öffentlich Bedienstete mittels einer „Testbatterie“ hinsichtlich ihrer Eignung für „Managementaufgaben“ im Staatsdienst geprüft, die u.a. aus Reiten, Bogenschiessen und Arithmetik bestand. Schuler (1996) zitiert ein weiteres Beispiel von Lavater (1775), welcher die Meinung vertrat, dass charakteristische physiognomische Merkmale eines Menschen, wie z.B. Körperform, Haar, Stimme, Sprache, Handschrift und Kleidung, äußert valide eignungsdiagnostische Indikatoren seien. Eine Systematisierung der Ausleseuntersuchungen erfolgte jedoch erst im Laufe des letzten Jahrhunderts (Hossiep, 1995). Ziel war hierbei, jeden Einzelnen gemäß seiner Eignung und einem Leistungsprinzip für die jeweilige Funktion auszuwählen.

Somit ist der Beginn der von psychologischen Theorien und Methoden geprägten Diagnostik auf den Anfang des 20. Jahrhunderts zu datieren. Zu dieser Zeit wurden Papier- und Bleistifttests, apparative Verfahren sowie Arbeitsproben, welche nach psychometrischen Prinzipien konstruiert und teilweise sogar bereits evaluiert wurden, für die schwerpunktmäßige Anwendung im Bereich der Produktions- und Dienstleistungsberufe eingesetzt. Die Eignungsprüfung mittels des Verfahrens der Arbeitsprobe entwickelte sich beispielsweise im Zuge der fortschreitenden arbeitsteiligen Produktionsweise in der Industrie. Die zahlreichen psychotechnischen Forschungs- und Prüfstellen versuchten mittels dieser Methode diejenigen zentralen psychischen Funktionen herauszufinden und zu messen, welche für eine bestmögliche Leistung und somit Ergebnis der Arbeit notwendig erschienen. Der Schwerpunkt der Anwendungen der genannten Verfahren lag hauptsächlich im Bereich der Produktions- und Dienstleistungsberufe, wobei eine Auswahl von Führungspositionen noch kaum stattfand (Schuler, 1996).

Schon zu Beginn des 20. Jahrhunderts wurden heute noch bekannte und gebräuchliche Verfahrenstypen eingesetzt. Anastasi (1985) verweist auf die Verwendung von Persönlichkeitstests zur Auswahl von Verkäufern in den USA, Owens (1976) auf die Anwendung zur Auswahl von Versicherungsvertretern. Außerdem begann in dieser Zeit die Anwendung von Intelligenztests, welche zunächst in Form von Schulreifetests und später erstmalig zur Selektion von amerikanischen Rekruten (Army-Alpha-Test) eingesetzt wurden. Zur Eignungsprüfung der Mitarbeiter wurden außerdem Einstellungsinterviews verwendet (Schuler, 1996).

Als Konglomerat aus Testverfahren, Einstellungsinterview und Arbeitsproben entwickelte sich das heute so bezeichnete und für die Auswahl von Führungskräften nicht mehr wegzudenkende „Assessment Center“, das auch als Vorläufer während der Weimarer

Republik zur Auswahl von Offizieren der Reichswehr eingesetzt wurde (Schuler, 1996). Eine differenziertere Betrachtung einzelner Auswahlinstrumente erfolgt in Kapitel 2.3.

Im Zuge der Verbreitung der Computer gestützte Datenverarbeitung entwickelte sich gegen Ende des letzten Jahrhunderts die computerunterstützte Eignungsdiagnostik, d.h. Auswahlverfahren werden mit Hilfe des Computers durchgeführt. Die Items werden standardisiert vorgegeben, ausgewertet und letztendlich Personalentscheidungen auf dieser Datenbasis getroffen. Schnell waren hinsichtlich der Durchführungsbedingungen ökonomische Vorteile erkennbar, zudem ließ sich eine Welt neuer Simulationsaufgaben für die Problemlöseforschung (vgl. Dörner, Kreuzig, Reither & Stäudel, 1983) erschließen. Darüber hinaus kam es durch die Möglichkeiten des adaptiven Testens auch zu einem Paradigmenwechsel in der Eignungsdiagnostik. So kann verhindert werden, dass ein Bewerber eine zuvor festgelegte Reihenfolge von vorgegebenen Aufgaben abarbeiten muss. Hier übernimmt der Computer die Aufgabe, optimale an den jeweiligen Probanden angepasste Untersuchungsschritte anzubieten (Hornke, 1991).

Angesicht der langen Tradition und geschichtlichen Wurzeln sowie der Weiterentwicklungen auf diesem Gebiet, erhebt sich die Frage, was heute unter berufseignungsdiagnostischen Entscheidungen bzw. Berufseignungsdiagnostik verstanden wird. Während der letzten Jahrzehnte haben viele Autoren versucht, geprägt von den sie beeinflussenden Werten und gesellschaftlichen Rahmenbedingungen, eine Definition des Begriffes abzugeben. Hossiep (1995) hat verschiedene definitorische Aussagen, wie er es formuliert, gesammelt und dargestellt, von denen exemplarisch zwei genannt sein sollen: Eckardt und Schuler (1992) verstehen unter Berufseignungsdiagnostik „…die Gesamtheit aller wissenschaftlichen und wissenschaftsgeleiteten praktischen Bemühungen …, die auf dem Wege über eine gedankliche Zuordnung von beruflichen Situationen zu Personen oder von Personen zu beruflichen Situationen die Ziele „Maximierung beruflicher Leistung“ und „Maximierung beruflicher Zufriedenheit“ anstreben. Schuler und Funke (1993) verstehen unter dem Konzept der Eignung, das Ausmaß der Übereinstimmung an Anforderungen des Arbeitsplatzes und der weiteren Arbeitsumgebung mit den Leistungsvoraussetzungen der Person.

Der Begriff der Eignung ist ein sehr facettenreicher und im Bedeutungsgehalt kontextabhängiger Begriff, enthält jedoch zwei wesentliche Zielkriterien: die berufliche Leistung und die berufliche Zufriedenheit einer Person. Es geht nicht nur um das bloße

Zusammenpassen zwischen Person und beruflicher Umwelt, sondern um ein transaktionales Verhältnis. Berufliche Zufriedenheit kann bspw. umgesetzt werden, indem der Arbeitnehmer in der Ausübung seiner Tätigkeit die Möglichkeit hat, bestimmte persönliche Bedürfnisse, wie z.B. nach Kontakten, zu befriedigen (Schuler, 1996). Somit erhöhen sich auch die Arbeitszufriedenheit und die Wahrscheinlichkeit, dass der Arbeitnehmer in der Organisation verbleibt. Über- oder Unterforderung, Nacht- und Schichtarbeiten, Arbeitsumgebungsfaktoren wie z.B. Lärm, Hitze, soziale Belastungen durch Vorgesetze und / oder Kollegen (Rosenstiel, 2001) gelten u.a. als Stressoren, die die psychische und physische Gesundheit beeinträchtigen und schädigen können (Semmer & Udris, 1995). Nach dem Report „Wettbewerbsfaktor Work-Life-Balance“ des wissenschaftlichen Institutes der AOK in Zusammenarbeit mit der Universität Bielefeld (2003) ist die Zahl der auf psychische Erkrankungen zurückgehenden Krankmeldungen seit 1994 um 74,4 % gestiegen.
90 % der Beschäftigungsverhältnisse scheitern nicht aufgrund fehlender fachlicher Kompetenzen des betroffenen Mitarbeiters, sondern wegen gegebener Unstimmigkeiten zwischen den Persönlichkeitsmerkmalen des Stelleninhabers und den Anforderungen der Position (Sarges, 2000). Das von Eckardt und Schuler (1992) genannte Zielkriterium der beruflichen Zufriedenheit scheint somit im Rahmen einer effizienten Personauswahl von entscheidender Bedeutung zu sein, ansonsten sind finanzielle Einbußen in Folge psychischer und physischer Fehlbelastungen möglich. Ein effektives Personalmarketing minimiert derartige finanzielle Einbußen auf Seiten der Organisation durch vorzeitiges Ausscheiden nicht geeigneter Arbeitnehmer und maximiert somit den Erfolg der Organisation.

Die berufliche Leistung wird in den meisten Fällen mittels eignungsdiagnostischer Verfahren gemessen (vgl. auch Kapitel 2.3). Die Relevanz eignungsdiagnostischer Entscheidungen ist immer dann hervorzuheben, wenn es einen merklichen Unterschied macht, welche Person für eine bestimmte Position ausgewählt wird, d.h. bestimmte Fähigkeiten von einem Bewerber erwartet werden müssen, um den Anforderungen der zu besetzenden Stelle gerecht werden zu können. Hier gibt es eine Vielzahl unterschiedlicher Instrumente, die zur Erfassung der Kompetenzen potentieller Bewerber zur Verfügung stehen. Neben Persönlichkeitstests, die das für einen Bewerber typische Verhalten erfassen sollen, stehen Leistungstests wie Konzentrations- oder Intelligenztests oben auf der Beliebtheitsskala von Praktikern, mittels denen maximales Verhalten, d.h. Leistungsfähigkeit erfasst werden soll. Darüber hinaus finden weitere Verfahren und Datenquellen wie Interviews, Arbeitsproben und Beurteilungen Dritter Anwendung. Der auswählenden Institution geht es folglich darum, einen Bewerber zu

finden, der die Anforderungen des Arbeitsplatzes möglichst optimal bewältigen und somit einen möglichst hohen Anteil zum Erfolg eines Unternehmens beitragen kann. Die Ergiebigkeit und Effizienz eines Unternehmens stehen jedoch in Abhängigkeit von der Produktivität, der Zufriedenheit und Belastung eines Arbeitnehmers. Je höher demnach die Passung zwischen den Leistungsmöglichkeiten eines Individuums und den Leistungsanforderungen eines Arbeitsplatzes, desto höher sind Ertrag und Erfolg einer Organisation sowie die Zufriedenheit und Kapazität des Arbeitnehmers.
Nach Rösler (1992) lassen sich zwei Wege einschlagen, wie Personen und Arbeitsplatz miteinander verbunden werden können: einerseits durch aktive Arbeitsplatzgestaltung und andererseits durch die Auswahl und Platzierung geeigneter Personen. Eine Maßnahme, Anforderungen an eine Person durch eine Tätigkeit einerseits und Leistungskapazitäten einer Person andererseits miteinander abzustimmen, ist die Arbeitsplatzgestaltung. Theoretische Grundlage dieses Vorgehens ist die Begrenztheit menschlicher Fähigkeiten als gegeben anzunehmen und die Bedingungen des Arbeitsplatzes auf diese abzustimmen. Mittels der Bereitstellung spezifischer Hilfsmittel sollen Leistungsdefizite aufgefangen und kompensiert werden.

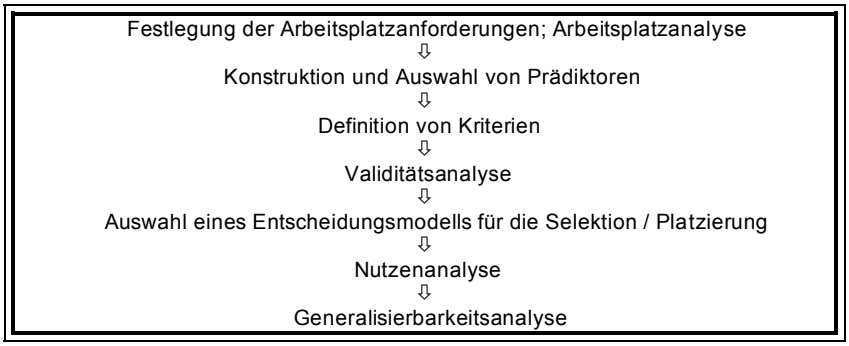
Arbeitsplatz und Person können ebenfalls anhand der Methode der Platzierung umgesetzt werden. In diesem Falle soll das vorhandene Leistungspotential eines Arbeitnehmers mittels Fort- und Weiterbildung gesteigert, gefördert und somit Fähigkeitsdefizite ausgeglichen werden. Im Rahmen der Eignungsdiagnostik stehen hier Verfahren zur Potentialfeststellung von Mitarbeitern zur Verfügung, die häufig jenen zur Personalauswahl gleichen (Lang von Wins & Rosenstiel, 1998). Die hier gemeinte klassische Personalentwicklung findet in erster Linie durch Lernen „ on the job“ (Sonntag, 1998), d.h. durch Maßnahmen, die unmittelbar am Arbeitsplatz während Ausübung der Tätigkeit stattfinden und durch Maßnahmen „off the job“ (Lammers, 1998), also im Rahmen einer traditionellen Weiterbildung.

Bei der Auswahl von Personen geht es darum, die Anforderungen eines Arbeitsplatzes als gegeben hinzunehmen und jene Personen auszuwählen, die den Anforderungen im weiteren Sinne entsprechen. Theoretische Grundlage dieses Ansatzes ist die These interindividueller Unterschiede: Jeder Mensch besitzt eine Palette unterschiedlicher Fähigkeiten und Fertigkeiten, so dass unterschiedliche Menschen ein und dieselbe Arbeit auch unterschiedlich gut ausführen. Je nachdem welche Fähigkeiten an einem Arbeitsplatz gefordert werden, ergibt sich für die Eignung eines Bewerbers auch ein unterschiedliches Fähigkeitsprofil. Mertens

(1974) hat in diesem Sinne den Begriff der Schlüsselqualifikation eingeführt. Gemeint sind damit Merkmale einer Person, die zur Bewältigung gegenwärtiger und zukünftiger beruflicher Anforderungen bedeutsam sind. Diese Merkmale bzw. das vorhandene Fähigkeitsprofil zu erfassen und mittels geeigneter Verfahren in Eignung oder Nichteignung umzusetzen, ist Ziel einer effizienten Personalsauswahl. Die Konzeption eines Auswahlverfahrens wird jedoch dadurch erschwert, dass neben den Fähigkeiten und Fertigkeiten einer Person ebenso Faktoren wie z.B. unterschiedliche Einstellungen, Erwartungen, Motive und an den Arbeitsplatz gestellte Forderungen Einfluss auf die Eignung oder Nichteignung einer Person haben. So wird beispielsweise ein Polizeibeamter, der das Führen von Mitarbeitern nicht als reizvoll und motivierend erlebt, möglicherweise über Jahre hinweg genau jene Situationen vermeiden, in denen aktiv geführt werden müsste. Zusätzlich reagieren Menschen unterschiedlich auf Veränderungen, Herausforderungen oder Monotonie, wirken aktiv auf ihre Umwelt und Entwicklung ein oder fühlen sich eher fremdbestimmt oder selbst aktiv einwirkend, etc. (vgl. Weinert, 1998). Bei der Konzeption eines effizienten Personalausleseverfahrens sollten folglich alle für die Ausübung der Tätigkeit erforderlichen Fähigkeiten und Fertigkeiten sowie die sie beeinflussenden Faktoren (Stress, Persönlichkeitsmerkmale, etc) erfasst werden und mittels bereits existierender oder / und neu zu entwickelnde Verfahren operationalisiert werden.

Bevor es jedoch zum Einsatz eignungsdiagnostischer Verfahren kommt, sind einige Vorbereitungen notwendig. Das folgende Schema nach Rösler (1992) zeigt die einzelnen Stadien, die bei der Etablierung wissenschaftlich begründeter Personalauslese durchlaufen werden müssen (Tabelle 2.1-1).

Tab. 2.1-1: Schematische Übersicht einzelner Stadien wissenschaftlich begründeter Personalauswahl nach Rösler (1992).



Anhand von Arbeits- und Anforderungsanalysen lassen sich die erforderlichen Fähigkeiten für die Ausübung der Tätigkeit feststellen, während die Fertigkeiten und Kenntnisse auf

Seiten der Person mittels eignungsdiagnostischer Verfahren überprüft werden können. Der Arbeitsplatz wird durch die Gesamtheit aller Aufgaben, die von dem jeweiligen Stelleninhaber bewältigt werden müssen, definiert. Eine Methode der Identifizierung der an einem Arbeits- bzw. Ausbildungsplatz auszuführenden Aufgaben oder auszuübenden Tätigkeiten, ihrer Ausführungsbedingungen sowie ihrer psychischen, physischen und sozialen Umfeldbedingungen ist die Arbeitsplatzanalyse. Nach Kannheiser und Frieling (1992) ist das Ziel der Arbeitsplatzanalyse u.a. die Erstellung einer empirischen Basis für die Konstruktion und die Auswahl von Prädiktoren und Kriterien sowie von Job-Typologien, anhand derer eine Validitätsgeneralisierung vorgenommen werden kann. Ebenso soll die Arbeitsplatzanalyse solche Faktoren ermitteln, die für die berufliche Zufriedenheit bedeutsam sind (Heyse und Kersting, 2004). In der Eignungsdiagnostik spielen unterschiedliche Arten von Anforderungen eine Rolle. Nennenswert sind Eigenschaftsanforderungen (z.B. Fähigkeiten und Interessen), Verhaltensanforderungen (z.B. Fertigkeiten und Gewohnheiten), Qualifikationsanforderungen (z.B. Kenntnisse und Fertigkeiten) und Ergebnisanforderungen (z.B. Problemlösungen und Qualitätsstandards) (Schuler, 1996).

Mit Hilfe diagnostischer Verfahren sollen überdauernde Eigenheiten des Arbeitsplatzes erfasst werden. Somit unterliegen sie, wie auch andere diagnostische Verfahren, den allgemeinen methodischen Anforderungen an Objektivität, Reliabilität und Validität. In der Regel besitzen derartige Instrumente die Form von Fragebögen, die entweder vom Arbeitsplatzinhaber selbst beantwortet werden oder mittels eines standardisierten Beobachtungsinterviews, das von fachkundigen und besonders geschulten Fachleuten, z.B. Psychologen, durchgeführt wird (Rösler, 1992).
Nur wenn Aussagen über die Passgenauigkeit zwischen den Fähigkeiten des Mitarbeiters und den Anforderungen des Arbeitsplatzes vorliegen, können Maßnahmen zur Personalentwicklung gezielt geplant, durchgeführt und evaluiert werden. In praxi wird die Arbeitsplatzanalyse jedoch oft aufgrund von Kostenüberlegungen vernachlässigt. Ebenso erscheint das häufig vorgebrachte Argument nicht unbedingt zwingend, dass die Bestimmung der Anforderungen eines Arbeitsplatzes in Zeiten eines raschen Wandels nicht sinnvoll möglich sei, da sich die Merkmale bereits in wenigen Jahren wieder verändert haben. Einerseits werden nicht alle Arbeitsplätze in der gleichen Art und Weise von den sich verändernden Entwicklungen beeinflusst und andererseits können die Veränderungen in ihrer Einflussnahme zumindest ansatzweise eingeschätzt werden.

Zur Bestimmung von Anforderungen[1] und Befriedigungspotentialen stehen drei Wege zur Verfügung: die arbeitsplatz-empirische, die personenbezogen-empirische und die erfahrungsgeleitet intuitive Methode, wobei sich die arbeitsplatzanalytisch-empirische Methode in den meisten Fällen als die Methode der Wahl herausgestellt hat (Eckardt & Schuler, 1992).
Anhand eines Beispieles der Deutschen Gesellschaft für Personalwesen (DGP) soll kurz der Weg zur Bestimmung von Anforderungen für den Polizeidienst erläutert werden. Ergebnisse zur arbeitsplatzanalytischen Methode lagen bereits aus einer Studie vor. Dieser verhaltensorientierte Ansatz wurde durch die Ergebnisse mehrerer polizeispezifischer DGP-Bewährungskontrollen um einen personenzentrierten Ansatz ergänzt. Auf Grundlage dieser Studienergebnisse konnten die Merkmale erfolgreicher Polizisten ermittelt werden. Diese Merkmale wurden dann in einem nächsten Schritt in ein Anforderungsprofil für zukünftige Polizisten transformiert. Damit auch mögliche Veränderungen berücksichtigt werden konnten, wurden die genannten Ansätze um die erfahrungsgeleitet intuitive Methode ergänzt. Das erstellte Anforderungsprofil wurde in Workshops polizeiinternen Experten zur Begutachtung und Modifikation vorgelegt (Kersting, 2004).

In der Forschung werden die Verfahren zur Arbeitsplatzanalyse in standardisierte und nicht standardisierte Verfahren unterteilt. Bei standardisiertem Vorgehen werden in den meisten Fällen Messinstrumente verwendet, die im Handel zu erwerben sind und unabhängig vom spezifischen Arbeitsplatz in immer der gleichen Weise eingesetzt werden. Zu den im deutschsprachigen Raum bekanntesten standardisierten Verfahren zur Arbeitplatzanalyse gehören u.a. der Fragebogen zur Arbeitsplatzanalyse (FAA) von Frieling und Hoyos (1978) und das Arbeitswissenschaftliche Erhebungsverfahren zur Tätigkeitsanalyse (AET) von Rohmert und Landau (1979).

Auf der anderen Seite stehen die nicht standardisierten Verfahren. Diese erlauben einen größeren Handlungsspielraum und somit eine größere Flexibilität dadurch, dass nur die grundlegenden Prinzipien des Vorgehens erläutert werden. Besonders in Berufen, in denen komplexe zwischenmenschliche Interaktionen eine Rolle spielen (nicht gewerblicher Bereich) oder der Arbeitplatz vielfältige und komplexe Anforderungen aufweist (z.B. Tätigkeit einer Führungskraft), sind den standardisierten Verfahren schnell Grenzen gesetzt und nicht standardisierte Verfahren kommen zum Einsatz. Kaufmann (1984) nennt als Beispiel für nicht

[1] Im Deutschen werden die Begriffe Arbeitsplatzanalyse und Anforderungsanalyse häufig synonym verwendet (vgl. auch Schuler, 2001).

standardisierte Verfahren Expertensitzungen aufgrund impliziter Annahmen über erfolgreiches Führungsverhalten. Zu nennen wären außerdem die Methode der kritischen Ereignisse nach Flanagan (1949) und die explorativ statistische Methode. Besonders in der erstgenannten Methode ist die Qualität der jeweiligen Analyse abhängig von der Qualifikation und Motivation der befragten Personen. Somit unterliegen diese Urteile, besonders bei der Befragung von Experten, den typischen Urteilsfehler (vgl. Kapitel 2.1.2). Die explorativ statistische Methode versucht diesen Fehler dadurch zu verringern, dass die Analyse weitestgehend mathematisch erfolgt (Kanning, 2002).

Auf Grundlage der Ergebnisse der Arbeitsplatzanalyse kann ein Anforderungsprofil erstellt werden, welches Persönlichkeitsmerkmale als „Sollangabe“ benennt (Fisseni, 1990). Enthalten könnte ein solches Profil u.a. die Hauptaufgaben und Ziele der Position, interne und externe Kontakte der Position (mit wem wird zusammengearbeitet, kommuniziert etc.), erforderliche Kenntnisse und Fähigkeiten (z.B. einschlägige Ausbildung, Weiterbildung, berufliche Fachkenntnisse, Sprachkenntnisse), erforderliche Kompetenzen (z.B. Belastbarkeit, Durchsetzungsfähigkeiten, Argumentation, Konfliktfähigkeit, analytisches Denken, sicheres Auftreten, Motivationskraft, Kundenorientierung) und wünschenswerte Erfahrungen (z.B. Führungserfahrung, Projekterfahrung, Interkulturelle Erfahrung).

Je klarer folglich ein Anforderungsprofil für die zu besetzende Stelle von den Auswahlverantwortlichen abgestimmt und festgelegt wurde, umso einfacher wird die Auswahl für einen bestimmten Kandidaten sein, umso klarer die Auswahlbegründung ausfallen, und umso größer das Verständnis der abgelehnten Kandidaten. Wer somit auf die Erstellung eines Anforderungsprofils im Vorfeld der Stellenausschreibung verzichtet und glaubt, im Laufe der verschiedenen Vorstellungsgespräche werde sowohl das Profil der Stelle als auch die Eignung der einzelnen Kandidaten deutlich werden, spart an der falschen Stelle.

Die Vorhersage von berufsrelevanten Variablen mit Hilfe unterschiedlicher Verfahren steht im Mittelpunkt der berufseignungsdiagnostischen Praxis. Der Einsatz der Verfahren dient stets dem Zweck eines Prädiktors auf Grundlage dessen das spätere Verhalten, insbesondere der Erfolg eines Teilnehmers am Auswahlverfahren, in der Ausbildung oder am Arbeitsplatz, vorhergesagt werden soll. Das Ergebnis einer Person in einem Testverfahren stellt somit den Prädiktor dar, anhand dessen auf den späteren beruflichen Erfolg (Kriterium) geschlossen wird. Nach Maukisch (1978) sind Prädiktoren diagnostische Verfahren, die

Persönlichkeitsmerkmale beliebiger Art in irgendeiner Weise operationalisieren und für die Vorhersage eines Kriteriums eingesetzt werden. Prädiktoren stellen nicht nur reine „Fähigkeitsaspekte“ wie die Leistung einer Person dar, sondern ebenso Motivations- und Einstellungsaspekte, denn durch diese Merkmale wird die Leistung mitbestimmt (Rösler, 1992). Prädiktoren müssen den empirischen Gütekriterien genügen, d.h. sie müssen objektiv, reliabel und valide sein. Besonders wichtig ist hier die prognostische Validität, denn ohne sie wäre eine Vorhersage nicht möglich.

In der eignungsdiagnostischen Praxis sind sehr viele unterschiedliche Testverfahren als Prädiktoren zur Vorhersage der beruflichen Leistung bzw. des beruflichen Erfolges anzutreffen, die in vielen Fällen auch miteinander kombiniert werden (vgl. Kapitel 2.3). Mehrere Methoden können dabei parallel nebeneinander oder sukzessive hintereinander eingesetzt werden. Im ersten Fall müssen die Testergebnisse der Methoden miteinander verrechnet werden, während im zweiten Falle jede Methode bereits für sich allein zu einer Entscheidung führt. Verfahren, die in der Eignungsdiagnostik eingesetzt werden, sind beispielsweise Arbeitsproben, Interviews, standardisierte Testverfahren oder Assessment Center (s.a. Kapitel 2.3).

Ein viel verwendetes und adäquates Bewährungsmaß in der berufseignungsdiagnostischen Praxis ist die Korrelation zwischen Prädiktor und Kriterium. Anhand der multiplen Korrelation lässt sich bspw. berechnen, welche Kombinationen und welche Gewichtung der Prädiktoren optimale Vorhersagen erlauben. Die Höhe des Korrelationskoeffizienten erlaubt einen Vergleich zwischen unterschiedlichen Prädiktoren. In praxi werden in der Berufseignungsdiagnostik Korrelationskoeffizienten in der Höhe von $r = .30$ erreicht (Schuler, 1996). Derartige Korrelationskoeffizienten finden sich z.B. zwischen den Werten, die Bewerber in einer gut konstruierten Arbeitsprobe erzielen und ihren späteren Leistungsbeurteilungen durch Vorgesetzte. Nach Schuler (1996) sind Korrelationskoeffizienten in der Berufseignungsdiagnostik in der Höhe von $r = .50$ bereits als gut zu klassifizieren. Beispielsweise fand sich in einer Untersuchung von Backhaus & Wagner (1994) ein Zusammenhang von $r = .48$ zwischen den Werten in einem Intelligenztest, der von einem großen Unternehmen zur Auswahl seiner Lehrlinge eingesetzt wird, und den Prüfungsergebnissen zum Abschluss ihrer Ausbildung. Eine differenzierte Betrachtung der prognostischen Validität findet sich unter Kapitel 2.3.

Bei der Analyse der korrelativen Beziehungen müssen jedoch nach Jäger (1966) u.a. einige Aspekte berücksichtigt werden. In den meisten Fällen der Berufseignungsdiagnostik handelt es sich um selegierte Stichproben. Die somit entstehende Varianzeinschränkung muss durch Minderungskorrekturen (vgl. Lienert, 1989) aufgefangen werden, da sonst der Vergleich der Vorhersagewerte der Prädiktoren zu falschen Schlüssen führt. Das Abschneiden der Extreme der Verteilung kann sich negativ auf die Korrelationen zwischen Prädiktoren und Kriterien auswirken. Vor allem bei kurvilinearen Beziehungen zwischen Prädiktoren und Kriterien kann eine ungünstige Selektionsgrenze zwischen angenommenen und abgelehnten Personen der Validitätskoeffizient gen Null gehen, obwohl der Prädiktor an sich hoch valide ist (Althoff, 1984).
Weiterhin müssen nach Jäger (1966) so genannte Suppressorvariablen in der Interpretation berücksichtigt werden. Wenn sich in der Analyse keine direkten Korrelationen der Prädiktoren mit den Kriterien zeigen, bedeutet dies nicht zwingend, dass sie für die Interpretation wertlos sind. Die Suppressorvariable kann den Vorhersagebeitrag einer oder mehrerer anderer Variablen erhöhen, indem sie irrelevante Varianzen in der anderen Prädiktorvariable unterdrückt. Durch die Absorbierung des störenden Merkmalanteils erhöht die Suppressorvariable die Nützlichkeit der anderen Prädiktorvariable (Bortz, 1993).

Wie bereits angesprochen, werden bei der Analyse der Vorhersageleistung die Testwerte der Prädiktoren in Beziehung zu einem Kriterium gesetzt. Die Kriterien dienen als Maßstab für die Qualität von Entscheidungen. Sie unterliegen, ebenso wie die Prädiktoren, den gleichen Anforderungen an die empirischen Gütekriterien der Objektivität, Reliabilität und Diskriminationsfähigkeit (Rösler, 1992). Nach Maukisch (1978) verstehen wir unter Kriterien Verhaltensweisen und Verhaltensergebnisse, die einer sozialen oder individuellen Bewertung unterliegen und geeignet sind, den Grad und die Art der Wertverwirklichung, der durch psychologische Intervention erreicht wird, zu erfassen. In der Berufseignungsdiagnostik versteht man unter einem Außenkriterium die objektiv bewertbare Leistung eines Probanden, die seine berufspraktische Bewährung reflektiert und zur Prüfung der Eignungsprognose herangezogen werden kann (Schmale & Schmidtke, 1969). Im Gegensatz zu den Prädiktoren, die an der Wirklichkeit geprüft werden sollen, entziehen sich die Kriterien der Validierung, da sie die Wirklichkeit selbst darstellen (Färber, 1995).
Hossiep (1995) unterscheidet drei Klassifizierungsmöglichkeiten: die Dauer der Zeit, die erfasst wird (kurzfristiges Verhalten nach Einstellung vs. Verhalten nach unterschiedlich langem Verbleib in der Organisation), das Ausmaß der angestrebten Spezifität (spezifisches

vs. globales Verhalten am Arbeitsplatz bzw. in der Organisation) und die Nähe zu den Organisationszielen (Nähe bzw. Abweichung zu festgelegten Aufgaben in der Funktion bzw. der Organisation insgesamt). Zentral ist bei der Definition der Kriterien die Frage, ob sie die wichtigsten und bedeutsamsten Aspekte einer Tätigkeit repräsentieren, d.h. für eine Tätigkeit relevant sind.

Eine weitere Unterscheidung orientiert sich an der Erfassbarkeit der Kriterien. Als so genannte „harte Kriterien" gelten jene, die der direkten Beobachtung zugängliche Verhaltensweisen wie Häufigkeit des Verspätens, die Anzahl der Krankheitstage, die Höhe der erreichten Verkaufszahlen oder die Anzahl ausgesprochener Kündigungen seitens des Arbeitgebers bzw. des Arbeitnehmers etc. erfassen. Unter den „weichen Kriterien" werden dagegen bspw. Beurteilungen des Verhaltens durch Vorgesetzte oder Kollegen verstanden. Doch auch harte Kriterien unterliegen der Gefahr von Verzerrungen, wenn das Ausmaß der Bewährung von Einflüssen abhängt, die nicht einer Person zugeschrieben werden können. Wenn bspw. die Zunahme von sich verkehrswidrig verhaltenden Autofahrern in einem Stadtteil nicht aufgrund der Unzuverlässigkeit des zuständigen Polizeibeamten erfolgt, sondern weil ein neu eingeführtes Parkverbot- und -gebotsystem nicht die Akzeptanz der Autofahrer findet. Bei Beurteilungen sind zudem immer Verzerrungen im Sinne bspw. eines Halo-Effektes zu bedenken, d.h. dass günstige oder weniger günstige Einschätzungen eines Merkmals sich auf die Einschätzung weiterer Merkmale übertragen können und ein sympathischer Kollege zugleich als leistungsfähig und intelligent eingeschätzt wird.

Die im Modell folgende Validitätsanalyse begründet eine wissenschaftlich fundierte Personalauswahl, d.h. die Klärung der Frage, in welchem Umfang ein Kriterium durch einen Prädiktor vorhersagbar ist. Durch den empirisch begründeten Nachweis über einen Zusammenhang zwischen Prädiktor und Kriterium, der sich durch die Höhe des Validitätskoeffizienten abbildet, kann aufgrund der Prädiktorinformation zwischen geeigneten und weniger geeigneten Bewerbern um eine Funktion in einer Organisation unterschieden werden. Generell bieten sich hier zwei Vorgehensweisen an: Zum einem wird der Prädiktor zum Zeitpunkt der Auswahl erhoben und das Kriterium erst zu einem späteren Zeitpunkt (prädiktive Validität), zum anderen können Prädiktor und Kriterium zum gleichen Zeitpunkt an einer Stichprobe von bereits ausgewählten Personen erhoben werden (konkurrente Validität). Auffällig waren hier bis in die siebziger Jahre die widersprüchlichen Ergebnisse unterschiedlicher Validitätsstudien. So ergaben Validitätsschätzungen für dasselbe Verfahren und für denselben Beruf unterschiedlich hohe Validitätskoeffizienten (Theorie der

situationalen Spezifität). Jedoch ist heute belegt, dass die meisten Unterschiede auf statistische Artefakte und Messfehler, wie z.B. zu kleine Stichproben, zurückzuführen waren (Schmidt & Hunter, 1977; Schmidt, Hunter, Pearlman & Shane, 1979). Rössler (1992) zieht hinsichtlich der Generalisierbarkeit vorliegender Validitätskoeffizienten folgendes Fazit: Validitätskoeffizienten sind in hohem Maße generalisierbar. Eine Aussage über die Validität hat dann einen Sinn, wenn sich eine Studie auf eine ausreichend große Stichprobe stützt. Für eine Validitätsgeneralisierung sind keine aufwendigen und minutiösen Arbeitsplatzanalysen erforderlich, da es innerhalb einer Job-Familie (z.B. Aufgaben im Verwaltungsdienst) kaum tatsächliche Unterschiede in der Validität gibt. Dennoch darf nicht unbeachtet bleiben, dass durch steten Wandel der Gesellschaft und einer sich rasch weiter entwickelnden Technik die Anforderungen einer Tätigkeit sowie die Tätigkeit an sich verändern. Es bleibt eine Herausforderung für eine wissenschaftlich orientierte Personalauswahl, alle eignungsrelevanten Anforderungen zum Zeitpunkt der Auswahl zu bestimmen und zu berücksichtigen, um eine zutreffende Prognose über die Eignung eines bestimmten Bewerbers zu geben. Aufgrund dessen erscheint es umso wichtiger zusätzlich zu der Betrachtung der Kongruenz zwischen den Anforderungen eines Arbeitsplatzes und den dafür notwendigen Fähigkeiten, das Entwicklungspotential einer Person möglichst gut einschätzen und vorhersagen zu können und im Sinne einer effektiven Personalentwicklung den Arbeitnehmern die Möglichkeit zu geben, berufsbezogene Fähigkeiten anhand von Weiter- und Fortbildungen zu verfestigen.
Die Frage nach der Strategie einer Entscheidung über geeignete bzw. nicht geeignete Bewerber wird in Kapitel 2.1.2 vertiefend dargestellt.

Die Auswahl geeigneter Mitarbeiter aus einem vorhandenen Bewerberpotential ist das primäre Ziel psychologischer Eignungsdiagnostik. Mittels der Datenerhebung, deren Interpretation und Kombination zu einem Gesamturteil werden Entscheidungen darüber gefällt, welche Bewerber für einen Arbeitsplatz geeignet sind und welche ungeeignet. Neben der Einbeziehung von Daten aus Testverfahren, bei denen das individuelle „Ankreuzverhalten“ einer Person mit einem Normwert verglichen wird, werden ebenso solche Informationen in das Gesamturteil über die Eignung einer Person integriert, welche auf Grundlage von Beobachtung und Interpretation des Verhaltens einer Person gewonnen wurden. Bei der Durchführung eines Interviews beispielsweise werden aufgrund von Beobachtungen Daten erhoben. Diese werden durch Beurteilungsprozesse seitens des Diagnostikers beeinflusst. Nach der Sammlung aller vorliegenden Informationen und deren

Bewertung und Kombination werden diese dann zu einem Urteil über eine Person integriert. Da es sich bei der Erstellung eines diagnostischen Urteils nach Sarbin, Taft und Bailey (1960) um einen kognitiven Prozess handelt, der zudem auf Grundlage einer komplexen Interaktion zwischen Proband und Diagnostiker basiert, kann es an zahlreichen Stellen dieses Prozesses der Urteilbildung zu Fehlerquellen, sprich Invalidierungen des Urteils kommen.

Prinzipiell wird einerseits die klinische von der aktuarischen oder statistischen Urteilbildung unterschieden (Kastner, 1995). Bei der klinischen Vorgehensweise basiert die Urteilsfindung auf Intuition und Erfahrung, auf deren Grundlage die Datenintegration zu einem Urteil erfolgt. Der Begriff „klinisch“ bezieht sich nicht auf ein bestimmtes Anwendungsgebiet, sondern auf die Art der Verarbeitung der Daten, nämlich der am Einzellfall orientierten Vorgehensweise. Im Gegensatz dazu werden die Daten bei der statistischen Methode nach bestimmten Systematiken und Regeln erhoben und verknüpft, so dass vorangegangene Invalidierungen des Urteils durch die Persönlichkeit des Diagnostikers weitestgehend minimiert werden können. Die Frage, welcher der beiden Vorgehen die effizientere sei, ist schon ab Mitte der 20er Jahre des letzten Jahrhunderts kontrovers diskutiert worden (vgl. z.B. Meehl, 1954; Sawyer, 1966). Letztendlich scheint die Frage der Überlegenheit einer der beiden Methoden in dieser allgemeinen Form nicht zu beantworten zu sein, da bei der Urteilsbildung ungleich mehr Variablen und Faktoren als die genannten eine Rolle spielen (Kastner, 1995).

Wie auch schon bei der Bewertung so genannter weicher Kriterien deutlich wurde, spielen Verzerrungen und subjektive Wahrnehmungen auch in der Personalauswahl eine besondere Rolle und beeinflussen die Güte von Entscheidungen und Bewertungen. Die Psychologie ist eine Wissenschaft vom menschlichen Verhalten und Erleben, die sich auch mit den Prozessen der Informationsverarbeitung auseinandersetzt. Bewertungen beziehen sich dabei sowohl auf die Beurteilung von Personen, wie auch die Beurteilung von Situationen, wobei deren Grenzen oft fließend verlaufen. In unzähligen empirischen Untersuchungen wurde versucht jene Fehler zu beleuchten, die einer Person bei der Beurteilung anderer Menschen und sozialer Situationen unterlaufen können. Nach Kanning (1999) legen viele Ergebnisse den Schluss nahe, dass die menschliche Informationsverarbeitung viel unvollkommener ist, als sie uns auf den ersten Blick hin erscheinen mag. Fehler sind schon in der Art und Weise, wie Menschen ihre Umwelt wahrnehmen, interpretieren, speichern und gedanklich verarbeiten, angelegt. So ist neben dem oben bereits erwähnten Halo-Effekt im Rahmen der Impliziten

Persönlichkeitstheorie die Annahmen zu erwähnen, die Menschen über die Beziehung zwischen zwei oder mehr Persönlichkeitsmerkmalen machen. Weist ein Bewerber eines dieser Merkmale auf, werden ihm auch die anderen zugeschrieben. Die umfassende Darstellung verschiedener Urteilsfehler und Urteilsverzerrungen, die Einfluss auch auf den diagnostischen Prozess haben, würde den Rahmen schnell sprengen. Deutlich werden sollte allerdings, dass eine Vielzahl unterschiedlicher Faktoren die diagnostische Urteilsbildung auch im Rahmen der eignungsdiagnostischen Personalauswahl beeinflussen kann. Neben den Einstellungen, Erwartungen, kognitiven Schemata, impliziten Persönlichkeitstheorien auf Seiten des Diagnostikers spielen ebenso die Art der Datengewinnung, die verwendeten Diagnoseverfahren sowie die Ziele und Konsequenzen der Beurteilung eine Rolle und stellen somit mögliche Fehlerquellen dar (vgl. Schuler, 1996; Kastner, 1995; Mattenklott, 1992).

So wird der Prozess der kognitiven Verarbeitung von erhobenen Daten und sich anschließenden Urteilen als schlussfolgerndes Denken bezeichnet und kann mit systematischen Fehlern einhergehen. Bei der Beurteilung eines Bewerbers sind alle Informationen sorgfältig abzuwägen, potentielle zusätzliche Informationen zu gewinnen und alles gemeinsam entsprechend zu gewichten und in Beziehung zu setzen. Die eignungsdiagnostische Untersuchung hat als Ziel, aus den Ergebnissen jene Aussagen zu gewinnen, die eine Antwort auf die diagnostische Frage (Eignung für den Beruf) erlauben. Inhaltlich vollzieht sich diese Ableitung während des gesamten diagnostischen Prozesses. Formell wird sie erst nach der Untersuchung abgeschlossen, wenn alle benötigten Informationen gewonnen werden konnten. Hier kann es schnell dazu führen, dass eine derart aufwendige Prozedur zur Vernachlässigung einzelner Informationen führt oder nicht alle relevanten Informationen integriert werden können. Beispielweise kann die Ursache für eine nicht zufrieden stellende Bewertung eines Bewerbers in einem Vorstellungsgespräch im gerade erst zur Kenntnis genommenen Trauerfall in der Familie liegen, den zu erwähnen, er jedoch aus verständlichem Schamgefühl vermied. Nach Fiedler (1997) kommen in solchen komplexen Situationen Heuristiken zum Einsatz, die als kognitive Werkzeuge, mit einem vergleichsweise geringen Aufwand, häufig zufrieden stellende Ergebnisse liefern. In der Folge wird bspw. die Wahrscheinlichkeit oder Häufigkeit eines Ereignisses eingeschätzt (Verfügbarkeitsheuristik) oder es werden einzelne Informationen, die zur Verfügung stehen, hinsichtlich ihrer Repräsentativität bewertet (Repräsentativitätsheuristik). Diese und weitere Heuristiken machen deutlich, dass Menschen dazu neigen, in einer hochkomplexen Welt durch Gestaltung ökonomischer Prozesse, sich ihre Handlungsfähigkeit zu bewahren. So

werden Individuen, analog der individuellen Wahrnehmung, die Verarbeitung, Abspeicherung und Erinnerung von Informationen gemäß ihrer individuellen Erfahrungen verarbeiten.

Ein anderes Beispiel für Urteilverzerrungen sind die Erwartungen von Beurteilern, die zu einer verzerrten Wahrnehmung der Realität führen können. Beispielhaft sei hier der Pygmalion-Effekt (Rosenthal & Jacobson, 1971) erwähnt, bei dem nachgewiesen werden konnte, wie sich die Erwartungen der Lehrer auf das tatsächliche Leistungsverhalten der Schüler auswirken. Erwartungen können auch im Rahmen des Labelings Auswirkungen auf die Eignungsbeurteilung haben. Bewerber, die sich mit einem bestimmten, mit Vorurteilen behafteten Etikett um eine Funktion bewerben, werden u. U. aufgrund gegebener Erwartungen bewertet. Zeigte ein Proband in der Vergangenheit ein bestimmtes Verhalten (z.B. Straffälligkeit), besteht die Gefahr, dass nicht ein einzelnes kriminelles Verhalten bewertet wird, sondern eine dahinter stehende kriminelle Persönlichkeit.

Das Bewusstsein, dass jede Form der Verzerrung in der Beurteilung anderer Menschen zu folgenschweren Beurteilungsfehlern führen kann, und neben den erwähnten Illustrationen ließe sich eine Vielzahl ergänzender Beispiele anfügen (vgl. a. Kanning, 1999), bietet die erste Grundlage dafür, den entsprechenden Fehlern zu begegnen. Denn die Interventionen zur Verringerung der Beurteilungsverzerrungen sollen bewirken, dass Automatismen unterbrochen werden und an ihre Stelle eine bewusst vorgenommene Beurteilung gestellt wird. Wichtig sind somit genaue Kenntnisse zu den allgemeinen Prinzipien menschlicher Urteilsbildung. So ist es beispielsweise wichtig, sich vor Augen zu führen, was genau im Beurteilungsprozess erfasst werden soll. Die vollständige Erfassung der gesamten Persönlichkeit eines Menschen ist selbst mit aufwendigen Untersuchungsmethoden kaum erreichbar. Für den Beurteiler bedeutet das, sich auf jene Verhaltensaspekte zu konzentrieren, die für die vorgeschriebenen Beurteilungsmerkmale von Bedeutung sind und Vernachlässigung jener, die keine Relevanz haben. Weiter ist eine strikte Trennung zwischen Beobachtung und Beurteilung ein wichtiger Aspekt. Durch die Beobachtung werden nur möglichst präzise Beschreibungen des Verhaltens erfasst, die jedoch keine Bewertungen erhalten. Erst in einem zweiten Schritt erfolgt anhand eines Maßstabes die Bewertung. Für die Prognose zukünftigen Verhaltens ist es bedeutsam, sich zu vergegenwärtigen, dass Vorhersagen nicht mit Sicherheit getroffen werden können, sondern nur mit einer gewissen Wahrscheinlichkeit. Dabei sind auch immer die Merkmale der jeweiligen Situation zu berücksichtigen, in denen das beobachtete Verhalten eines Bewerbers auftrat. Denn das

Verhalten eines Menschen wird neben den mitbestimmenden Persönlichkeitsmerkmalen oder Verhaltensgewohnheiten auch durch die jeweilige Situation mit beeinflusst.

2.1.2 Entscheidung und Nutzen

Unabhängig von der verwendeten Methode der Urteilsbildung geht es in der psychologischen Eignungsdiagnostik um eine Entscheidungsoptimierung. Anhand der Daten, welche über den Prozess der Urteilsbildung gewonnen werden konnten, sollten möglichst einwandfrei die geeigneten Bewerber von den ungeeigneten differenziert werden können. Auf mögliche Fehler dieses Prozesses wurde im vorausgehenden Kapitel differenzierter eingegangen. Falsche Entscheidungen zugunsten ungeeigneter Bewerber werden mit großer Wahrscheinlichkeit negative Konsequenzen für beide Seiten haben. Auf Seiten des Bewerbers können aufgrund von Unter- oder Überforderung physische und psychische Probleme resultieren (vgl. psychophysiologische Erkrankungen; Comer, 1995). Auf Seiten der Organisation resultieren erhebliche Kosten durch u.a. krankheitsbedingte Ausfälle, Minderleistungen, Kosten eines weiteren Auswahlverfahrens.

Anhand diagnostischer Strategien soll diese Entscheidungsoptimierung umgesetzt werden. Jäger und Petermann (1992) definieren eine diagnostische Strategie als eine auf diagnostischen Daten aufbauende Konzeption, mit deren Hilfe der Diagnostiker ein vorher festgelegtes Ziel erreichen kann. In praxi umgesetzt bedeutet das zunächst, eine Fragestellung festzulegen, die es zu untersuchen gilt. Diese differieren u.a. in ihrer Spezifität, im Thema, in der Art ihrer Umsetzung, etc.. Als allgemeine Fragestellung könnte beispielsweise gelten: Ist ein Bewerber für eine bestimmte Position geeignet oder nicht? Im Vergleich dazu handelt es sich um eine spezifische Fragestellung, wenn es darum geht, welche Werte eine bestimmte Bewerberstichprobe in einem speziellen Testverfahren erreicht.

Auf Grundlage der Antworten werden Entscheidungen für einzuleitende Maßnahmen getroffen, beispielsweise anhand welcher Methoden geeignete Bewerber in der Population effektiver angesprochen werden können. Das in der Ausgangsfragestellung definierte Ziel (effektive Auswahl geeigneter Bewerber) soll anhand der festgelegten Maßnahmen erreicht werden. Sofern dieses in einen diagnostischen Prozess eingebettet ist und diagnostische Daten die Auswahl der Maßnahme steuern, sprechen Jäger und Petermann (1992) von diagnostischer Strategie.

Im diagnostischen Prozess werden anhand diagnostischer Strategien Entscheidungen zwischen zwei oder mehr Alternativen gefällt. Da im Grunde nicht davon ausgegangen werden kann, dass das angewendete diagnostische Instrumentarium optimal auf die Fragestellung passt, sind die daraufaufbauenden Entscheidungen immer mit einem Fehler behaftet (Tabelle 2.1-2).

Nach Amelang und Zielinski (1994) werden bei Entscheidungen institutioneller Art alle Personen anhand eines standardisierten Vorgehens ausgewählt. Alle Bewerber bearbeiten z.B. das gleiche Verfahren. Man sucht eine Entscheidungsstrategie aus, bei der davon auszugehen ist, dass sie den größtmöglichen Nutzen und das geringste Fehlerrisiko darstellt. Bei individuellen Entscheidungen hingegen geht es um eine Person, welche anhand von diagnostischen Daten einem bestimmten „treatment“ (Behandlung) unterzogen wird. In diesem Falle geht es speziell um den größtmöglichen individuellen Nutzen.

Tab. 2.1.-2: Arten von Entscheidungen nach Cronbach und Gleser (1965)

1. Nutzen von Entscheidungen geht zugunsten	Institution	vs.	Individuum
2. Annahme	festgelegt	vs.	variabel
3. Behandlungen	singulär	vs.	multipel
4. Möglichkeit von Ablehnungen	ja	vs.	nein
5. Informationsdimensionen	univariat	vs.	multivariat
6. Entscheidungen	terminal	vs.	investigatorisch

Um festgelegte Annahmequoten handelt es sich dann, wenn feststeht, wie viele Personen z.B. eine Institution im Kontext der Personalauswahl aufnehmen will. In diesem Fall können die Geeigneten nur dann festgestellt werden, wenn von allen Personen diagnostische Daten vorliegen. Andererseits handelt es sich um nicht festgelegte oder variable Annahmequoten, wenn die Entscheidungen über einzelne Personen nicht von den anderen Personen abhängen. Die Eltern aller Kinder beispielsweise, die im 7. Lebensjahr noch einnässen, werden einer psychologischen Exploration unterzogen.
Unter dem Begriff der „Behandlung“ werden im Allgemeinen unterschiedliche Interventionen verstanden. Einerseits geht es eher um definierte Maßnahmen, wie z.B. die Therapie eines Klienten. Andererseits geht es um die Kombination vieler einzelner Treatments, wie z.B. den Einbezug der Familie in die Therapie. Cronbach & Gleser scheint es vornehmlich um die Unterscheidungen zwischen einstufigen und mehrstufigen bzw. sequentiellen Testungen zu gehen. Im einstufigen Fall erfolgt die Zuordnung z.B. aufgrund eines Testergebnisses,

während im mehrstufigen Falle mehrere unterschiedliche Ergebnisse zu einer Zuordnung führen. Erfolgt aufgrund von Ergebnissen eine Ablehnung von Personen, handelt es sich um eine Selektion, während ohne Ablehnung eine Zuordnung der Person zu einer Behandlung erfolgt.
Liegen bei einer Entscheidung allein die Daten eines Prädiktors vor, spricht man von der univariaten Beschaffenheit. Bestehen die Informationen jedoch aus mehreren Kennwerten, so sind sie multivariat beschaffen. Zur Erhöhung der Validität werden in den meisten Fällen mehrere Prädiktorendaten zur Analyse herangezogen, weil damit verschiedene Facetten des Kriteriums abgedeckt werden.
Ist man aufgrund der diagnostischen Informationen zu einer Entscheidung gekommen und hat eine Person einem Treatment zugewiesen (z.B. Aufnahme in ein Weiterbildungsprogramm), so spricht man von einer terminalen Entscheidung und die diagnostische Aufgabe ist abgeschlossen. Ist jedoch noch keine endgültige Entscheidung gefallen (Einstellung auf Probe), handelt es sich um eine investigatorische (vorläufige) Entscheidung.

In den meisten Fällen sind Entscheidungen eher investigatorischer Natur und weniger terminal. Man denke nur an den Kontext Universität, Behörde oder Betrieb. Hier werden im Sinne des investigatorischen Vorgehens fortwährend diagnostische Informationen über den Leistungs- und Motivationsstand jeder Person gesammelt, bis eine Person ihren endgültigen Platz in der Institution gefunden hat (oder andererseits abgelehnt wurde). In diesem Zusammenhang kommt aus finanziellen und zeitlichen Gründen das einstufige Vorgehen recht häufig zum Einsatz. Hier ist einerseits die nichtsequentielle Batterie zu nennen, in welcher die gesamte Testbatterie den Probanden vorgegeben wird. Diejenigen werden ausgewählt, die den aus allen Testergebnissen errechneten optimalen Summenscore erzielt haben. Ähnlich dem sogenannten „single screen“, bei dem alle weiteren Entscheidungen auf einem Testwert basieren.

Innerhalb des sequentiellen Vorgehens sind die folgenden drei Grundmuster der Entscheidungsstrategie möglich (Amelang & Zielinski, 1994): In der „pre-reject-Strategie“ werden die Probanden, die nach dem ersten Test einen bestimmten Wert nicht erreicht haben, von dem weiteren Verfahren ausgeschlossen. Die übriggebliebenen Probanden werden weiteren Testungen unterzogen, wobei deren Ablehnung oder Annahme aus der Kombination der Erst- und Zweittests errechnet wird. Diese Strategie spart der Organisation erhebliche Kosten und entspricht den in der DIN 33430 formulierten Leitsätzen, wonach die Bewerber

nicht mehr beansprucht werden sollen, als es für den eigentlichen Untersuchungszweck notwendig ist. Bei der Vorentscheidungs- oder „pre-accept-Strategie“ werden alle Probanden, die nach einem ersten Testteil einen bestimmten Wert erreicht haben, bereits akzeptiert. Die verbleibenden werden entsprechend der Vorauswahlstrategie behandelt. Nach Bestimmung der Punkte in einem Test wird die Gruppe der Probanden bei der vollständig sequentiellen Strategie dreigeteilt, eine Gruppe, die terminal akzeptiert wird, die andere wird definitiv abgewiesen und eine dritte wird mit einem Folgetest untersucht.

Wichtig bei der Auswahl einer geeigneten Strategie ist die Berücksichtung des von Cronbach (1990) unter dem Stichwort „bandwidth-fidelity-dilemma“ publik gewordenen Abwägungsproblems. „Bandwidth“ bedeutet in diesem Zusammenhang die Bandbreite oder Komplexität der Informationen, die durch ein diagnostisches Verfahren gewonnen wird. „Fidelity“ hingegen die Akkuratheit dieser gewonnenen Informationen (Schuler & Höft, 2001). Werden nun bei der Auswahl von Personen eine Vielzahl von Verfahren eingesetzt, um einen möglichst breiten Überblick über das Leistungspotential der Bewerber zu gewinnen, leidet darunter die „Fidelity“, die Detailtiefe. Spezifische einzelne Informationen können gezwungenermaßen nicht erhoben werden. Geht es andererseits darum, schwerpunktmäßig detaillierte Informationen zu gewinnen, reduziert sich zwangsweise der Anteil an gewonnenen globaleren bzw. komplexeren Informationen über die Probandenstichprobe. Zur Lösung des Problems nennen die Autoren die Symmetriehypothese von Cronbach und Gleser (1965), nach dem das Abstraktionsniveau von Prädiktor und Kriterium in etwa gleich sei soll. Ein spezifischer Prädiktor erfordert ein eng umgrenztes Kriterium und im Gegensatz dazu erfordert ein globaleres Kriterium einen Prädiktor mit einem vergleichbaren Abstraktionsniveau.

Die Zuordnung von Personen gemäß den beschriebenen Entscheidungsstrategien basiert auf einem Pool an diagnostischen Daten. Diese Informationsbasis ist jedoch meistens weder erschöpfend noch völlig zuverlässig. Außerdem ist die Validität der Entscheidungsregeln unsicher. Folglich sind alle Arten von Entscheidungen mit mehr oder weniger weitreichenden Fehlern behaftet.
Im Fall von zwei möglichen Klassen (z.B. geeignet vs. nicht geeignet) lassen sich nach Kallus und Jahnke (1992) vier mögliche Kombinationen von Übereinstimmung und fehlender Übereinstimmung unterscheiden, die in Tabelle 2.1-3 dargestellt werden.

Tab. 2.1-3: Arten richtiger und falscher Zuordnung nach Kallus & Jahnke (1992)

	ZUORDNUNG: GEEIGNET	ZUORDNUNG: UNGEEIGNET
TATSÄCHLICH: GEEIGNET	**richtige Zuordnung** richtig als geeignet ausgewählt Risiko: 1 - α	**falsche Zuordnung** fälschlich als ungeeignet bezeichnete Geeignete Risiko α
TATSÄCHLICH: UNGEEIGNET	**falsche Zuordnung** fälschlich als geeignet ausgewählte Ungeeignete Risiko β	**richtige Zuordnung** richtig als ungeeignet ausgewählte Ungeeignete Risiko: 1 β

Im Zweifelsfall lassen sich somit zwei Arten von Zuordnungsfehlern unterscheiden. Einerseits der Fehler erster Art, d.h. falsch Positive. Teilnehmer an einem Auswahlverfahren werden als geeignet bezeichnet, obwohl sie ungeeignet sind. Zum anderen der Fehler zweiter Art, d.h. falsch Negative. Die eigentlich geeigneten Teilnehmer werden als ungeeignet klassifiziert. Die Effizienz der Auslese bemisst sich nach dem Anteil der Geeigneten an allen Ausgewählten, wobei dieser Term auch selektiver Eignungskoeffizient genannt wird (Kallus & Jahnke, 1992).

Dieser Parameter ist das Effizienzkriterium für eine Entscheidungsstrategie (Taylor & Russell, 1939). Um diesen berechnen zu können, benötigt man die Grundrate, die Selektionsquote und die Prädiktorvalidiät. Dabei bildet die Grundrate den Anteil der Geeigneten in der Bewerberpopulation und die Selektionsquote den Anteil der Personen, die von allen zu untersuchenden ausgewählt werden sollen. Anhand der von Taylor und Russel (1939) veröffentlichten Tabellen lässt sich die Erfolgsquote in Abhängigkeit verschiedener Ausprägungskombinationen dieser Einflussgrößen ablesen.

Die folgende Tabelle 2.1-4 zeigt in Anlehnung an Taylor und Russel (1939) exemplarisch das Zusammenspiel der drei Größen Grundquote, Selektionsquote und Testvalidität (Petersen, 2002).

Tab. 2.1-4: Zusammenwirken von Grundquote, Selektionsquote und Prädiktorvalidität (Petersen, 2002)

FREIE STELLEN IM UNTERNEHMEN = 50 **ANZAHL DER GEEIGNETEN** BEWERBER = 90 VON 100			
Validität des Auswahlverfahrens	**r = .00**	**r = .35**	**r = .65**
Prozentsatz der Geeigneten unter den Eingestellten	90 %	95 %	98 %
FREIE STELLEN IM UNTERNEHMEN = 50 ANZAHL DER GEEIGNETEN BEWERBER = 50 VON 100			
Validität des Auswahlverfahrens	**r = .00**	**r = .35**	**r = .65**
Prozentsatz der Geeigneten unter den Eingestellten	50 %	61 %	73 %
FREIE STELLEN IM UNTERNEHMEN = 10 ANZAHL DER GEEIGNETEN BEWERBER = 50 VON 100			
Validität des Auswahlverfahrens	**r = .00**	**r = .35**	**r = .65**
Prozentsatz der Geeigneten unter den Eingestellten	50 %	74 %	92 %

Die dargestellten Beispiele zeigen, dass der Einsatz eines validen Verfahrens umso wichtiger ist, je geringer der Anteil Geeigneter unter den unausgelesenen Bewerbern und je größer die Selektionsquote ist. Andererseits ist die Prädiktorvalidität weniger wichtig, wenn die Grundrate hoch und die Selektionsquote niedrig ist. Folglich kann auch ein Test mit geringer Validität z.B. im Rahmen einer Testbatterie durchaus zur Effektivität eines Auswahlverfahrens beitragen.

Die Taylor-Russel Tafeln waren ein bedeutender Schritt in Richtung kosten- und nutzenorientierter Personauswahl, jedoch weist dieses Modell ebenso Schwachpunkte auf. Das Modell bezieht sich ausschließlich auf einen Prädiktor und geht von dem dichotomen Kriterium „erfolgreich / nicht erfolgreich“ aus. Der Nutzen, der mit dem Grad des Erfolges zunimmt, wird außer Acht gelassen (Schuler, 1996). Ebenso fehlt die Möglichkeit den Nutzen eines Testverfahrens in Geldeinheiten auszudrücken.

Genau um diesen monetären Aspekt erweiterten Brodgen (1949) sowie Cronbach und Gleser das Taylor-Russell Modell (Höft, 2001). Neben den bereits oben genannten Informationen (Grundrate, Selektionsrate und Validität) werden in diesem Modell ebenso Faktoren wie u.a. die Kosten des Auswahlverfahrens pro Bewerber und die geschätzte Verweildauer der Bewerber in einer Institution mit berücksichtigt.

Die so genannte Payoff-Funktion (Abb. 2.1-1) errechnet sich wie folgt:

$$\Delta U = N_E \cdot T \cdot SD_Y \cdot r_{xy} \cdot \bar{z}_x - C \cdot N_B$$

Abb. 2.1.1: Payoff-Funktion des Brodgen-Cronbach-Gleser-Modells (Höft, 2001)

Der Nutzenzuwachs der durch das Testverfahren entsteht wird mit ΔU bezeichnet und in Geldeinheiten angegeben. Größen, die in die Formel eingehen, sind die Anzahl der Eingestellten (N_E), die Anzahl der berücksichtigten Zeiteinheiten, welche die Bewerber schätzungsweise in Jahren in einer Institution verbringen (T), die Standardabweichung des Kriteriums in Geldeinheiten (SD_Y), der Validitätskoeffizient (r_{xy}), der durchschnittliche standardisierte Testwert (Prädiktorwert) der Ausgewählten ($\bar{z}_x$), die Kosten für das Testverfahren pro Anwendung (C) und die Anzahl der Bewerber (N_B). Ein Problem der Nutzenanalyse nach dem Brodgen-Cronbach-Gleser-Modell ist die Schätzung der Standardabweichung der Leistung (SD_y), wenn keine objektiven Daten, wie z.B. die Anzahl der verkauften Objekte, vorliegen.

Methoden zur Bestimmung der Standardabweichung sind bspw. Kostenrechnungsverfahren, globale Schätzverfahren, proportionale Regeln und individuelle Schätzungen (vgl. Höft, 2001).

Nach dem von Höft (2001) berechneten fiktiven Anwendungsbeispiel errechnet sich für die Verwendung eines Testverfahrens für die Auswahl von Bankkaufleuten im Vergleich zu einer bloßen Zufallsauswahl ein beachtlicher inkrementeller Nutzenzuwachs für die nächsten zehn Jahre von ca. 9½ Millionen DM. Ein weiteres anschauliches Beispiel zur Darstellung der Kalkulation mittels der Payoff-Funktion geben Barthel und Schuler (1989). Sie errechneten nach einem Modell von Boudreau (1983) für den Einsatz eines biographischen Fragebogens zur Auswahl von Außendienstmitarbeitern in einer Versicherungsgesellschaft ein Nutzen von immerhin netto DM 473.265. An dem von Höft (2001) definierten Quantitätsanteil - Produkt aus N_E und T -, ist weiter erkennbar, dass je mehr Personen ausgewählt werden und je größer ihre Verweildauer in der Organisation ist, desto höher ist auch der monetäre Nutzen des eingesetzten Verfahrens.
Neben dem wichtigen Beitrag der Nutzenfunktionen für die psychologische Diagnostik, nach der sich die Diagnostik auch „ rechnen “ muss, ergeben sich dadurch auch Schwierigkeiten (Amelang & Zielinski, 1994). Zusätzlich zu den Hindernissen in der Schätzung von einzelnen Parametern, wie z.B. der Standardabweichung des Kriteriums (vgl. auch Höft, 2001), fordern sie die individuelle Perspektive, die gegenüber der institutionellen vernachlässigt wird, stärker zu berücksichtigen. Denn auch der individuelle Nutzen und vor allem die Kosten, die für einen Sozialstaat entstehen, wenn der Einzelne bei Ablehnung seinem „Schicksal“ überlassen wird, sollte nach Meinung der Autoren Beachtung finden.

Menschen unterscheiden sich, wie in Abschnitt 2.1.1 bereits angesprochen, nicht nur hinsichtlich ihrer Fertigkeiten, Fähigkeiten und Kenntnisse, sondern auch hinsichtlich ihrer beruflichen Leistung. Aussagekräftige berufsdiagnostische Beurteilungen bilden diese interindividuellen Unterschiede ab und erlauben somit einer Institution die geeigneten Bewerber auszuwählen und die vorhandenen Mitarbeiter gemäß ihrem Potenzial adäquat einzusetzen. Valide Eignungsurteile sind somit grundlegend für den wirtschaftlichen Gewinn einer Organisation. Der Nutzen von Auswahlverfahren kann unmittelbar anhand der berechneten Kosten für die Ausbildung und der Vorhersageleistung der Prädiktoren für den Ausbildungserfolg bestimmt werden. Trotz dem der Ausbildungserfolg nicht mit dem Berufserfolg gleichgesetzt werden kann, ist der Ausbildungserfolg doch in vielen Berufen

eine notwendige Voraussetzung des zukünftigen Berufserfolgs. Bricht demnach eine Person vorzeitig die Ausbildung aufgrund schlechter Prüfungsergebnisse oder anderer Faktoren ab, entstehen erhebliche finanzielle Kosten.

Nach Kersting (2004) beträgt der finanzielle Aufwand für einen technisch wenig aufwendigen und relativ kurzen Ausbildungsgang in der Regel mindestens 50.000 Euro pro Person. Wenn man annimmt, dass ein Lehrgang 20 Auszubildende beinhaltet, entspräche dies einer Investition von einer Million Euro. Üblicherweise wird für eine effiziente Personalauswahl ein Prozentsatz von 5% der Investitionssumme, im beschriebenen Fall demnach 50.000 Euro, investiert. Die Personalauswahl wird in vielen Fällen jedoch eher kostengünstig und nach „bestem Wissen und Gewissen“ betrieben.

Unterschiede in den Ausbildungsleistungen von Personen, wie bspw. Ergebnisse in Zwischenprüfungen, sind erheblich leichter zu bestimmen und in Zahlen umzusetzen als die beruflichen Leistungen. Noch schwieriger ist es, die finanziellen Konsequenzen dieser beruflichen Leistungsunterschiede zu bestimmen. In manchen Berufen kann beispielsweise die Anzahl und der Wert von Vertragsabschlüssen, die Quantität und Qualität von hergestellten Produkten als Maß für die Arbeitsleistung herangezogen werden. In Organisationen, bei denen sich der finanzielle Nutzen nicht über objektive Kennzahlen darstellen lässt, wie z.B. bei der Polizei, wird häufig anhand von Expertenratings eingeschätzt, wie viel Material und Personal ein durchschnittlicher Mitarbeiter für die Bewältigung einer definierten Aufgabe benötigt. Nach Kersting (2004) zeigt sich, dass überdurchschnittliche Mitarbeiter ungleich weniger Personal und Material für die Bewältigung einer Aufgabe benötigen als schwache Mitarbeiter. Eine Analyse der Kosten und Nutzen eines Verfahrens basiert auf der Grundlage unterschiedlicher Leistungen der Mitarbeiter, diese Leistungsunterschiede werden anhand des Kennwertes der „Standardabweichung“ quantifiziert. Bestünden keinerlei Unterschiede in den Leistungen, die Standardabweichung wäre dementsprechend gleich Null, wäre die Durchführung eines Auswahlverfahrens überflüssig.

Um die Variabilität der Leistungen der einzelnen Mitarbeiter bestimmen zu können, ist es erforderlich, deren Output zu bestimmen. Alle vom Mitarbeiter erbrachten Leistungen, wie z.B. die Fertigstellung von Produkten oder die Abgabe von Dienstleistungen, müssen

bestimmt werden, so dass zwischen der Leistung unterdurchschnittlicher, durchschnittlicher und leistungsstarker Mitarbeiter unterschieden werden kann.

Nach bisherigen Forschungsergebnissen (vgl. Schmidt, Hunter & Pearlman, 1982) ist der jährliche Output (Fertigstellung von Produkten oder Abgabe von Dienstleistungen) eines durchschnittlichen Mitarbeiters in der Regel doppelt so hoch wie sein jährliches Entgelt. Dabei bezieht sich der beschriebene Unterschied auf die folgende Einteilung: Die schwächsten Mitarbeiter sind die unteren 15 %, während die überdurchschnittlichen Mitarbeiter die oberen 15 % der Verteilung abschneiden. Die Standardabweichung der in Geldwert umgerechneten Leistung zwischen überdurchschnittlichen und durchschnittlichen Mitarbeitern beträgt zwischen 40 % und 70% des jährlichen Entgelts (Kersting, 2004).

Ein effizientes Auswahlverfahren unterscheidet nicht nur geeignete von nicht geeigneten Bewerbern (hohe Qualität), sondern ist auch hinsichtlich der anfallenden Kosten für Durchführung, Auswertung etc. angemessen. Jedes Auswahlverfahren verursacht Kosten, egal ob es sich um ein hauseigenes oder ein externes Verfahren handelt. Beispielsweise kostete bei Veröffentlichung der DIN 33430 eine Testung mit 30 Personen mit einem Verfahren inklusive Auswertung, Interpretation sowie Materialkosten und Kosten für die Testdurchführung um die 1500 Euro (Kersting, 2004). Interessanterweise werden derartige Angebote in vielen Fällen als zu kostenintensiv abgelehnt. Alternativ werden dann oft Vorstellungsgespräche geführt. Bei Berücksichtigung von Fremdkosten, Zeitaufwand und den damit verbundenen Arbeitszeitausfall sowie den verwaltungstechnischen Aufwand, ist oft eine mehrstufige Strategie einer einstufigen vorzuziehen. Rechnet man bei Durchführung eines Vorstellungsgespräches, als einstufige Maßnahme, bspw. pro Gespräch inkl. Vorbereitung und Auswertung eine minimale Zeit von 30 Minuten, ergibt sich für 30 Bewerber ein aufzubringendes Zeitvolumen von $30 * 30 = 900$ Minuten, bzw. zwei Arbeitstage. Grundsätzlich gilt, dass Gruppenverfahren kostengünstiger sind als Einzelverfahren und gute Testverfahren in der Regel mehr Informationen pro Zeiteinheit erbringen als mündliche Verfahren und dadurch zu geringeren Kosten führen.
Bei der Wahl eines Verfahrens sollte jedoch primär der qualitative Aspekt eine Rolle spielen, denn auch wenn ein Verfahren kostengünstig ist, erbringt es nicht notwendigerweise auch eine valide Vorhersage des beruflichen Erfolges. In vielen Fällen sind die empirischen Gütekriterien Objektivität, Reliabilität und Validität, welche die Grundlage jeglicher Entscheidung für oder gegen ein Auswahlverfahren sein sollten, oft nicht bekannt oder

werden gegenüber ökonomischen Bewertungen hinsichtlich der Testdurchführung zurück gestellt (Steck, 1997).

In Tabelle 2.1.5 sind die Einsatzhäufigkeiten und die Gültigkeit einiger personalpsychologischer Auswahlverfahren aufgeführt (Kersting, 2004). Siehe hierzu auch Kapitel 2.3 zur Validität eignungsdiagnostischer Instrumente.

Tab. 2.1-5: Einsatzhäufigkeit und Gültigkeit eignungsdiagnostischer Verfahren (Kersting, 2004)

AUSWAHLVERFAHREN	EINSATZHÄUFIGKEIT	GÜLTIGKEIT
Arbeitsprobe	44 %	.54
Intelligenztest	34 %	.51
Strukturiertes Eignungsinterview	70 %	.51
Unstrukturiertes Eignungsinterview	57 %	.38
Assessment Center	39 %	.37
Referenz	71 %	.26
Bewerbungsunterlagen	98 %	keine Angabe
Medizinische Begutachtung	64 %	keine Angabe
Gruppengespräch	51 %	keine Angabe
Leistungstest	47 %	keine Angabe

Wie den Angaben aus der Tabelle 2.1-5 zu entnehmen, steht die Häufigkeit des Einsatzes der Verfahren nicht unbedingt in Abhängigkeit ihrer Vorhersageleistung. Zu ähnlichen Ergebnissen kommen auch andere Untersuchungen (vgl. a. Fruhner, Schuler, Funke & Moser, 1991; Scheinecker & Wallner, 2003; Steck, 1997). Klassisch scheint hierbei die Vorliebe für das unstrukturierte Einstellungsinterview im Gegensatz zum Intelligenztest, obwohl es eine geringere prognostische Validität besitzt. Auch die Auswahl von Personen aufgrund ihrer Bewerbungsunterlagen findet in 98% der Fälle statt, obwohl die Forschung zur klinischen Urteilsbildung zeigt, dass derartige Interpretationen mit erheblichen Reliabilitätsmängeln behaftet sind, was ebenso zu eher niedrigen Validitätswerten führt (Schuler & Marcus, 2001). Die Einbeziehung der empirischen Gütekriterien der Auswahlverfahren, soweit vorhanden, sollte im Sinne von Kosten-Nutzen Abwägungen eine größere Rolle spielen als es bisher in praxi der Fall ist.

Anhand eines von Kersting (2004) verwendeten Beispieles können die einzelnen Parameter der Kosten-Nutzen Funktion differenziert betrachtet werden. In seiner Darstellung wurde der Nutzen von standardisierten und unstandardisierten Einstellungsinterviews miteinander verglichen. Eine Organisation will zur Einstellung von zwei neuen Mitarbeitern insgesamt 14 Bewerber einladen. Das Jahresentgelt der eingestellten Person beträgt 50000 Euro. Entsprechend den Erklärungen zur Berechnung der Variabilität der Berufsleistung der

Mitarbeiter und dessen Berechnung, wird von einem durchschnittlichen jährlichen Output von 100000 Euro pro Mitarbeiter ausgegangen (jährlicher Wert eines durchschnittlichen Mitarbeiters ist in der Regel doppelt so hoch wie sein jährliches Entgelt). Überdurchschnittliche Mitarbeiter leisten im Vergleich zum Durchschnitt mindestens 40 % mehr an Output, so dass die geschätzte Standardabweichung der Berufsleistung auf 40000 Euro beziffert wird. Es wird außerdem davon ausgegangen, dass die eingestellten Personen 10 Jahre in der Institution verbleiben und im Auswahlverfahren einen Wert erreichen, der eine Standardabweichung über dem Gruppendurchschnitt liegt. Tabelle 2.1-6 gibt die beschriebenen Werte wieder.

Tab. 2.1-6: Nutzen zweier verschiedener Personalauswahlverfahren (Kersting, 2004)

Parameter der Kosten-Nutzen-Funktion		Unstandarisiertes Eignungsinterview	Standarisiertes Eignungsinterview
N_A	Anzahl der ausgewählten Bewerber	2	2
T	Verweildauer in Jahren	10	10
R_{XY}	Gültigkeit des Verfahrens	.38	.51
Z_X	erzielter durchschnittlicher Prädiktorwert	1	1
SD_Y	Standardabweichung der Berufsleistung in Euro	40.000	40.000
C	Kosten des Verfahrens	250	500
N_B	Anzahl der untersuchten Personen	14	14
Nutzen	Zusätzlicher Nettonutzen im Vergleich zur Zufallsauswahl (in Euro)	300.500	401.000

Das Beispiel zeigt, dass sich allein durch die Verwendung eines standardisierten Einstellungsinterviews im Vergleich zum unstandardisierten Interview, welches vermeintlich als das kostengünstigere erscheint, ein zusätzlicher finanzieller Nutzen von 105.500 Euro ergibt.

Stehle und Barthel (1984) untersuchten den Nutzen der Einführung eines Assessment Center gegenüber der Anwendung unstandardisierter Eignungsinterviews für den Bereich der Einstellung von Führungsnachwuchskräften für die chemische Industrie. Für die Einstellung von 12 Hochschulabsolventen für ein Traineeprogramm ergaben sich unter den von den Autoren gesetzten Bedingungen durch die Einführung eines Assessment Centers Einsparungen von einer halben Million Euro. Schuler, Funke, Moser und Donat (1995) analysierten den langfristigen Nutzen der Einführung eines umfassenden Auswahlverfahrens zur jährlichen Einstellung von 25 Personen im Bereich Forschung und Entwicklung mit circa 3,8 Millionen Euro. Barthel und Schuler (1989) haben im Bereich der Auswahl von Außendienstmitarbeitern in einer Versicherungsgesellschaft den Nutzen eines weiteren prädiktiv validen eignungsdiagnostischen Instrumentes belegen können. Für den Einsatz eines

biographischen Fragebogens errechneten die Autoren aufgrund der Kalkulation nach einem Modell von Boudreau (1989) ein Nutzensbetrag von DM 473.265. Welche Kosten durch Kündigungen auf ein Unternehmen zu kommen, die häufig eine Ursache falscher Personalentscheidungen bei der Einstellung von nicht geeigneten Mitarbeitern sein können, belegte Cascio (1991). Als Grundlage seiner fiktiven Berechnung wurde ein 200-Betten-Krankenhaus mit 1200 Angestellten und einer monatlichen Kündigung von 2% herangezogen. Für jeden dieser Mitarbeiter entstand der Organisation ein monatlicher Unkostenbetrag von 7426 US Dollar.

In der heutigen Zeit werden jährlich Millionen von Personalentscheidungen getroffen, wobei deren Qualität sicherlich genauso unterschiedlich ist wie die Verfahren, mittels derer diese Entscheidungen getroffen werden. In sicherlich nicht wenigen Fällen werden personelle Entscheidungen sogar noch auf Grundlage der „Intuition“ oder des „persönlichen ersten Eindruckes“ getroffen, ohne auf bewährte und empirisch belegte diagnostische Methoden zurückzugreifen. Jedoch scheint sich in diesem Bereich der Schwerpunkt von der Bestenauslese auf die Methode der Klassifikation zu verschieben, also der optimalen Verteilung der Arbeitskräfte auf die bestehenden Ressourcen, sowie die Identifizierung der gänzlich Ungeeigneten.
In einer wissenschaftlich begründeten Personalauswahl stehen die Personalverantwortlichen vor der Aufgabe, ein ökonomisches aber gleichermaßen zuverlässiges Auswahlverfahren zusammenzustellen. Dabei ist zum einen zu bedenken, dass das Verfahren in seiner Außenwirkung der Organisation und entsprechend seinem Aufwand der zu besetzenden Position angemessen ist.
Verfahren zur berufsbezogenen Eignungsbeurteilung gibt es sehr viele; diejenigen die empirischen Bewährungskontrollen unterzogen wurden und somit nachweisbare Beweise für ihre Treffsicherheit und prognostische Validität besitzen, dagegen nur in Relation wenige. Es gibt sogar Institutionen, die seit Jahren Verfahren zur Eignungsbeurteilung einsetzen, ohne die Aussagekraft und Gültigkeit je geprüft zu haben. Das subjektive Evidenzgefühl (Kersting, 2004) ist in diesem Fall entscheidend, obwohl dies irrational ist und wissenschaftlichen wie auch wirtschaftlichen Grundsätzen widerspricht. Empirische Kontrollen der Aussagekraft von Eignungsbeurteilungen sind der Garant dafür, dass Fehler im Verfahren rechtzeitig erkannt und beseitigt werden, dass Veränderungen schnell berücksichtigt werden können und dass die Verfahren somit kontinuierlich verbessert werden können. Folglich sind empirische Bewährungskontrollen eine wesentliche Voraussetzung für die Qualität der Verfahren und für

Verhaltensoptimierungen. Nach Hossiep (1995) wurden in den sechs einschlägigen deutschen Fachzeitschriften innerhalb eines Zeitraumes von 44 Jahren lediglich 82 Arbeiten zum Thema empirische Kontrollen der Gültigkeit berufseignungsdiagnostischer Verfahren veröffentlicht. Das diese Zahl im Vergleich zu dem geschätzten Einsatz von berufseignungsdiagnostischen Verfahren eine extrem geringe Zahl ist, ist offensichtlich.

Empirisch kontrollierte psychologische Testverfahren, welche speziell zur Ermittlung der Eignung für eine bestimmte Laufbahn, wie beispielsweise für den gehobenen Polizeivollzugsdienst konstruiert wurden, haben gegenüber konventionellen Auswahlverfahren (Zeugnisnoten, Leistungsbeurteilungen durch den Vorgesetzten, persönliches Gespräch) u.a. den Vorteil, dass hinterher ziemlich exakte Aussagen über das Ausmaß der Übereinstimmungen zwischen den Testergebnissen und dem Berufs- oder Ausbildungserfolg ermittelt werden kann.

Wesentlich festzuhalten ist, dass mit keinem noch so ausgeklügelten Testverfahren eine hundertprozentige perfekte Vorhersage der beruflichen Bewährung möglich ist. Dies hat mehrere Gründe, von denen Althoff (1977) zwei der wichtigeren nennt: Generell erscheint es unmöglich, Testmethoden in der Art zu entwickeln, dass sie zum Zeitpunkt der Eignungsuntersuchung sämtliche Aspekte des Bewerbers erfassen können, welche spätere relevante Fähigkeiten und Persönlichkeitsmerkmale abbilden. Außerdem können Faktoren, die zum Zeitpunkt des Auswahlverfahrens noch nicht wirksam werden, wie z.B. familiäre Bedingungen, Krankheit, die berufliche Leistung beeinflussen. Wenn die Einstellungsentscheidung wesentlich auf den Ergebnissen eines Eignungstestes basiert, lässt sich der Zusammenhang zwischen Testergebnissen und Erfolgskriterien nur für die Personen erschließen, die gute Testergebnisse erzielt haben und demnach eingestellt worden sind. Denn nur für diese Gruppe von Personen liegen auch Daten über deren berufliche Bewährung in einem Unternehmen vor, im Gegensatz zu Personen, die nicht durch das Einstellungsverfahren gekommen sind. Folglich entstehen erhebliche Varianzeinschränkungen (selegierte Stichprobe), die durch Wechsel und Entlassungen von den ursprünglich geeigneten Personen über die Zeit zu einer weiteren Einschränkung der Stichprobe führen. Dies ist bei der Bewertung eines empirisch durchgeführten Vergleiches von Testergebnissen und Kriterien des Berufserfolges zu berücksichtigen.

Weiter ist bei der Beurteilung der Güte von Testverfahren zu bedenken, dass Prädiktoren nur befriedigende und vor allem valide Aussagen machen können, wenn die Erfolgskriterien ebenfalls valide sind (siehe auch Kapitel 2.3).
Folglich lässt sich die tatsächliche Qualität eines Testverfahrens erst dann adäquat einschätzen, wenn ebenso zuverlässige und gültige Kriterien für die berufliche Bewährung vorliegen. Aus einem geringen Zusammenhang zwischen Testergebnis und Erfolgskriterium lässt sich also per se nicht schließen, dass der Test untauglich ist, denn der geringe Zusammenhang könnte auch an invaliden Kriterien liegen.
Weiter ist bei der Beurteilung der Güte von Testverfahren zu bedenken, dass Prädiktoren nur befriedigende und vor allem valide Aussagen machen können, wenn die Erfolgskriterien ebenfalls valide sind (siehe auch Kapitel 2.3).
Folglich lässt sich die tatsächliche Qualität eines Testverfahrens erst dann adäquat einschätzen, wenn ebenso zuverlässige und gültige Kriterien für die berufliche Bewährung vorliegen. Aus einem geringen Zusammenhang zwischen Testergebnis und Erfolgskriterium lässt sich also per se nicht schließen, dass der Test untauglich ist, denn der geringe Zusammenhang könnte auch an invaliden Kriterien liegen.

2.2 Personalauswahl am Beispiel des Polizeivollzugsdienstes

Vor der Beschreibung eines Auswahlverfahrens für den Polizeivollzugsdienst gilt folgendes festzuhalten: Die Regelung der Einstellung in den Polizeivollzugsdienst ist Ländersache. Somit sind die Voraussetzungen für eine Einstellung in den Polizeivollzugsdienst wie auch die jeweiligen Einstellungsverfahren in jedem Bundesland unterschiedlich. Auch wenn verschiedene Parallelen erkennbar sind, gilt das an dieser Stelle beispielhaft dargestellte Eignungsfeststellungsverfahren allein für das Bundesland Hamburg.

2.2.1 Allgemeine Grundlagen für die Einstellung

Übergeordnete Richtlinien für die Auswahl von Bewerbern, die sich um eine Einstellung in den mittleren oder gehobenen Polizeivollzugsdienst bemühen, sind in der Regel durch Gesetze und Verordnungsblätter festgelegt und organisiert. Nach den Vorschriften dieser Verordnungen steht jedem Polizeivollzugsbeamten entsprechend seiner Eignung, Befähigung und fachlichen Leistung der Aufstieg in alle Ämter des Polizeivollzugsdienstes offen. Dabei

umfasst der Polizeivollzugsdienst die Dienstzweige der Schutzpolizei, der Kriminalpolizei und der Wasserschutzpolizei.
Konkret wird die Einstellung in den Polizeivollzugsdienst durch die §§ 7 HmbLVOPol (Hamburger Laufbahnverordnung für Polizeivollzugsbeamte) für den mittleren Dienst und 17 für den gehobenen Dienst geregelt. Theoretisch wäre auch eine direkte Einstellung in den höheren Dienst nach § 24 HmbLVOPol möglich, allerdings wird dieser „Direkteinstieg" derzeit nicht praktiziert und ist deswegen auch nicht durch ein weitergehendes Auswahlverfahren geregelt. Zurzeit können nur Beamte des gehoben Polizeivollzugsdienstes über ein internes Auswahlverfahren nach § 22 HmbLVOPol für die Ausbildung im höheren Polizeivollzugsdienst ausgewählt werden. Beamte des mittleren Dienstes haben nach § 15 HmbLVOPol die Möglichkeit, sich in einem Auswahlverfahren für die Ausbildung zum gehobenen Dienst – Studium an der Fachhochschule für Öffentliche Verwaltung – zu qualifizieren.
In der Laufbahnverordnung werden weiter die formellen Voraussetzungen und Rahmenbedingungen hinsichtlich Alter, Schulbildung, Teilnahme an einer Einstellungsprüfung, Vorliegen einer gesundheitlichen Polizeivollzugstauglichkeit, Ausnahmen u.ä. für eine Einstellung in den Polizeivollzugsdienst durch politischen Beschluss geregelt.

Auf Basis der gesetzlichen Grundlagen wird die Einstellung von Bewerbern für den mittleren und gehobenen Dienst in den „Richtlinien des Polizeivollzugsdienstes" praktisch geregelt. In diesen Verordnungen wird der Ablauf des Einstellungsverfahrens - Fristen und Inhalt der Bewerbung, Ausschluss vom Verfahren, Inhalte des Testverfahrens zur Eignungsfeststellung, Wiederbewerbungsmöglichkeiten - verfügt. Weiter werden hier Detailfragen zur Vorauswahl, Bewertung der Testergebnisse, zu erbringende Mindestanforderungen u.ä. präzisiert und konkretisiert. Wie im Öffentlichen Dienst obligatorisch, sind bei Veränderungen gemäß Personalvertretungsgesetz die Personalräte zu beteiligen.

Nach der Verordnung über die Laufbahn der hamburgischen Polizeivollzugsbeamten (HmbLVOPol) gliedert sich die Einheitslaufbahn der Polizeivollzugsbeamten in die Laufbahnabschnitte I bis III.
Der Laufbahnabschnitt I beschreibt den mittleren Polizeivollzugsdienst und umfasst theoretisch die Ämter vom Polizeiwachtmeister bis zum Polizeihauptmeister. Die Einstellung in die Kriminalpolizei ist im Laufbahnabschnitt I nicht vorgesehen, d.h. in den mittleren

Dienst werden nur Bewerber für den Schutzpolizeivollzugsdienst und den Dienst bei der Wasserschutzpolizei eingestellt. Der gehobene Polizeivollzugsdienst erstreckt sich auf den Laufbahnabschnitt II undbildet sich in den Ämtern vom Polizei- oder Kriminalkommissar bis zum Ersten Polizei- oder Ersten Kriminalhauptkommissar ab. Schließlich erfasst der Laufbahnabschnitt III den höheren Polizeivollzugsdienst für Schutz-, Wasserschutz- und Kriminalpolizei. Hierzu zählen die Ämter des Polizei- oder Kriminalrates bis zum Leitenden Polizei- bzw. Kriminaldirektor.

2.2.2 Einstellung in den gehobenen Polizeivollzugsdienst

Im dritten Abschnitt des HmbLVOPol (§§ 15-21) werden die Modalitäten für die Einstellung in den Laufbahnabschnitt II des Polizeivollzugsdienstes geregelt. Wie dargestellt, können sowohl Beamte des mittleren Polizeivollzugsdienstes durch ein Auswahlverfahren für den gehoben Dienst ausgewählt, wie auch externe Bewerber direkt in den gehoben Polizeidienst eingestellt werden. In der vorliegenden Arbeit wurde ausschließlich das Verfahren für die Einstellung von externen Bewerbern berücksichtigt. Das im Auswahlverfahren erreichte Gesamtergebnis (Formelle Voraussetzungen, Testergebnis, gesundheitliche Eignung, charakterliche Eignung) entscheidet letztendlich über die abschließende Zulassung zum Studium an der Fachhochschule für Öffentliche Verwaltung (FHÖV), Fachbereich Polizei.

Zu den formellen Voraussetzungen für die Zulassung als „Direkteinsteiger“ in den gehobenen Dienst zählen:

Bewerber dürfen das 25. Lebensjahr noch nicht vollendet haben oder das 32., wenn sie eine abgeschlossene oder in wesentlichen Teilen vollendete Berufsausbildung, Dienstzeit in der Bundeswehr von mindestens vier Jahren, Kinderbetreuungszeit von mindestens 4 Jahren oder besondere Fähigkeiten und Kenntnisse für den Polizeivollzugsdienst nachweisen können.

Weiter muss der Nachweis einer zu einem Hochschulstudium berechtigten Schulbildung oder einen von der zuständigen Behörde als gleichwertig anerkannten Bildungsstand vorliegen (Fachhochschulreife bzw. Abitur).

Abschließend ist der erfolgreiche Abschluss der Einstellungsprüfung sowie der Nachweis der gesundheitlichen Tauglichkeit durch den Personalärztlichen Dienst zu bescheinigen.

Der Vorbereitungsdienst dauert nach § 18 der HmbLVOPol drei Jahre und sechs Monate. Er gliedert sich in eine sechsmonatige Praxiseinweisung auf einem Polizeikommissariat und einem dreijährigen Studium an der Fachhochschule für Öffentliche Verwaltung, Fachbereich Polizei. Nach erfolgreicher Ausbildung wird die Laufbahnprüfung II abgelegt und bei Bestehen werden die Anwärter unter Berufung auf das Beamtenverhältnis auf Probe zum Polizeikommissar oder zum Kriminalkommissar ernannt.
Zwar können Bewerber sich mit ihrer Bewerbung konkret für einen Dienstzweig bewerben (z.B. Kriminalpolizei oder Schutzpolizei), jedoch entscheidet die zuständige Behörde vor Beginn der Ausbildung entsprechend dem dienstlichen Bedarf über die Zuweisung der Beamten zu den einzelnen Dienstzweigen. Soweit dabei für die Einstellung in den Dienstzweig der Wasserschutzpolizei keine Bewerber zur Verfügung stehen, können Beamte der anderen Dienstzweige zur Ausbildung zugelassen werden.

2.2.3 Auswahlverfahren für den gehobenen Polizeivollzugsdienst

Das Ziel des Auswahlverfahrens ist die Feststellung der geistigen, körperlichen und persönlichen Eignung der Bewerber für die Anforderungen des Polizeivollzugsdienstes, die Feststellung ihrer gesundheitlichen Tauglichkeit für die Tätigkeit sowie eine bedarfsorientierte Bestenauslese, wenn mehr geeignete Bewerber vorliegen als Studienplätze angeboten werden können. Der Personalbedarf richtet sich nach der mittelfristigen Personalbedarfsplanung und wird von der Landespolizeiverwaltung unter Berücksichtigung der Anforderungen des polizeilichen Vollzugsdienstes für jeden Dienstzweig einzeln erhoben.

2.2.3.1 Schriftliche Bewerbungen (Vorauswahl)

Jeder Bewerber für ein Studium an der Fachhochschule hat seine Bewerbung schriftlich zu formulieren. Die Bewerbung muss dabei folgendes enthalten bzw. beachten:

- einen vollständig ausgefüllten standardisierten Bewerbungsbogen mit Anlagen und einem individuell zu gestaltenden Anschreiben,
- einen Lebenslauf mit einem aktuellen Foto,
- eine Fotokopie des Abschlusszeugnisses einer allgemeinbildenden Schule,
- falls vorhanden, einen Nachweis über eine abgeschlossene oder in wesentlichen Teilen abgeleistete Berufsausbildung,

- ein ärztliches Attest über den allgemeinen Gesundheitszustand,
- bei vorhandener Sehhilfe einen Befund des Augenarztes,
- einen Nachweis über die Fähigkeit schwimmen zu können,
- einen Nachweis über den Besitz des Führerscheins Klasse 3,
- soweit vorhanden Arbeitszeugnisse früherer Arbeitgeber.

Weiter muss die Bewerbung innerhalb einer bestimmten Frist eingehen und die Körpergröße eines Bewerbers für die Schutzpolizei darf nicht unter 160 cm liegen.

Bewerber für die Wasserschutzpolizei haben darüber hinaus zusätzlich geforderte Befähigungsnachweise (seemännische Patente) einzureichen.
Vor Einladung zum Auswahlverfahren wird die Erfüllung der genannten Voraussetzungen bei den Bewerbern überprüft und ggf. über mögliche Ausnahmegenehmigungen, d.h. wenn ein besonderes dienstliches Interesse an der Einstellung vorliegt, entschieden.

Erfüllen die Bewerber ansonsten die genannten Voraussetzungen nicht oder besteht kein Bedarf, im Einzelfall eine Ausnahmegenehmigung zu erteilen, wird die Bewerbung abgelehnt. Liegen dennoch mehr Bewerbungen vor als Bewerber zum Auswahlverfahren eingeladen werden können, kann zusätzlich aufgrund von dokumentierten Schulnoten in vorliegenden Zeugnissen und sonstigen biographischen Daten (z.B. Ausbildungsverlauf) eine weitere Vorauswahl getroffen werden.

2.2.3.2 Auswahlverfahren

Alle Bewerber, welche die grundsätzlichen Voraussetzungen für die Einstellung erfüllen oder denen eine Ausnahmegenehmigung erteilt wurde, werden anschließend zum Auswahlverfahren eingeladen.

Das Auswahlverfahren für den gehobenen Polizeivollzugsdienst unterteilt sich in drei Abschnitte: Der erste Abschnitt umfasst das eigentliche Auswahlverfahren, mit dem festgestellt werden soll, ob die Bewerber die erforderlichen „geistigen, körperlichen und persönlichen Voraussetzungen“ für den Polizeivollzugsdienst besitzen. Dabei werden die Leistungen der Teilnehmer in den unterschiedlichen Testverfahren anhand einer neunstufigen Skala bewertet. Ein Punktewert von eins bildet die schlechteste und ein Punktewert von neun die beste Bewertung ab. Die Transformation der einzelnen Bewertungen (Rohwerte) auf die

neunstufige Skala erfolgt in Abhängigkeit von den Vorgaben des jeweiligen Testverfahrens. Zu beachten ist hierbei, dass das Auswahlverfahren grundsätzlich im Sinne einer sequentiellen pre-reject Auswahlstrategie durchgeführt wird. Dies bedeutet, dass das erste Auswahlinstrument allen am Tag eingeladenen Probanden gemeinsam vorgegeben wird. Diejenigen Teilnehmer, welche nicht einen Mindestwert von vier Punkten erreichen, scheiden aus dem weiteren Verfahren aus und werden verabschiedet. Die verbleibenden Bewerber bekommen den nächsten Testabschnitt zur Bearbeitung präsentiert. Diese Prozedur wiederholt sich für alle folgenden Komponenten des Auswahlverfahrens.

Die Inhalte des ersten Abschnittes des Auswahlverfahrens sollen unterschiedliche, für den Polizeivollzugsdienst berufsrelevante Leistungsmerkmale, registrieren. Konkret fanden zum Zeitpunkt der Untersuchung folgende Instrumente Verwendung: ein Lückendiktat, die Anfertigung eines hinsichtlich Inhalt und Deutschleistung bewerteten Berichtes sowie zwei unterschiedliche standardisierte Leistungstests zur Erfassung der kognitiven Leistungsfähigkeit.
Bei dem Lückendiktat handelt es sich um einen vom Tonband abgespielten standardisierten Text mit Wortlücken, welche vom Bewerber orthographisch fehlerfrei zu ergänzen sind. Die dokumentierten Rechtschreibfehler werden gezählt und einer neunstufigen Punkteskala zugeordnet. Für die Berichtsanfertigung wird den Teilnehmern eine Bildergeschichte mit sechs Fotos vorgelegt, aus denen unterschiedliche Abläufe eines Tatbestandes erkennbar sind. Aufgabe der Bewerber ist es dabei, mit eigenen Worten einen wahrgenommenen Tatablauf zu rekonstruieren und zu beschreiben. Neben der auch hier bewerteten orthographischen und grammatikalischen Leistung, werden die Anzahl der wiedergegeben standardisierten inhaltlichen Details erfasst. Die Bewertung erfolgt hier durch einen Deutschlehrer des Ausbildungsinstitutes und die Ergebnisse werden auf eine neunstufige Skala übertragen. Im dritten Testverfahren zur Prüfung der „geistigen Eignung“ wird mit Hilfe eines kulturfairen Intelligenztestes das allgemeine intellektuelle Leistungsniveau abgebildet. Die Einbindung eines kulturfairen nicht sprachgebundenen Intelligenztestes im Auswahlverfahren begründet sich durch die Vorgabe, auch ausländische Bewerber für den Polizeivollzugsdienst zu gewinnen. Ziel dabei ist es, durch einen höheren Anteil ausländischer Polizeivollzugsbeamter eine größere Anzahl unterschiedlicher ethnischer Gruppen in der Gesellschaft anzusprechen und diese verstehen zu können. Somit soll letztlich eine stärkere und effizientere Ansprechbarkeit und Mobilität der Polizei gewährleistet werden. Bewerber ausländischer Herkunft sind jedoch in der Regel mit einer anderen Muttersprache aufgewachsen, auch wenn

viele von ihnen bereits in Deutschland geboren oder zumindest hier zur Schule gegangen sind. Demzufolge haben sie vergleichsweise mehr Schwierigkeiten, den Anforderungen eines sprachgebundenen Intelligenztestes zu entsprechen. Eine Minimierung möglicher Benachteiligung der ausländischen Mitbewerber und diese gleichzeitig für den Polizeivollzugsdienst zu gewinnen, rechtfertigte daher den Einsatz eines kulturfairen Intelligenzverfahrens. Zur Bewertung herangezogen werden die in Stanine-Werte transformierten Rohwerte.
Allerdings erfolgt im Anschluss ebenfalls die Durchführung eines standardisierten nicht kulturfairen Testverfahrens, dessen Stanine-Wert ebenfalls in das Gesamtergebnis einfließt.

Die Prüfung der „körperlichen Eignung“ bildet den zweiten Teil des ersten Abschnittes. Hierbei soll von den Bewerbern in fünf verschiedenen Sportübungen auf der Grundlage bundeseinheitlicher Vorgaben ihre körperliche Leistungsfähigkeit nachgewiesen werden. Die fünf Übungen und deren jeweilige Bewertung sind in Tabelle 2.2-1 dargestellt.

Tab. 2.2-1: Übungen zur Erfassung der körperlichen Leistungsfähigkeit

ÜBUNG	PRÜFUNG	ABLAUF	BEWERTUNG
Cooper-Test	Aerobe und anaerobe Ausdauer und Willenskraft.	Die Testperson versucht auf einer Rundstrecke, die alle 50 m durch Pfosten markiert ist, in der Zeit von 12 Minuten eine möglichst weite Strecke durch Laufen zurückzulegen.	Bewertet wird die innerhalb von 12 Minuten zurückgelegte Strecke in Metern.
Wendelauf	Schnelligkeit, Schnellkraft, Beweglichkeit und Gewandtheit.	Zwei Kästen stehen auf einer Linie 10 m auseinander. Die Strecke ist viermal zu durchlaufen.	Bewertet wird die benötigte Zeit nach dem 4. Durchgang in Sekunden und Zehntelsekunden.
Klimmzüge	Kraft und lokale Kraftausdauer	Männer: Klimmzüge aus Streckhang. Frauen: Klimmzüge im Liegehang.	Bewertet wird die Anzahl der Klimmzüge.
Standweit-sprung	Sprungkraft, Schnellkraft und Koordination.	Aus Hockstellung im Schlusssprung soweit wie möglich nach vorne.	Von drei Versuchen wird jeweils der weiteste in Zentimetern gewertet.
Kasten-Bumerang-Test	Gewandtheit, Gleichgewichtssinn, Kopplungs- und Orientierungsfähigkeit.	Die Testperson beginnt mit einer Rolle vorwärts, läuft hinter einem Medizinball 90° nach rechts, überspringt das dort stehende Kastenteil und durchkriecht dasselbe zurück. Nachdem der Medizinball wieder in einer 90°-Rechtsdrehung umlaufen worden ist, sind in der gleichen Abfolge zwei andere Kastenteile zu überspringen und zu durchkriechen. Anschließend führt der Weg um den Medizinball zurück ins Ziel.	Bewertet wird die benötigte Zeit in Minuten und Sekunden.

Die durch die Bewerber erzielten Rohwerte werden auf eine neunstufige Skala übertragen. Die zu erreichende Mindestleistung entspricht dabei dem 4er-Wert. Allerdings ist dann ein Ausgleich einzelner Übungen möglich, wenn insgesamt eine Punktzahl von 20 erreicht wurde

(Durchschnittswert = 4) und dabei kein einzelner Wert unter zwei liegt. Bei der Bewertung des Sporttestes werden geschlechts- und altersspezifische Normen herangezogen.

Im zweiten Abschnitt erfolgen die Untersuchung der gesundheitlichen Tauglichkeit sowie die Überprüfung der so genannten „charakterlichen Eignung“ des Bewerbers für den Polizeiberuf. Alle bisher erfolgreichen Bewerber, die aufgrund ihrer erreichten Ergebnisse grundsätzlich für eine Einstellung in Betracht kommen, werden nach bundeseinheitlicher Vorschrift ärztlich hinsichtlich ihrer Polizeitauglichkeit untersucht. Weiter werden im Rahmen der „charakterlichen Eignung“ durch polizeiliche Ermittlungen Tatsachen (z.B. polizeiliche oder staatsanwaltschaftliche Ermittlungsergebnisse) hinterfragt, die gegen eine Einstellung in den Polizeivollzugsdienst sprechen könnten. Hierzu werden Auskünfte über in der Vergangenheit möglicherweise begangenen Straftaten o.ä. beim Bundeszentralregister, den zuständigen Polizeidienststellen und den Landeskriminalämtern eingeholt.

Die Prüfung der „persönlichen Eignung“ ist der dritte und letzte Abschnitt des Auswahlverfahrens und besteht aus einem Vorstellungsgespräch und einem Rundgespräch. Hier ist das Ziel, dass der Bewerber vor einer fünf Personen umfassenden Kommission zeigt, dass er aufgrund seiner Persönlichkeit für die Einstellung in den gehobenen Polizeivollzugsdienst geeignet ist.
Das Vorstellungsgespräch wird als unstrukturiertes Gespräch mit der Kommission geführt, wobei es in der Regel 30 Minuten nicht überschreitet. Am Rundgespräch nehmen die Bewerber in Gruppen bis zu sechs Personen teil. Sie sollen anhand eines vorgegebenen Problems eine gemeinsame Lösung erarbeiten. Die Gesprächsdauer wird mit 60 Minuten festgelegt.
Im Vorstellungs- wie im Rundgespräch werden anhand einer 15-stufigen Skala verschiedene nicht näher operationalisierte Verhaltensdimensionen bewertet, wobei der Mittelwert der einzelnen Bewertungen der Kommissionsmitglieder den Gesamtpunktwert ergibt. Erst wenn die Bewerber in dieser zweiten Stufe mindestens 10 Punkte, aus beiden Gesprächen addiert, erreicht haben, erscheinen sie für die Einstellung in den gehobenen Polizeivollzugsdienst geeignet.
Nach Feststellung der Eignung werden die Bewerber für den Laufbahnabschnitt II als Polizeikommissar - oder Kriminalkommissaranwärter in den Vorbereitungsdienst eingestellt.
Zur Veranschaulichung ist das gesamte Auswahlverfahren noch einmal zusammenfassend in Tabelle 2.2-2 dargestellt.

Tab. 2.2-2: Zusammenfassung der Elemente des Auswahlverfahrens

ELEMENT	ERFASSTE MERKMALE	BEWERTUNG UND SKALA
Lückendiktat	• Rechtschreibefähigkeit	Anzahl Fehler auf neunstufiger Skala
Berichtsfertigung	• Inhaltliche Details • Rechtschreibung • Grammatik • Ausdruck	Mittelwert aus Deutschleistung und erwarteter inhaltlicher Details auf neunstufiger Skala
Kulturfairer Leistungstest	• Reihenfortsetzen • figurale Klassifikation • Matrizenaufgaben • topologische Schlussfolgerungen.	Gesamtergebnis auf Stanine-Skala
Nicht kulturfairer Leistungstest	• Textanalyse • Schätzen • Buchstabenreihen • Entscheidungskriterien	Gesamtergebnis auf Stanine-Skala
Sporttest	• Cooper-Test • Wendelauf • Klimmzüge • Standweitsprung • Kasten-Bumerang-Test	Mittelwert der einzelnen Übungen auf neunstufiger Skala
Vorstellungsgespräch	• Mündliche Ausdrucksfähigkeit • Urteilsvermögen • Argumentationsfähigkeit • Allgemeinwissen, Wissen um bedeutsame politische und gesellschaftliche Zusammenhänge im In- und Ausland	Mittelwert von Vorstellungs- und Rundgespräch aller Merkmale auf fünfzehnstufiger Skala
Rundgespräch	• Mündliche Ausdrucksfähigkeit • Urteilsvermögen • Argumentationsfähigkeit • Gruppenaktivität	
Endergebnis: Gesamtsummenwert aller Testteile		

2.2.3.3 Ausbildung zum gehobenen Polizeivollzugsdienst

Die Ausbildung für den gehobenen Polizeivollzugsdienst an der Fachhochschule für Öffentliche Verwaltung (FHÖV) dauert drei Jahre und gliedert sich in drei Studienabschnitte. Im ersten Studienabschnitt sollen grundlegende Fähigkeiten und Kenntnisse für den gehobenen Polizeivollzugsdienst erworben werden. Dieser endet nach zwei Semestern mit der Zwischenprüfung, deren Ergebnis zeigen soll, ob das Studium weiter erfolgreich fortgesetzt werden kann. Nach erfolgreichem Abschneiden in der Zwischenprüfung folgt ein berufspraktisches Studienhalbjahr, in welchem die Studenten einen Einblick in die Tätigkeitsfelder der Polizei gewinnen und ihre im Studium vermittelten theoretischen Kenntnisse in praktische umsetzen sollen. Der dritte und letzte Studienabschnitt endet nach zwei weiteren Semestern mit der Laufbahnprüfung II, mit welcher erfasst werden soll, ob der Student das Ziel der Ausbildung erreicht hat. Inhaltlich setzt sich das Studium aus Lehrveranstaltungen an der Fachhochschule für Öffentliche Verwaltung - Fachbereich Polizei - und Berufspraktika an den Polizeidienststellen zusammen. Die Studieninhalte bestehen aus Pflicht-, Wahlpflicht- oder Wahlfächern in unterschiedlichen Fachgebieten. Die Beamten der

Wasserschutzpolizei haben außerdem eine sechsmonatige Zusatzausbildung abzulegen. Tabelle 2.2-3 gibt eine Übersicht über die im Studium vermittelten Fachgebiete und deren Unterteilung in die einzelnen Fächer.

Tab. 2.2-3: Fachgebiete und Fächer im Studium des gehobenen Polizeivollzugsdienstes

FACHGEBIETE	FÄCHER
Polizei- und Kriminalwissenschaften	Kriminalistik Einsatzlehre Kriminologie Schifffahrtslehre / Gefahrgut / Umweltschutz Verkehrslehre / Verkehrsrecht Informationsverarbeitung
Rechtswissenschaften	Allgemeines Verwaltungsrecht Polizeirecht Recht des öffentlichen Dienstes Staatsrecht / Verfassungsrecht Straf-, Ordnungswidrigkeiten-, Bürgerliches Recht Strafverfahrensrecht
Organisations- und Gesellschaftswissenschaften	Politologie Führungslehre / Public Management Psychologie Soziologie
Einsatzausbildung	Selbstverteidigung Dienstausbildung Polizeitechnik Waffenausbildung

2.3 Zur Bedeutung der Validität in der Personalauswahl

Die allgemeinen Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen wurden 2002 vom Arbeitsausschuss 4.4 „Psychologische Eignungsdiagnostik“ im Normenausschuss Gebrauchstauglichkeit und Dienstleistungen (NAGD) erarbeitet und als DIN 33430 verabschiedet. Die hier festgelegten Leitsätze ergeben Hinweise für die sach- und fachgerechte Entwicklung von Auswahlverfahren, die dem Stand von Wissenschaft und Technik Rechnung tragen. Mit der Norm werden Qualitätskriterien und – standards für berufsbezogene Eignungsbeurteilungen sowie Qualitätsanforderungen für die an der Eignungsbeurteilung beteiligten Personen beschrieben. Wissenschaftlicher Hintergrund der DIN ist u.a. der Anspruch nach Erfüllung von gegebenen Gütekriterien eines Auswahlinstrumentes. In Anlehnung an Lienert (1989) soll ein gutes Auswahlinstrument objektiv, reliabel und valide sein.

2.3.1 Grundlagen und Formen der Validität

Die Validität eines eignungsdiagnostischen Instrumentes ist in der Personalauswahl das wichtigste Gütekriterium. Die Validität eines Auswahlverfahrens gibt den Grad der Genauigkeit an, mit dem ein Auswahlverfahren dasjenige Persönlichkeitsmerkmal oder diejenige Verhaltensweise, die es zu erfassen vorgibt bzw. erfassen soll, tatsächlich auch erfasst. Vollkommen valide ist ein Verfahren, wenn seine Ergebnisse einen unmittelbaren und fehlerfreien Rückschluss auf den Ausprägungsgrad des zu erfassenden Persönlichkeits- oder Verhaltensmerkmales zulassen.
Die einfachste Form der Validitätsprüfung besteht nach Kanning (2002) darin, per Augenschein zu überprüfen, ob ein bestimmter Test in etwa das erfassen kann, was er zu erfassen beabsichtigt. So wäre bspw. anzunehmen, dass ein Testverfahren zur Erfassung der Konzentrationsfähigkeit Aufgaben beinhaltet, bei denen die konzentrative Leistungsfähigkeit der Probanden unter bestimmten zeitlichen Bedingungen gemessen werden kann und keine, die Hinweise auf das Vorliegen von Persönlichkeitsstörungen erlauben. Damit wird deutlich, dass die hier vorliegende „face validity“ eine wichtige Grundlage für die weitere Bestimmung und zur Erschließung der Validität ist.

Nach Michel und Conrad (1982) erlauben die Ergebnisse eines Tests oder Fragebogens bei gegebener hoher Validität einen Schluss aus dem beobachteten Verhalten in der Testsituation auf Verhalten außerhalb der Testsituation. Zur Erschließung der Validität werden unterschiedliche Zugänge unterschieden. Dabei ist zwischen drei Hauptarten zu unterscheiden: der Kriteriums-, der Inhaltlichen- sowie der Konstruktvalidität. Die Validität eines Tests wird jedoch nicht direkt gemessen, sondern aus Messungen erschlossen (Jäger, 1986). Bei der Bestimmung der Inhaltlichen Validität oder Kontentvalidität ist der Test bzw. sind seine Elemente so beschaffen, dass sie das zu erfassende Persönlichkeitsmerkmal oder die in Frage stehende Verhaltensweise repräsentieren. Der Test selbst stellt das optimale Kriterium für das Persönlichkeitsmerkmal oder die Verhaltensweise dar. Inhaltliche Validität wird einem Test in der Regel durch ein Rating von Experten zugebilligt (Lienert & Raatz, 1994, S. 10 ff.).

Ob ein Test ein bestimmtes Konstrukt erfasst, wird aufgrund theoretischer Erwägungen und anhand sich daran anschließender empirischer Untersuchungen entschieden. Die Konstruktvalidität zielt direkt auf die psychologische Analyse der einem Test zugrunde

liegenden Eigenschaften und Fähigkeiten ab. Für die Erfassung der Inhaltlichen Validität sowie der Konstruktvalidität lässt sich zwar im Allgemeinen kein einzelner Parameter für den Grad der Validität ermitteln, jedoch liefert beispielsweise das Multi-Trait-Multi-Method-Verfahren (ursprünglich von Campell & Fiske, 1959, eingeführt) den diskriminanten und den konvergenten Validitätskoeffizienten, die vergleichend herangezogen werden könnte. Hierbei werden mehrere Konstrukte simultan mit mehreren Messungen erhoben. Für die Darstellung der konvergenten Validität wird gefordert, dass die Messungen eines Konstruktes über die unterschiedlichen Methoden hinweg hoch miteinander korrelieren, während die Messungen unterschiedlicher Konstrukte beim Einsatz desselben Messinstrumentes einen möglichst niedrigen Zusammenhang aufweisen sollten (diskriminante Validität).
Die Vorhersage von berufsrelevanten Variablen mit Hilfe unterschiedlicher Verfahren steht im Mittelpunkt der eignungsdiagnostischen Praxis. Der Verfahrenseinsatz hat demnach die Funktion eines Prädiktors, mit dem der berufliche Erfolg einer Person, das Kriterium, vorhergesagt werden soll. Die Untersuchung dieser Beziehung, die in den meisten eignungsdiagnostischen Untersuchungen die herausragende Rolle spielt, ist durch die Analyse der kriterienbezogenen Validität möglich. Bei diesem Aspekt der Validität werden die Testergebnisse einer Stichprobe von Probanden mit einem Außenkriterium, so z.B. die spätere berufliche Leistung dieser Stichprobe, korreliert. Vorteilhaft im Gegensatz zu den beiden oben genannten Aspekten der Inhaltlichen Validität und der Konstruktvalidität lässt sich hier ein einzelner Parameter als Maßzahl für den Grad der Validität ermitteln. Außerdem ist ein Vergleich der angegebenen prädiktiven Validitäten anderer eignungsdiagnostischer Instrumente in vielfältig vorliegenden Untersuchungen aus dem berufseignungsdiagnostischen Kontext möglich. Angesichts ihrer Bedeutung im Rahmen der Personalauswahl, wird auf die kriterienbezogene Validität unter Punkt 2.3.1.2 näher eingegangen.

2.3.1.1 Objektivität und Reliabilität

Ein weiteres Hauptgütekriterium eines Testverfahrens und damit auch eines Auswahlverfahren im Rahmen der wissenschaftlich gestützten Personalauswahl ist die Objektivität. Nach Lienert (1989) ist hierunter der Grad zu verstehen, in dem die Ergebnisse eines Auswahlverfahrens unabhängig vom Untersucher bzw. dem Personalverantwortlichen sind. Ein Auswahlverfahren wäre vollkommen objektiv, wenn unterschiedliche Untersucher bei demselben Bewerber zu gleichen Ergebnissen gelangten, d.h. ihn für die vakante Position als geeignet bzw. nicht geeignet beschreiben. Im Rahmen der DIN 33430 (Anforderungen an

Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen; 2002) wird gefordert, dass die zur Eignungsbeurteilung eingesetzten Verfahren eine größtmögliche Durchführungs-, Auswertungs- und Interpretationsobjektivität besitzen müssen. Die Verfahren, ihre Materialien und die dem Verfahren beigefügten Instruktionen für den Kandidaten sowie die Vorgehensweise bei der Eignungsbeurteilung müssen so beschaffen sein, dass die Ergebnisse so wenig wie möglich durch den Kandidaten selbst verfälscht werden können.

Neben der Validität kann auch die Reliabilität nach Lienert und Orlik (1965) als ein Sonderfall experimenteller Präzision verstanden werden. In der DIN 33430 wird gefordert, dass die eingesetzten Auswahlverfahren eine der jeweiligen Art des Verfahrens und der angestrebten Aussage entsprechend hohe Zuverlässigkeit aufweisen müssen. Der erforderliche Grad der Zuverlässigkeit richtet sich u.a. nach dem untersuchten Merkmal, der Bedeutsamkeit der angestrebten Entscheidung sowie den jeweiligen Anwendungs- und Untersuchungsbedingungen. Sofern der Ausprägungsgrad von Personenmerkmalen aufgrund von mündlich gewonnenen Informationen bzw. Verhaltensbeobachtungen eingeschätzt wird, ist dabei sicherzustellen, dass verschiedene Beurteiler bei gleicher Beobachtungsgrundlage möglichst übereinstimmen. Dabei ist zu dokumentieren, nach welchen Gesichtspunkten, in Bezug auf die Zuverlässigkeit, die Verfahren zur Eignungsbeurteilung ausgewählt wurden.

Allgemein ist unter Reliabilität das Verhältnis von systematischer, d.h. theoretisch erklärter Varianz, zur Fehlervarianz, d.h. durch die jeweils betrachteten Bedingungen nicht erklärte Varianz, zu verstehen (Feger,1983). Im Reliabilitätskoeffizienten findet das Verhältnis von wahrer Varianz zu der beobachteten Varianz ihren Ausdruck. Je größer der Anteil der wahren Varianz ist, umso geringer fällt dabei der Anteil der Fehlervarianz aus. Der Reliabilitätskoeffizient hat einen Wertebereich von 0 (der Messwert besteht nur aus Messfehlern) bis 1 (der Messwert ist identisch mit dem wahren Wert). In der klassischen Testtheorie wird unter Reliabilität nach Lienert (1989) der Grad der Genauigkeit verstanden, mit dem bspw. ein Test ein bestimmtes Persönlichkeits- oder Verhaltensmerkmal misst, unabhängig davon, ob der Test dieses Merkmal auch zu messen beansprucht. Letzteres ist Gegenstand der Validität. Nach Guthke, Böttcher und Sprung (1990) ist die Reliabilität eine notwendige Bedingung für die Validität. Nach der Validität ist in der klassischen Testtheorie die Messgenauigkeit das wichtigste Gütekriterium. Als vollkommen reliabel wäre ein Test dann anzusehen, wenn durch seine Messergebnisse ein Proband oder eine Verhaltensweise fehlerfrei beschrieben werden kann, d.h. ein Test wäre in der Lage, den wahren Wert ohne

jeden Messfehler zu erfassen. Die Reliabilität wäre demnach umso höher, je kleiner der zu einem Messwert gehörende Fehleranteil ist. In der Praxis lassen sich jedoch Einflüsse durch diverse Störungen (situative Bedingungen, Müdigkeit des Probanden, Missverständnisse o.ä.) nie ganz ausschließen, so dass o.g. Idealfall praktisch nicht auftritt. Der Grad der Reliabilität wird durch einen Reliabilitätskoeffizienten bestimmt, der angibt, in welchem Maß unter gleichen Bedingungen gewonnene Messwerte über ein und denselben Probanden übereinstimmen und damit, in welchem Maß ein Testergebnis reproduzierbar ist (Lienert, 1989, S.15). Eine reliable Messung liegt dann vor, wenn der Fehleranteil klein ist; eine unreliable Messung, wenn ein großer Fehleranteil gegeben ist. Die Reliabilität per se existiert jedoch nicht, sondern nur verschiedene methodische Zugänge (Paralleltest, Retest, Interne Konsistenz) über die sie erschlossen werden kann (Lienert, 1989). Erst wenn die Untersuchung eine hohe Reliabilität besitzt, ist gewährleistet, dass die erhobenen Testwerte durch mögliche Störbedingungen während der Untersuchungssituation nicht beeinflusst werden (Bortz, 1989, S.18).

2.3.1.2 Prognostische Validität

Wenn beschrieben werden soll, inwieweit ein Auswahlinstrument in der Lage ist, die Eignung eines Bewerbers um einen Ausbildungsplatz bzw. eine vakante Position in der Organisation vorherzusagen, wird eine Maßzahl benötigt, mittels derer die Enge des linearen Zusammenhanges zwischen dem Instrument (Prädiktor) und Kriterium (z.B. Berufserfolg) bestimmt werden kann. Dabei wird das mit dem Auswahlinstrument erreichte Ergebnis des Bewerbers (z.B. IQ-Wert im Intelligenztest) mit einem zuvor erhobenen Merkmal des Berufserfolges (z.B. verkaufte Policen pro Monat bei einer offenen Stelle in der Versicherungsbranche) in Beziehung gesetzt und auf statistische Signifikanz überprüft. Damit ist Hauptgegenstand der Untersuchung, inwieweit die Ergebnisse der ersten Datenreihe eine Vorhersage der Ergebnisse der zweiten Datenreihe erlauben. Dabei bleibt allerdings zu betonen, dass eine hohe Korrelation nicht zwangsläufig mit einer Erklärung für die Wirkungszusammenhänge einhergeht, das heißt keine Angaben zur Ursache des Zusammenhanges gemacht werden können. Allerdings hängt die erschlossene Höhe des Validitätskoeffizienten im wesentlichen von drei Faktoren ab: Vom Grad dessen, was an Gemeinsamkeit durch den Test und das Kriterium erfasst und oft als „Zulänglichkeit“ des Tests bezeichnet wird, von der Reliabilität des Tests und von der Reliabilität des Kriteriums. Die konkurrente und die prädiktive Validität lassen sich als Unterformen der kriteriumsbezogenen Validität voneinander unterscheiden. Bei der konkurrenten Validität

werden Test- und Kriteriumswerte nahezu gleichzeitig erhoben, bei der prädiktiven Validität dagegen die Testergebnisse zu einem Zeitpunkt t1 und zu einem späteren Zeitpunkt t2 die Punktwerte des Kriteriums. Eine trennscharfe Unterscheidung zwischen diesen beiden Arten zur Bestimmung der Validität ist allerdings nicht immer möglich (Amelang & Zielinski, 1994). Kleinevoss und Sonnenberg (1987) weisen darauf hin, dass die Prognostizität von wissenschaftlich fundierten Auswahlverfahren generell eher unter- als überschätzt wird, da allgemein angenommen werden kann, dass der wahre Zusammenhang zwischen Prädiktor und Kriterium immer unterschätzt wird, da bei beiden mit Messfehlern gerechnet werden muss.

Da es in der eignungsdiagnostischen Praxis in der Regel darum geht, aus einer Vielzahl möglicher diagnostischer Instrumente diejenigen auszuwählen, die mit einem geringen Aufwand und wenig Kosten eine größtmögliche Enge des Zusammenhanges zwischen gewählten Prädiktoren und Kriterien aufzeigen, wird der Wert der prognostischen Validität daran gemessen, wie hoch der prozentuale Anteil der Varianz des Kriteriums ist, der durch die Varianz des Prädiktors aufgeklärt werden kann. Dieser im Determinationskoeffizienten seinen Ausdruck findende Prozentwert wird bestimmt durch das Quadrat von Prädiktor und Kriterium. Korreliert bspw. ein zur Auswahl eingesetztes Instrument zu r = .60 mit dem gewählten Kriterium, so ermöglicht dieses Verfahren, bei gleicher Varianz von Prädiktor und Kriterium, die Aufklärung von 36 % der Kriteriumsvarianz. Nach Lienert (1989) sollte ein Verfahren in einem solchen Umfang valide sein, dass seine Anwendung eine bessere Vorhersage ermöglicht als seine Unterlassung. Für die praktische Anwendung sind allerdings niedrige Validitätskoeffizienten unter r = .30 relativ bedeutungslos, zumindest, wenn sie nicht im Rahmen einer Testbatterie Verwendung finden. Schuler (1989) weist zudem auf unterschiedliche Fehlerquellen hin, aufgrund derer es zu „falschen“ Varianzaufklärungen kommen kann: Stichprobenfehler bei geringer Stichprobengröße, Ungenauigkeiten der Messungen von Prädiktor und Kriterium und Messbereicheinschränkungen aufgrund der Stichprobenauswahl sowie der Kompetenz der Testdurchführenden.
Neben den genannten Einflussgrößen kann aber auch die Zielgruppe eine entscheidende Einflussgröße darstellen. Nach Funke, Kraus, Schuler und Stapf (1987) wurde bei Wissenschaftlern mit biographischen Fragebogen metaanalytisch eine prognostische Validität von r = .47 bestimmt, bei Jugendlichen berichtet Funke (1986) dagegen nur eine Validität von r = .15. Genau andersherum sehen die Relationen jedoch hinsichtlich der prädiktiven Validität des eignungsdiagnostischen Instrumentes Intelligenztest aus. In einer Vielzahl unterschiedlicher Untersuchungen konnte die prädiktive Validität kognitiver Leistungstests

belegt werden. So haben Schmidt und Hunter (1998) die Messung der Intelligenz, mit einer prädiktiven Validität von r = .51, als das wichtigste eignungsdiagnostische Instrument für Einstellungsentscheidungen bezeichnet. Doch Schuler erklärt eine vergleichbar geringe Validität von r = .20 bei Wissenschaftlern durch eine starke Vorselektion hinsichtlich kognitiver Fähigkeiten und einer damit in der Folge geringen aufklärbaren Varianz im Erfolgskriterium.
Abschließend kann in diesem Zusammenhang noch einmal auf die DIN 33430 verwiesen werden. Hier wird gefordert, dass die eingesetzten Auswahlverfahren eine für die Fragestellung möglichst hohe Gültigkeit ausweisen müssen. Dabei muss die Gültigkeit grundsätzlich aufgrund von empirischen Analysen zur Konstrukt-, Kriteriums- oder Inhaltsvalidität nachgewiesen werden. Die Art der Gültigkeitsbestimmung muss dem Zweck des Verfahrens und der vorliegenden Fragestellung angemessen sein.

In der nachfolgenden Tabelle 2.3-1 sind die wichtigsten Auswahlkriterien und ihre prädiktiven Validitäten nach einer metaanalytischen Studie von Schmidt und Hunter (1998) dargestellt.

Tab. 2.3-1: Prognostische Validität von Auswahlkriterien (Schmidt & Hunter, 1998)

AUSWAHL KRITERIEN	PROGNOSTISCHE VALIDITÄT
Intelligenztest	r = . 51
Arbeitsprobe	r = . 54
strukturiertes Interview	r = . 51
unstrukturiertes Interview	r = . 38
Biographische Daten	r = . 35
Assessment Center	r = . 36
Interessen	r = . 10
Test zum Berufswissen	r = . 48
Alter des Bewerbers	r = -. 01

2.3.1.3 Inkrementelle Validität

Gerade weil in Auswahlverfahren häufig unterschiedliche Datenerhebungsinstrumente Anwendung finden, ist der Begriff der inkrementellen Validität an dieser Stelle erwähnenswert. In der Praxis dient häufig nicht ein Test allein als einziger Hinweisgeber

dafür, inwieweit eine auszuwählende Person für eine vakante Position geeignet erscheint, sondern es werden Ergebnisse unterschiedlicher Informationsquellen (Schulnoten, Leistungstests, Interviews u.ä.) miteinander in Beziehung gesetzt, um das gewählte Kriterium möglichst umfassend vorhersagen zu können. So erklären Guthke et al. (1990) die inkrementelle Validität als den gegenüber den anderen angewandten Datenerhebungsmethoden zusätzlichen Beitrag eines Tests zur Erklärung der Kriteriumsvariablen. Der Validitätskoeffizient eines Testverfahrens ist nicht isoliert zu bewerten, sondern im Zusammenhang mit weiter Anwendung findenden Verfahren, d.h. hinsichtlich der Frage, in welchem Grad er gegenüber den weiteren Verfahren zusätzliche diagnostische Erkenntnisse liefert. Ein zusätzlich eingesetztes Testverfahren besitzt demnach inkrementelle Validität, wenn dessen Aufnahme in eine Testbatterie als zusätzlicher Prädiktor zur Vorhersage des Kriteriums, den Anteil der aufgeklärten Varianz am Kriterium erhöht. Dies hat zur Folge, dass auch ein hoch valider Leistungstest nicht für die Kombination im Rahmen einer Testbatterie von Auswahlinstrumenten nützlich sein könnte, da die durch ihn gewonnenen Informationen bereits durch andere Komponenten erfasst werden.
Beispielhaft soll die Auswirkungen der Aufnahme unterschiedlicher Prädiktoren in eine Testbatterie und die damit einhergehende inkrementelle Validität bzw. fehlende inkrementelle Validität anhand der folgenden Tabellen (2.3-2 – 2.3-5) veranschaulicht werden.

In Tabelle 2.3-2 wird der Fall dargestellt, wenn zwischen drei Variablen (Prädiktor 1, Prädiktor 2 und Kriterium) kein nachweisbarer statistischer Zusammenhang besteht.

Tab.2.3-2: Korrelationsmatrix mit zwei Prädiktoren und Kriterium
ohne statistischen Zusammenhang

	KRITERIUM	PRÄDIKTOR 1	PRÄDIKTOR 2
Kriterium	1.0	.00	.00
Prädiktor 1		1.0	.00
Prädiktor 2			1.0

Keiner der beiden Prädiktoren ist geeignet für die Vorhersage des Kriteriums, da die drei Variablen nicht miteinander korrelieren (Tabelle 2.3-2).

In Tabelle 2.3-3 wird dargestellt, wenn von den beiden zur Verfügung stehenden Prädiktoren, nur einer mit dem Kriterium einen statistisch nachweisbaren Zusammenhang aufzeigt.

Tab.2.3-3: Korrelationsmatrix mit statistisch nachweisbarem Zusammenhang zwischen Prädiktor 1 und Kriterium

	KRITERIUM	PRÄDIKTOR 1	PRÄDIKTOR 2
Kriterium	1.0	.50	.00
Prädiktor 1		1.0	.00
Prädiktor 2			1.0

Von den beiden zur Verfügung stehenden Prädiktoren korreliert nur der erste mit dem Kriterium (Tab. 2.3-3). Somit ist eine Vorhersage des Kriteriums nur durch den Prädiktor 1 möglich. Prädiktor 2 ist dagegen kein sinnvoller Prädiktor für das Kriterium.

In Tabelle 2.3-4 wird aufgezeigt, wie sich die inkrementelle Validität bei zwei vorhandenen Prädiktoren auf die Vorhersage des Kriteriums auswirkt.

Tab.2.3-4: Korrelationsmatrix mit statistisch nachweisbarem Zusammenhang zwischen beiden Prädiktoren und Kriterium

	KRITERIUM	PRÄDIKTOR 1	PRÄDIKTOR 2
KRITERIUM	1.0	.50	.60
PRÄDIKTOR 1		1.0	.40
PRÄDIKTOR 2			1.0

Beide Prädiktoren korrelieren mit dem Kriterium (Tab. 2.3-4). Dabei wird deutlich, dass Prädiktor 2 einen höheren statistisch nachweisbaren Zusammenhang mit dem Kriterium aufzeigt als Prädiktor 1. Durch die zusätzliche Verwendung des ersten Prädiktors kann jedoch die Vorhersage des Kriteriums verbessert werden. Damit verfügt Prädiktor 1 über inkrementelle Validität, da durch den weiteren Prädiktor zusätzliche Varianz des Kriteriums aufgeklärt werden kann. In Abbildung 2.3-1 wird beispielhaft die zusätzliche Varianzaufklärung durch den Einsatz des zweiten Prädiktors belegt.

Durch die Bestimmung des Determinationskoeffizienten lässt sich zeigen, dass der Prädiktor 2 alleine (r_{x2y} = .60;: r^2_{x2y} = .36) 36 % der Varianz des Kriteriums, der Prädiktor 1 alleine (r_{x1y} = .50 ; r^2_{x2y} = .25) 25 % aufklären kann (Abb. 2.3-1). Beide Prädiktoren zusammen klären

jedoch ($R^2_{x.12}$ = .436) etwa 43 % der Varianz des Kriteriums auf. Damit wird die inkrementelle Validität des zweiten Prädiktors belegt und es ist sinnvoll, diesen zusätzlich für eine bessere Vorhersage des Kriteriums aufzunehmen.

$$R_{c.12} = \sqrt{\frac{r^2c1 + r^2c2 - 2r_{c1}r_{c2}r_{12}}{1 - r^2 12}}$$

$$R_{c.12} = \sqrt{\frac{.50^2 + .60^2 - 2 * .50 * .60 * .40}{1 - .40^2}}$$

$$R_{c.12} = \sqrt{\frac{.25 + .36 - .24}{.84}}$$

$$R_{c.12} = .66 \qquad R^2 c.12 = .436$$

$R_{c.12}$ = multipler Valitätskoeffizient beider Prädiktoren gegenüber dem Kriterium
r_{c1}, r_{c2} = Validitätskoeffizienten der beiden Prädiktoren gegenüber demselben Kriterium
r_{12} = Interkorrelationskoeffizient der beiden Kriterien

Abb.2.3-1: Beleg der inkrementellen Validität des zweiten Prädiktors

Anders sähe die Korrelationsmatrix aus, wenn sich keine inkrementelle Validität ergibt (Tab. 2.3-5).

Die drei Variablen (Prädiktor 1, Prädiktor 2 und Kriterium) korrelieren in gleicher Höhe miteinander (Tabelle 2.3-5). Durch die Aufnahme des Prädiktors 2 zur Vorhersage des Kriteriums, wird diese nicht verbessert. Wenn beide Prädiktoren perfekt korrelieren würden (r_{12} = 1), hätte der zweite Prädiktor keine zusätzliche inkrementelle Validität und könnte aus dem Testverfahren genommen werden bzw. durch einen anderen Testteil ersetzt werden. Bei diesem Beispiel handelt es sich um einen in praxi selten auftretenden Idealfall, der wie bereits gesagt nur zutreffen würde, wenn die Korrelation zwischen Prädiktor 1 und 2 gleich 1 wäre.

Tab.2.3-5: Korrelationsmatrix mit zwei Prädiktoren ohne inkrementelle Validität

	KRITERIUM	PRÄDIKTOR 1	PRÄDIKTOR 2
KRITERIUM	1.0	.50	.50
PRÄDIKTOR 1		1.0	.50
PRÄDIKTOR 2			1.0

Spätestens seit der durch Schuler (1990) im deutschsprachigen Raum verbreiteten Forschung und Bedeutung der sozialen Akzeptanz von Auswahlinstrumenten sollten nur solche Methoden Verwendung finden, deren Nützlichkeit einerseits empirisch dokumentiert werden konnte, deren Anwendung andererseits auch notwendig ist und nicht zu einer einen Bewerber möglicherweise abschreckenden Belastung führt. Illustriert wird die Bedeutung der inkrementellen Validität gut am Beispiel verwendeter psychologischer Leistungstests zur Vorhersage von Berufserfolgen, deren Untersuchung und Nachweis einer gegebenen Validität eine längere wissenschaftliche Tradition hat, als der Nachweis für jede andere Methode (Hunter, 1986; Schmidt & Hunter, 1981). Entsprechend liegen heute eine große Zahl Untersuchungen der Validität kognitiver Fähigkeitstests vor. Hunter und Hunter (1984) unterstreichen die Bedeutung von kognitiven Fähigkeitstest mit dem Fazit, dass weitere Verwendung findende Verfahren häufig als Ergänzungen zu Intelligenztests gesehen und im Zusammenhang mit der Erschließung der inkrementellen Validität bewertet werden können. In nachfolgender Tabelle 2.3-5 stellen Schmidt und Hunter (1998) die prädiktive Validität von eignungsdiagnostischen Prädiktoren bei Kombination mit einem zweiten Prädiktor und somit die inkrementelle Validität gegenüber Intelligenztests dar.

Wie durch Tabelle 2.3-6 deutlich wird, besitzen eine Vielzahl unterschiedlicher Verfahren zwar eine relativ hohe prädiktive Validität, führen jedoch im Sinne der inkrementellen Validität nur zu einem geringen Validitätszuwachs. Erklären lässt sich dieser Umstand am besten am Beispiel der biographischen Daten, die bei zusätzlicher Anwendung eines standardisierten kognitiven Fähigkeitstest nur einen Validitätszuwachs von r = .01 beitragen. Dieses Ergebnis kann dahingehend interpretiert werden, dass biographische Daten in beträchtlichen Umfang (r = .50) mit allgemeinen kognitiven Fähigkeiten korrelieren und somit vermutlich auch ein indirektes Maß für mentale Fähigkeiten darstellen.

Tab. 2.3-6: Eignungsdiagnostische Prädiktoren kombiniert mit einem zweiten Prädiktor nach Schmidt und Hunter (1998)

EIGNUNGSDIAGNOSTISCHE INSTRUMENTE	VALIDITÄT	INKREMENTELLE VALIDITÄT
Intelligenztests	. 51	---
Arbeitsproben	. 54	. 12
Interview (strukturiert)	. 51	. 14
Interview (unstrukturiert)	. 38	. 07
Biographische Daten	. 35	. 01
Assessment Center	. 36	. 01
Interessen	. 10	. 01
Test zum Berufswissen	. 48	. 07
Alter	- . 01	. 00

Die Untersuchung der inkrementellen Validität ist insbesondere in der psychologischen Grundlagenforschung von Bedeutung. Die Ergebnisse lassen wiederum Schlussfolgerungen für die Zusammenstellung von eignungsdiagnostischen Verfahrenskomponenten zu. So untersuchten bspw. Bühner und Schmitz-Atzert (2003) die inkrementelle Validität von Aufmerksamkeit und Arbeitsgedächtnis gegenüber Intelligenz im Hinblick auf die Prädiktion von Schulleistung und Problemlöseleistung. Hierbei zeigten sich in konfirmatorischen Faktorenanalysen nur moderate Zusammenhänge zwischen Arbeitsgedächtnis- und Aufmerksamkeitsfaktoren. Weder Aufmerksamkeits- noch Arbeitsgedächtnistests leisteten über die Intelligenz hinaus einen signifikanten Beitrag zur Prädiktion von Wissensanwendung in Problemlöseszenarien. Lediglich bei der Prädiktion des Wissenserwerbs konnten Aufmerksamkeits- und Arbeitsgedächtnistests zusätzliche Varianz gegenüber Intelligenz aufklären.

2.3.1.4 Weitere Aspekte der Validität

Orientiert man sich am Konzept der klassischen Testtheorie sollte erwartet werden, dass höhere Reliabilitäten von Test und Kriterium auch höhere Validitäten erbringen. Fisseni (1990) macht dies am Beispiel des Reliabilitätsindex deutlich. Hiernach gibt dieser die Obergrenze der kriterienbezogenen Validität an, denn enger als mit dem wahren Wert können die beobachteten Werte eines Tests nicht korrelieren. Die Obergrenze bestimmt sich als

Wurzel des Reliabilitätskoeffizienten. An Beispiel veranschaulicht bedeutet dies: Liegt die Reliabilität bei r = .78 kann die kriteriumsbezogene Validität höchstens den Wert r. = .88 erreichen. Ein Grund, warum die Validität eines Tests ihre Obergrenze nicht erreichen kann, könnte in der geminderten Reliabilität des Kriteriums zu finden sein. Da der wahre Kriteriumswert wie auch der wahre Testwert nicht beobachtbar sind, ist deren Korrelation nicht unmittelbar bestimmbar. Gemäß den Axiomen der klassischen Testtheorie lassen sich jedoch ihre Werte durch Äquivalente schätzen und hochrechnen. Mittels dieser so genannten einfachen Minderungskorrektur kann die Minderung korrigiert werden, die der Validitätskoeffizient zufolge mangelnder Kriteriumsreliabilität erlitten hat (Lienert, 1989). Die nachfolgende Formel (Abb. 2.3-2) ermöglicht eine Hochrechnung a posteriori, die von der Fiktion eines vollkommen reliablen Kriteriums ausgeht.

$$crit\ .corr\ .r_{tc} = \frac{r_{tc}}{\sqrt{r_{cc}}}$$

Abb.2.3-2: Einfache Minderungskorrektur

Angewandt werden sollte die einfache Minderungskorrektur, oder wie Spearman sie nannte, die Attenuationskorrektur, immer dann, wenn angenommen werden kann, dass das gemessene Persönlichkeitsmerkmal eine größere Konstanz aufweist als das gewählte Validitätskriterium. Ein Beispiel wäre nach Lienert (1989) die Testvalidierung aufgrund eines individuellen Schätzverfahrens, wenn man zur Feststellung der allgemeinen Intelligenz von Schülern als geringer reliables Kriterium die Einschätzung von Lehrern heranziehen würde und nicht das Ergebnis eines hoch reliablen standardisierten Intelligenztests. Lienert (1989) verweist auch darauf, dass für Zwecke der Testinterpretation möglichst nur diese Formel Anwendung finden sollte und nicht jene, die neben der Unreliabilität des Kriteriums auch die Unreliabilität des Tests, d.h. des Prädiktors, mit einbezieht (Abb. 2.3-3).

$$crit\ .et\quad test\quad corr.r_{tc} = \frac{r_{tc}}{\sqrt{r_{tt} \bullet r_{cc}}}$$

Abb.2.3-3: Doppelte Minderungskorrektur

Zum einen ist die Rechtfertigung psychologisch fragwürdig, zum anderen lässt sich belegen, dass der doppelminderungskorrigierte Validitätskoeffizient praktisch kaum eine Bedeutung hat, da er kaum höher wird als der einfach minderungskorrigierte. Darüber hinaus kann die Anwendung der doppelten Minderungskorrektur durchaus zu einem paradoxen Effekt führen. Um den Effekt anschaulich zu machen, wird im Beispiel in Tabelle 2.3-7 die beobachtete Validität = r_{tc} von drei Tests (a – c) konstant gehalten. Steigen die Reliabilitäten von Test = r_{tt} und Kriterium = r_{cc} (von a nach c) an, dann sinkt die Korrelation zwischen wahren Test und wahren Kriterium = crit. et test corr. r_{tc}.

Tab. 2.3-7: Beispielhafte paradoxe Auswirkungen der doppelten Minderungskorrektur.

	TEST A	TEST B	TEST C
Validität Test r_{tc}	**.60**	**.60**	**.60**
Reliabilität Test r_{tt}	**.60**	**.65**	**.70**
Reliabilität Krit. r_{cc}	**.60**	**.65**	**.70**
doppelte Minderungskorrektur $.crit.et.test.corr\, r_{tc} = \frac{r_{tc}}{\sqrt{r_{tt} \bullet r_{cc}}}$	**1.0**	**.92**	**.86**

In der vorliegenden Untersuchung sind die zur Bestimmung der prädiktiven Validität Verwendung findenden Kriterien überwiegend als reliabel einzuschätzen, demnach ist keine Minderungskorrektur zu rechtfertigen.

Wie schon in Kapitel 2.3.1.3 angedeutet, ist im deutschsprachigen Raum spätestens seit der durch Schuler (1990) verbreiteten Forschung, die soziale Akzeptanz von Testverfahren ein weiterer wichtiger Aspekt der Validität.

Wie Petersen (2002) verweist, stellt die Auswahlsituation häufig den Erstkontakt zwischen Bewerber und Organisation dar und aufgrund fehlender anderer Erfahrungen können Bewerber in diesem Stadium den Eindruck gewinnen, dass die Auswahlsituation und das von der Organisation gezeigte Verhalten repräsentativ für die Organisation seien. Somit wird dem Bewerber durch die Teilnahme am Auswahlverfahren ein Eindruck der Institution vermittelt, welcher, insbesondere durch abgelehnte Bewerber, die keine Gelegenheit finden, diese

selektive Wahrnehmung zu kompensieren, an andere potentielle Bewerber weiter getragen wird. Verfahren mit einer hohen Akzeptanz dürften die Kooperationsbereitschaft und die Motivation der Teilnehmer und somit die Validität der Eignungsdiagnose positiv beeinflussen (Nevo, 1986).
Allerdings verweist Kersting (1998) darauf, dass das Konzept der sozialen Akzeptanz nicht mit zu hohen Erwartungen überfrachtet werden darf. „Eine hundertprozentige Akzeptanz von Auswahlverfahren wird es solange nicht geben, solange Leistungsmessungen anstrengend und stressig sind, solange mit Hilfe dieser Verfahren Personen zurückgewiesen werden, solange Bewerber mit zum Teil unrealistischen Erwartungen in die Auswahlverfahren gehen und solange sich gegensätzliche Interessen von Bewerbern und Organisationen im Auswahlverfahren manifestieren“ (S.73).

In Deutschland eingeführt wurde das Konzept der sozialen Akzeptanz von Schuler und Stehle (1983) und von Schuler (1990) weiterentwickelt. In Tabelle 2.3-8 sind die wichtigsten Aspekte dieses Konzeptes zusammengefasst.

Tab. 2.3-8: Komponenten der sozialen Validität nach Schuler und Stehle (1983)

Partizipation und Transparenz bei Entwicklung und Einführung Durch	
⇓	⇓
Sozialpsychologische Anforderungen	Kommunikation Inhalt und Form

Mit sozialpsychologischen Anforderungen sind Informationen über das Sozialgefüge einer Organisation gemeint, d.h. das Organisationsklima und der gelebte Führungsstil. Partizipation bedeutet die Einbeziehung von Organisationsmitgliedern, wenn es um die Entwicklung und Durchführung von Auswahlverfahren geht.

Wie Fruhner, Schuler, Funke und Moser (1991) andeuten, lassen bisherige Untersuchungsergebnisse vermuten, dass das Erleben der Bewerber nicht in erster Linie durch deren subjektive Belastung im Auswahlverfahren beeinflusst wird, sondern eher durch Situationsparameter wie Informationsmöglichkeiten, Kontrollierbarkeit und

Augenscheinvalidität eines Auswahlverfahrens. So wünschen zwei Drittel der Bewerber mehr Vorinformationen zur Eignungsuntersuchung (Graudenz, 1987). Fruhner et al. (1991) befragten Hochschulabsolventen mit welchem Verfahren sie ausgewählt werden möchten (Tab. 2.3-9).

Tab. 2.3-9: Reihenfolge beliebter Auswahlinstrumente bei Hochschulabsolventen nach Fruhner, Schuler, Funke und Moser (1991)

1.	Vorstellungsgespräch
2.	Arbeitsprobe
3.	Praktikumsleistung
4.	Zeugnisnoten
5.	Psychologische Testverfahren
6.	Lebenslauf
7.	Schriftproben
8.	Losverfahren

Die relative schlechte Platzierung psychologischer Testverfahren wird durch die Autoren dahingehend erklärt, da sich die Ablehnung vielfach auf Persönlichkeitstests und weniger auf kognitive oder Intelligenztests bezieht. Insbesondere jene Beurteiler, die noch keine Erfahrungen mit einem genannten Auswahlinstrument machten, beurteilten dieses als weniger positiv.

Interessant ist in diesem Zusammenhang auch das Ergebnis einer Untersuchung von Steck (1997). In einer schriftlichen Umfrage an eintausend zufällig ausgewählten Mitgliedern des Berufsverbandes Deutscher Psychologen, gaben von 271 Befragten 169 die Antwort, dass sie Testverfahren überhaupt in ihrer Berufspraxis einsetzen. Unter den Informationsquellen, die für die Entscheidung über die Anwendung von Tests herangezogen werden, wurde neben wissenschaftlichen Zeitschriften am häufigsten der Katalog der Testzentrale genannt. Unter den Qualitätsmerkmalen, die für die Anwendung eines Tests als entscheidend erachtet werden, nahmen die Gütekriterien (49,7 %) und die diagnostische Relevanz (32,3 %) nur den zweiten und dritten Platz ein. Im Vordergrund stand für die meisten Anwender die Ökonomie der Testdurchführung (63,4 %).

Zurzeit fehlen jedoch insbesondere Untersuchungen zu Beantwortung der Frage, inwieweit sich Bewerber aufgrund weniger beliebter Auswahlverfahren von einer Bewerbung abhalten lassen. Poortinga, Coetsier, Meuris, Miller, Samsonowitz, Seisdedos und Schlegel (1982; zit.

n. Moser und Zempel, 2001) schätzen den Anteil der Bewerber, die lieber auf eine Bewerbung verzichtet hätten als sich einem psychologischen Test auszusetzen auf nur ca. 2 %. Nach Fruhner et al. (1991) würden 84, 3 % der befragten Studenten an einem Vorstellungsgespräch teilnehmen, jedoch nur 45 % an einem psychologischen Test. Damit bleibt jedoch offen, ob die Bewerber wirklich auf eine ihnen lukrativ erscheinende Position verzichten würden oder die Anwendung nur ein negatives Licht auf das Unternehmen wirft.

2.3.2 Zur Prognose der beruflichen Bewährung

2.3.2.1 Prädiktoren

In der Personalauswahl werden unter Prädiktoren eignungsdiagnostische Variablen eines Auswahlverfahrens verstanden, mit denen ein statistisch nachweisbarer Zusammenhang mit dem vorherzusagenden Kriterium (Ausbildungs- oder Berufserfolg) beschrieben werden kann. In den folgenden Abschnitten ist eine Übersicht der wichtigsten eignungsdiagnostischen Prädiktoren dargestellt. Zur umfassenden und vertiefenden Darstellung sei zudem auf die gängige Literatur (z.B. Fisseni, 1990; Jetter, 1996; Kanning & Holling, 2002; Sarges, 1995; Sarges & Wottawa, 2001; Schuler, 1995, 2001) verwiesen.

2.3.2.1.1 Bewerbungsunterlagen

Bei der Auswahl neuer Mitarbeiter stellt die Sichtung und Auswertung der Bewerbungsunterlagen gewöhnlich den ersten Schritt dar (Schuler, 1996). Als erste Informationsquelle für den Diagnostiker dienen sie der Überprüfung der formalen Voraussetzungen für die jeweiligen Stellen (Hollmann & Reitzig, 1995) und erfüllen somit eine primäre Filterfunktion. Aus den beigelegten Materialien erhält die Organisation erste Informationen über einen potentiellen neuen Mitarbeiter und kann sich entscheiden, inwieweit eine weitere Beschäftigung mit der vorliegenden Bewerbung sinnvoll ist oder nicht. Bewerbungsunterlagen bestehen in den meisten Fällen aus einem Anschreiben, Lebenslauf, Angaben und entsprechende Zertifikate über Schulabschlüsse, gesundheitliche Voraussetzungen, Arbeitszeugnisse und Referenzen, Nachweise über evtl. berufliche (Vor-) Erfahrungen oder andere Qualifikationen o.ä.. Dem Lebenslauf wird in den meisten Fällen

eine besondere Bedeutung zugemessen und zählt somit nach Hollmann und Reitzig (1995) zu den ergiebigsten Informationsquellen der Bewerbungsunterlagen. Die aus den Unterlagen gewonnenen Fakten und Daten werden analysiert und dienen als Grundlage weiterer Interpretationen.

Das Ziel der Sichtung der Bewerbungsunterlagen ist es, sich anhand der Fakten und Daten, ein Bild des jeweiligen Bewerbers zu machen. Einerseits soll aus dem vergangenen Verhalten und der Leistung des Bewerbers auf das zukünftige berufliche Verhalten geschlossen werden (beispielsweise abzulesen aus Arbeitszeugnissen und Referenzen). Andererseits werden anhand von Indikatoren, z.B. Schulnoten, auf nicht direkt beobachtbare Eigenschaften geschlossen, die einen Zusammenhang mit der erforderlichen Leistung aufweisen sollen (Raststetter, 1999). Aus Indikatoren wie Auslandsaufenthalt, Vielfalt der bisherigen Stellen und außerberuflichen Aktivitäten (politisches und kulturelles Engagement) wird auf Selbstsicherheit, Selbstständigkeit, Mobilität, Initiative sowie Durchsetzungs- und Entscheidungsvermögen geschlossen. Neben der reinen Informationssammlung dient die Analyse des Lebenslaufes jedoch auch dazu, sich auf das möglicherweise folgende Vorstellungsgespräch vorzubereiten. Die Informationen über die „wahren“ Hintergründe, wie z.B. eines Arbeitswechsels o.ä., werden vermutlich eher durch Befragungen erzielt. Das Anschreiben ist der einzig individuell zu gestaltende Teil der Bewerberunterlagen. Dieser ist somit insbesondere dann von Bedeutung, wenn Biographien von Berufsanfängern vorliegen (Raststetter, 1999). Es erfüllt in diesem Fall eine Art Selbstmarketingfunktion. Neben der eigenen Selbstdarstellung und dem ersten Abgleich zwischen Anforderungen der ausgeschriebenen Funktion und den Fähigkeiten des Bewerbers, werden eher formale Aspekte, wie z.B. Stil und Inhalt beurteilt.

Der Bewerbungsprozess verläuft in den meisten Fällen mehrstufig (Seibt & Kleinmann, 1991), so dass mittels der Analyse der Bewerberunterlagen eine große Anzahl von Bewerbern, die in einer nächsten Phase zu bewältigen sind, reduziert werden können. Dabei werden die Grenzen des praktisch Leistbaren insbesondere bei kleineren Organisationen schnell überschritten. Schnell übersteigt die Nachfrage das Angebot um ein Vielfaches. Kanning (2002) schreibt beispielhaft von 10 ausgeschriebenen Stellen und 1500 Bewerbern. Damit stellt allein die zu bewältigende Masse die Verantwortlichen teilweise vor Probleme, die fast zwangsläufig zu einer nicht mehr sachgerechten Bewertung der Bewerbungsunterlagen führt.

Unter dem Aspekt der Effizienz, also der Kosten-Nutzen-Relation, ist zu berücksichtigen, dass die erste Sichtung und die darauf fußende Ablehnung von Bewerbern in vielen Fällen in relativ kurzer Zeit vonstatten gehen. Seibt und Kleinmann (1991) gehen von einem Zeitraum zwischen durchschnittlich fünf bis fünfzehn Minuten aus. In vielen Fällen liegt die Ablehnungsquote von Bewerbungen bei bis zu 98%. Da der Personalverantwortliche, wie oben beschrieben, eine Großzahl an Bewerbungen zu sichten hat, erscheint sogar der genannte Zeitrahmen fast zwingend erforderlich. Anhand der beurteilten vorliegenden Fähigkeiten und Leistungen in den Unterlagen muss ein subjektives Urteil über die potentielle berufliche Bewährung des Kandidaten gemacht werden. Dabei sollte möglichst kein Geeigneter übersehen oder Ungeeigneter positiv bewertet werden. Beim so genannten Fehler der ersten Art wird der Bewerber nach der Analyse seiner Unterlagen abgelehnt und entsprechend nicht im weiteren Auswahlprozess (Einladung zum Vorstellungsgespräch und Teilnahme am Testverfahren) berücksichtigt, obwohl er für die ausgeschriebene Position geeignet wäre. Für den Bewerber ist dieser Fehler ungerecht, da er trotz seiner bestehenden Eignung nicht weiter berücksichtigt wird. Für die Organisation ist dieser Fehler insbesondere dann bedenklich, wenn insgesamt nur wenige geeignete Bewerber zur Verfügung stehen (Grundquote gering) und somit die Auswahl durch die fehlerhafte Entscheidung zusätzlich erschwert wird (Kanning, 2002). Genau entgegengesetzt verhält es sich bei dem Fehler der zweiten Art. Hier werden zu den weiteren Auswahlschritten jene Bewerber eingeladen, bei denen aufgrund der Sichtung der Bewerbungsunterlagen schon hätte deutlich werden müssen, dass sie für die fragliche Tätigkeit nicht geeignet sind. Für den Bewerber eröffnen sich hier zwar im ersten Moment weitere Chancen, da er nicht schon aufgrund der vorliegenden Schriftlage abgelehnt und zum Einstellungsverfahren eingeladen wurde. Andererseits dürften messgenauere Verfahren seine unberechtigten Hoffnungen zerschlagen oder er findet, im Falle einer Einstellung, aufgrund der Über- bzw. Unterforderung im Beruf keine Berufszufriedenheit. Für die Organisation wirkt sich beim zweiten Fehler insbesondere negativ aus, dass zum einen mehr Bewerber in den nächsten Auswahlschritt eingeladen werden müssen und somit die Ökonomie leidet. Zum anderen sollten nachfolgende Auswahlinstrumente nicht entsprechend reliabel und valide sein, es zur Fehlbesetzung und damit einhergehenden unnötigen zusätzlichen Kosten kommt.

Der Prozess der Eindrucks- und Urteilsbildung aus Bewerbungsunterlagen ist bisher kaum erforscht (Hollmann & Reitzig, 1995). Zwar wurden Versuche unternommen, die

Urteilsfindung der Auswerternden zu simulieren (vgl. Schuler, 1996), jedoch erbrachte dies wenig empirisch gestützte Ergebnisse.
Bei der Aufnahme und Interpretation von Informationen, sei es aus Bewerbungsunterlagen oder im Interview, spielen, wie bei jeder Informationsverarbeitung, kognitive Prozesse des Beurteilers eine zentrale Rolle. Die Kriterien, anhand derer die erhaltenen Informationen bewertet werden, sind in den meisten Fällen nicht explizit formuliert, geschweige denn schriftlich festgehalten. Die individuelle Bewertung der jeweiligen Verantwortlichen hat somit einen starken Einfluss; man denke nur an die kognitiven Prozesse der Schemabildung und Prototypen, welche unter diesen Umständen stark zum Tragen kommen dürften. Die wenigsten der verwendeten Kriterien sind somit objektiv nachprüfbar (vgl. Rasstetter, 1999).
In der vorliegenden Arbeit spielen besonders die Schulnoten als ein potentieller Prädiktor für die berufliche Bewährung eine entscheidende Rolle. Als ein zentraler Teil der Bewerbungsunterlagen wird auf den genannten Punkt an dieser Stelle zwar differenzierter eingegangen, eine detaillierte Diskussion der prognostischen Validität der Schulnoten erfolgt unter Punkt 2.3.2.3.

Weitere wichtige Erkenntnisse für den Auswahlprozess werden ferner den Arbeitszeugnissen zugeschrieben, deren Aussagekraft und Bedeutung jedoch beschränkt wird, da aufgrund des geltenden Arbeitsrechtes kaum ausdrücklich negative Formulierungen und Bewertungen über den Arbeitnehmer erfolgen dürfen. Erfahrene und in der „Zeugnissprache“ (Schuler, 1996) versierte Fachleute können zwar aus diesen wichtige Informationen gewinnen, jedoch werden in vielen Fällen Zeugnisse auch von „Laien“ erstellt, so dass dann keine einheitliche Handhabung z.B. hinsichtlich der Interpretation der Formulierungen gegeben ist. Die Aussagefähigkeit dieser Komponente der Bewerbungsunterlagen wird somit in Frage gestellt. Jedoch spielen Arbeitszeugnisse neben Referenzen in mündlicher Form, insbesondere bei der Auswahl für höhere Positionen, eine große Rolle. Diesem Teil der Bewerbungsunterlagen kommt nach Hunter und Hunter (1984) eine prognostische Validität von r = .26 zu, ein höherer Wert als den Bewerbungsunterlagen insgesamt zukommt, wobei jedoch Reilly und Chao (1982) eine durchschnittliche Validität von .14 ermittelten. Zwischen Referenzen und Leistungsbeurteilungen durch Vorgesetzte ermittelten Moser und Rhyssen (2001) eine Korrelation von r = .20. Als Hauptursache für die eingeschränkte Validität vermuten die Autoren die eingeschränkte Prädiktorvarianz durch die Vorselektion, die mangelnde Reliabilität der Referenz und Mildefehler im Kriterium. Für die gesamten Bewerbungsunterlagen geben Reilly und Chao (1982) eine mittlere prognostische Validität

von r = .18 an. Auch wenn den Schulzeugnissen innerhalb der Bewerbungsunterlagen ein relativ großes Gewicht zukommt, ist die Aussagekraft der Bewerbungsunterlagen insgesamt allerdings begrenzt (Schuler, 1996). Vielfältige Fehlerquellen, wie z.B. auf Seiten des Diagnostikers wirkende Faktoren wie implizite Persönlichkeitstheorien, Prototypenbildung, Schemata, etc., verfälschen die Aussagekraft.

2.3.2.1.2 Einstellungsinterview

Das Einstellungsinterview (Auswahl-, Vorstellungs-, Bewerbungsgespräch) ist wohl nach der Analyse der Bewerbungsunterlagen das am weitesten verbreitete Instrumentarium der Personalauswahl und –beurteilung und dient dem Ziel, die Passung und Eignung der Kandidaten zu klären.
Die Durchführung der Auswahlgespräche wird in den meisten Fällen sehr heterogen gestaltet: Sie reicht von der völlig freien Gesprächsform über teilstrukturierte bis zu vollstrukturierten Varianten mit standardisierten Abläufen und Fragestellungen. Die von den Interviewern gestellten Fragen beziehen sich insbesondere auf die bisherigen beruflichen Erfahrungen und Ausbildungen sowie einzelne relevante Aspekte des Lebenslaufes. Erfragt werden ebenfalls persönliche Aspekte, wie die familiäre Situation, zukünftige Lebensvorstellungen, Einstellungen und Meinungen. Mittels der „klinischen Urteilsbildung“, einer intuitiven Kombination der Informationen aus den Antworten des Bewerbers, seiner nonverbalen Kommunikation und anderer Eindrücke während des Gespräches, wird ein Gesamteindruck gebildet und ein Urteil gefällt (Schuler, 1996).

Das Einstellungsinterview ist sorgfältig vorzubereiten und daher sehr zeitintensiv. Unter anderem müssen die Bewerbungsunterlagen gründlich ausgewertet und einzelne wichtige Aspekte, die einer Nachfrage bedürfen, notiert werden. Außerdem sollte ein roter Faden innerhalb des Gespräches vorhanden sein, welcher sich aus den Anforderungen der jeweiligen Stelle ergibt. Zur Reduzierung der Unsicherheit der Bewerber sollte eine ruhige und ungestörte Gesprächsatmosphäre geschaffen werden.
Das Interview gehört für Bewerber und Diagnostiker zu der am meisten geschätzten und bevorzugten Form der Personalauswahl (Schuler, 1996). Die Bewerber haben die Möglichkeit eine eigene aktuelle Leistung (im Unterschied z.B. zu Schulnoten oder Arbeitszeugnissen) zu erbringen und die Situation gewährt ihnen gewisse Kontrollmöglichkeiten.

Hinsichtlich der empirischen Untermauerung ergibt sich jedoch eine Diskrepanz zu dieser hohen Wertschätzung der Anwender. Im Vergleich mit anderen Auswahlverfahren erbringt das Interview eine relativ geringe prognostische Validität. Arvey & Campion (1982) beziffern in einem Sammelreferat die prognostische Validität bei großer Streuung auf r =.05 bis r =.25. Auch metaanalytische Studien von Hunter und Hunter (1984) erbrachten nur Werte um r =.15. Die Autoren weisen allerdings 1987 darauf hin, dass die in den meisten älteren Validierungsstudien errechneten niedrigen Validitätskoeffizienten möglicherweise aus dem Grunde zustande gekommen sind, da sie eigentlich die inkrementelle Validität, also die Verbesserung der Validität durch zusätzliche Aufnahme des Interviews gegenüber anderen Auswahlinstrumenten, erfassen.

Eine Ursache für die in einigen Studien nur geringe Aussagekraft und Qualität des Interviews geht nach Schuler und Funke (1989) auf eine unzureichende Informationsverarbeitung und Urteilsbildung seitens des Befragenden zurück. Hier spielen Faktoren wie Wahrnehmungseffekte (z.B. Primacy-Recency-Effekt), implizite Persönlichkeitstheorien und Prototypen eine große Rolle und tragen womöglich zu einer geringen Vorhersagekraft der Interviews bei. Auch wenn beispielsweise innerhalb eines standardisierten Gespräches stets die gleichen Fragen gestellt werden, die Bewertung der Informationen (Antworten) hängt immer von dem jeweiligen Interviewer ab.
Andererseits sehen Schuler und Funke (1989) in einer unzweckmäßigen inhaltlichen und fragetechnischen Gestaltung des Interviews eine weitere Schwäche. Weitere Informationen über den „Fehlerfaktor“ Urteilskompetenz des Interviewers findet der interessierte Leser u.a. bei Schuler und Funke (1989) und Sarges (1995).
Qualitätsreduzierend wirkt sich außerdem aus, dass bei den meisten durchgeführten Interviews die Durchführungs-, Auswertungs- und Interpretationsobjektivität kaum in dem erforderlichen Maße gegeben ist.

Trotz dieses Mangels erfüllt das Interview, bezogen auf das o.g., nach Schuler (1996) eine Reihe wichtiger Funktionen. Darunter fallen u.a. die Vorhersage beruflichen Erfolges, Information des Bewerbers über die Stelle, den Betrieb, die Anforderungen, Erwartungen o.ä., persönliche Bekanntschaft zwischen potentiellem Arbeitnehmer und Arbeitgeber, Festlegung der Vereinbarungen hinsichtlich der Stelle, etc.. Diese Funktionen werden von keinem der anderen Komponenten eines Auswahlverfahrens in dem Maße geleistet; das Interview erfüllt somit einige wichtige Funktionen, die durch andere Verfahren in diesem Maße nicht

kompensiert werden. Die niedrige prognostische Validität sollte zwar nicht außer Acht gelassen werden, eher sollte an Möglichkeiten gedacht werden, das Interview unter Objektivitäts-, Reliabilitäts- und Validitätsaspekten effizienter zu gestalten.

Besonders in den 70er und 80er Jahren bemühte sich die Forschung, das Interview aufgrund seines Stellenwertes in der Personalauswahl methodisch zu verbessern. Schuler (1996) bezieht sich auf eine metaanalytische Studie von McDaniel, Whetzel, Schmidt, Hunter, Mauerer und Russel (1986), in welcher für eine anforderungsbezogene Gestaltung des Interviews eine Steigerung der Validität im Vergleich mit traditionellen psychologischen Interviews auf r = .30 errechnet wurden. Auch die Strukturierung kann zur Verbesserung der Validität beitragen: In einer metaanalytischen Studie von Wiesener und Cronshaw (1988) wird von einem mittleren Validitätskoeffizient von r = .40 für strukturierte Interviews im Gegensatz zu einem mittleren Validitätskoeffizient von r = .13 für unstrukturierte Interviews berichtet. McDaniel et al. (1994) errechneten sogar einen korrigierten Validitätskoeffizienten von r = .47 für strukturierte und r = .40 für weniger strukturierte Interviews.

Eine wichtige Möglichkeit, die Validität des Interviews zu erhöhen, ist die Verwendung geprüfter und verankerter Skalen während der Durchführung. Die unterschiedliche Urteilskompetenz der Interviewer könnte somit als Fehlerquelle minimiert werden und den weniger erfahrenen Diagnostiker wird eine Entscheidungshilfe an die Hand gegeben. Zusätzlich kommen u.a. die Vorbereitung der Interviewer durch ein sorgfältig konzipiertes und kompetent geführtes Training, die Trennung von Informationssammlung und Entscheidung, Ergänzung der Interviewfragen durch validierte Fragen aus anderen Testverfahren und / oder biographischen Fragebögen (z.B. situative Fragen) in Betracht (Schuler, 1996). Die aufgeführten Maßnahmen können zur methodischen Verbesserung des Auswahlgespräches beitragen und somit die prognostische Validität dieses Instrumentes erhöhen.
Eine Methode, welche die in der Forschung festgestellten Defizite des traditionellen Interviews zu überwinden und die belegten Verbesserungsmöglichkeiten der Forschungsarbeiten zu integrieren versucht, ist das von Schuler (1992a) entwickelte Multimodale Einstellungsinterview mit einer feststehenden Abfolge von sieben Komponenten. Nach einem kurzen informellen Gespräch, in dem keine Beurteilungen erfolgen dürfen, folgt die Selbstvorstellung des Bewerbers. Anschließend werden standardisierte Fragen zu Berufsinteressen und Berufswahl gestellt, welche gleichzeitig auf

verankerten Einstufungsskalen beurteilt werden. Nach einem weiteren nun freien Gesprächsteil folgen biographiebezogene Fragen, die aus den Anforderungsanalysen der jeweiligen Stelle abgeleitet wurden. Außerdem werden während des knapp 30-minütigen Interviews situative Fragen gestellt, welche ebenfalls anforderungsbezogen entwickelt wurden und deren Beurteilung ebenfalls anhand verankerter Skalen vorgenommen wird. Der Wechsel zwischen freien und standardisierten Interviewkomponenten soll zwischen den Teilnehmern und den Interviewern zu einer positiven und als harmonisch erlebten Gesprächsatmosphäre führen.
Studien zur prognostischen Validität (Schuler, 1996) dieses Interviews bei Bewerbern um Ausbildungsplätze und bei Studierenden erbrachten Validitätskoeffizienten zwischen r = .27 und r = .51. Es wurden Beurteilerübereinstimmungen, selbst bei untrainierten Interviewern, über r = .70 erreicht.

Diese recht zufrieden stellenden Validitätskoeffizienten machen deutlich, dass eine derartige Zusammenstellung aus anforderungsbezogenen, biographiebezogenen und situativen Fragen sowie standardisierten und „freien“ Teilen des Interviews zu einer methodischen Verbesserung und somit Verbesserung der Validität des Interviews beitragen kann. Wichtig ist, dass diese Art der Interviewgestaltung versucht, auch diejenigen Aspekte, welche das Einstellungsinterview zu einem der gebräuchlichsten Instrumente der Personalauswahl macht, zu erhalten und damit die soziale Akzeptanz zu erhöhen (wie z.B. Schaffen einer harmonischen und persönlichen Gesprächsatmosphäre). Denn eine Steigerung der Validität nur auf Basis stärkerer Standardisierung und Strukturierung herbeiführen zu wollen, führt letztendlich dazu, dass das Interview anderen schriftlichen Verfahren ähnelt und durch diese ersetzt werden kann.
Von dieser speziellen Form des Interviews einmal abgesehen, zeigen auch die Ergebnisse der metaanalytischen Studien der letzten Jahre, dass besonders das strukturierte Auswahlgespräch, welches auf anforderungsbezogenen Analysen der jeweiligen Tätigkeit basiert, nachweisbar reliable und valide Ergebnisse erbringt, welche nicht mit anderen Verfahren gewonnen werden (Westhoff, 2000).

2.3.2.1.3 Biographische Fragebogenverfahren

Biographische Daten haben im Bereich der Personalauswahl eine lange Tradition. Historische Belege, wie z.B. die ersten Anwendungen bei der Auswahl von Verkäufern in einer

amerikanischen Versicherungsgesellschaft und weitere Anwendungsbeispiele finden sich seit etwa 70 Jahren in der Literatur (Bliesener, 1991).
Biographische Fragebogen werden u.a. von Stehle (1995) als standardisierte Instrumente, welche der Erfassung soziodemographischer Variablen, Einstellungen, bisherigen Erfahrungen, schulischen und beruflichen Entwicklungen sowie Aktivitäten und Interessen dienen, definiert. Nach Owens (1976), dem einflussreichsten Vertreter dieses Ansatzes in der Organisationspsychologie, zeichnen die biographischen Informationen ein Bild von dem bisherigen Lebensmuster der jeweiligen Person, sozusagen ein Bericht über ihren bisherigen Lebensweg. Unter der Annahme einer relativen Konsistenz des Verhaltens über die Zeit, könnte man somit davon ausgehen, dass anhand der bisherigen beruflichen und persönlichen Erfahrungen zukünftiges Verhalten und Ziele verlässlich prognostiziert werden.
Beispiele häufig in der Forschung verwendeter biographischer Items sind u.a. persönliche Angaben (Alter, Familienstatus), sozialökonomischer Status (regelmäßige Ausgaben, Einkommen, etc.), Interessen (Hobbys, sportliche Aktivitäten, etc.) und berufliche Erfahrungen, wie z.B. frühere Anstellungen, Wechsel von Arbeitsplätzen, Anzahl früherer Berufe u.v.m. (Weinert, 1998).

Inhaltlich sind die Items der biographischen Fragebogen mit denen des Einstellungsinterviews identisch, jedoch erlauben sie im Allgemeinen durch die statistische Urteilsbildung höhere Prognosen des beruflichen Erfolges als Einstellungsinterviews (Stehle, 1995).
Hinsichtlich der Konstruktion von biographischen Fragebogen ist in den letzten Jahren eine Vielzahl unterschiedlicher Verfahrensweisen entwickelt worden. Eine wichtige und häufig Anwendung findende Funktion in der Personalauswahl besitzen die biographischen Fragebögen, vielfach um größere Bewerberzahlen zu reduzieren. Bewerber, die aufgrund ihrer Angaben im Vergleich zu Mitbewerbern im Fragebogen nicht oder weniger für die Organisation oder für die ausgeschriebene Stelle geeignet erscheinen, können frühzeitig, d.h. vor der Anwendung aufwendig durchgeführter eignungsdiagnostischer Instrumente, abgewiesen werden. Daneben können die durch die Teilnehmer im Fragebogen gemachten Angaben auch als ein zusätzliches Instrument zur Eignungsprüfung im traditionellen Vorstellungsgespräch genutzt werden. Darüber hinaus finden biographische Fragebögen im Rahmen von Klassifikationsmaßnahmen sowie Beruf- und Karriereberatungen Anwendung.

Für die recht häufige Anwendung im organisationspsychologischen Umfeld spricht einerseits ihre leichte Zugänglichkeit und erhebliche Augenscheinvalidität (Weinert, 1998). Ein weitaus

wichtigerer Punkt ist jedoch die in vielen metaanalytischen Studien nachgewiesene prädiktive Validität der biographischen Daten (Reilly & Chao ,1982; Hunter & Hunter, 1984; Schuler & Funke, 1989). Reilly und Chao (1982) bescheinigten eine kriterienbezogene Validität in einem Bereich von r =.30 bis r =.50. Stehle (1990) beschreibt sogar einen Validitätsbereich zwischen r =.40 und r =.70. Er zählt die biographischen Fragebögen im Vergleich zu anderen Verfahren zu den validesten Einzelverfahren. Bei der Anwendung biographischer Fragebogen für die Auswahl von Bewerbern für den mittleren Polizeivollzugsdienst konnte Petersen (2002) eine prognostische Validität von r = .50 erschließen, die sich auch in der Kreuzstichprobe bestätigen ließ.

Trotz dieser recht beachtlichen Validitätswerte sollten einige Nachteile dieses Verfahrens ebenfalls berücksichtigt werden. Aufgrund der in meisten Fällen durchgeführten itemweisen Validierung der biographischen Fragebögen (vgl. a. Hollmann, 1991; England, 1971, Petersen, 2002) wird einerseits zwar eine hohe Anpassung an die Stichprobe möglich. Andererseits entsteht somit eine eingeschränkte Anwendbarkeit auf andere betriebliche Situationen, so dass bei Anwendung unter anderen Bedingungen (z.B. andere Organisationen) der Fragebogen erneut validiert werden muss.
Weiterhin stellt sich die Frage nach der Generalisierbarkeit bzw. Spezifität der aus den faktorenanalytischen Berechnungen gewonnen biographischen Dimensionen. Einige Prädiktoren stehen sicherlich für mehrere Situationen, wie z.B. Selbstvertrauen und Extraversion für den „allgemeinem beruflichen Erfolg“, während andere eher spezifischere nur in ganz klar umrissenen Situationen, wie z.B. Erziehungs- und Ausbildungserfolg, ihren Einfluss haben (Weinert, 1998).
Weiterhin ist u.a. über eine mögliche Verfälschbarkeit und eine generelle Eignung der biographischen Daten als Prädiktoren nach Weinert (1998) viel zu wenig bekannt. Ebenso werden aufgrund zu geringer Bewerberzahlen oft die notwendigen Kreuzvalidierungen und Replikationen der Untersuchungen nicht durchgeführt.

2.3.2.1.4 Standardisierte Testverfahren

Die Anwendung von standardisierten Testverfahren in der Eignungsdiagnostik hat eine lange Tradition. Standardisierte Test- und Fragebogenverfahren unterliegen einer Reihe von Gütekriterien (insbesondere hinsichtlich der Reliabilität, Validität und Objektivität), denen sie genügen müssen, und sind das Ergebnis eines aufwendigen wissenschaftlich fundierten

Entwicklungsprozesses. Psychologische Testverfahren stellen den Prototyp konstruktorientierter Diagnoseverfahren dar und sind nach Brandstätter (1979) standardisierte, routinemäßig anwendbare Verfahren zur Messung individueller Verhaltensmerkmale, aus denen Schlüsse auf Eigenschaften bestimmter Personen oder deren Verhalten in bestimmten Situationen gezogen werden können. Die Notwendigkeit der Anwendung standardisierter Testverfahren folgte der in vielen empirischen Untersuchungen belegten Anfälligkeit der menschlichen Informationsverarbeitung für systematische Fehler und Verzerrungen (Dörner, 1996, Kanning, 1999). Als Tests werden in erster Linie solche Messinstrumente bezeichnet, mit denen Leistungen wie Konzentrationsfähigkeit, Vigilanz oder Intelligenz erfasst werden können. Neben den klassischen Leistungstests, in denen Antworten eindeutig als falsch oder richtig festgelegt sind, finden auch standardisierte Fragebogen Anwendung, die sich auf die Beschreibung einer Person bzw. ihres Verhaltens beziehen. Der wesentliche Unterschied zwischen Leistungstests und standardisierten Verfahren zur Selbstbeschreibung besteht darin, dass die Bewerber bei der Bearbeitung eines Selbstbeschreibungsfragebogens ihr hypothetisches Handeln in einer bestimmten Situation beschreiben. D.h. die Bewerber geben durch ihre Antworten an, wie sie in beschriebenen Momenten handeln würden. Dagegen handeln die Probanden bei der Bearbeitung eines Leistungstests tatsächlich.

Gerade bei der Personalauswahl stellt sich bei der Durchführung von Testverfahren auch immer die Frage nach dem Wahrheitsgehalt der gemachten Angaben. Bewerber, die mit Mitbewerbern um einen Ausbildungsplatz in Konkurrenz stehen, sind bemüht, sich besonders vorteilhaft darzustellen. Einige mögen sogar bestimmte Angaben verschweigen oder Antworten verfälschen. In der Psychologie wird dieses Verhalten unter dem Begriff „Sozial erwünschtes Verhalten“ behandelt (s.a. Mummendey, 1987). Bei professionellen Leistungstests spielen derartige Verfälschungstendenzen jedoch nur eine geringe Rolle. Ihre Aufgaben sind in der Regel so konzipiert, dass durch Raten nur eine geringe Trefferquote erzielt werden kann. Dazu werden die erfassten Merkmale in der Regel über verschiedene Items erfasst, so dass der Einfluss des Rateverhaltens relativ gering bleibt.

Die Vielfalt standardisierter psychologischer Testverfahren im Bereich der Personalauswahl ist heute sehr umfassend. Zum Überblick aktuell bedeutsamer eignungsdiagnostischer Verfahren sei der interessierte Leser auf Kanning und Holling (2002) oder Sarges und Wottawa (2001) verwiesen. Nach Schorr (1991) ist es jedoch in der Praxis nur eine geringe Zahl von Testverfahren, die regelmäßig zum Einsatz kommen. An erster Stelle stehen bei der

Personalauswahl mit 46,8% der befragten Arbeits- und Organisationspsychologen die Intelligenztests. Weiterhin werden insbesondere Persönlichkeitstests, spezielle Funktionsüberprüfungs- und Eignungstests sowie auch klinische Verfahren angewandt (Tabelle 2.3-10).

Tab. 2.3-10: Testverfahren in der Arbeits- und Organisationspsychologie nach Schorr (1991)

VERFAHREN	ANWENDUNG DURCH A & O PSYCHOLOGEN
Intelligenztest	46,8 %
Persönlichkeitstest	31,2 %
Spezielle Funktions- und Eignungstests	24,7 %
Klinische Testverfahren	23,4 %
Allgemeine Leistungstests	14,3 %
Projektive Verfahren	5,2 %

Aus gleicher Quelle ist zu erfahren, dass 14,3 % der Wirtschaftspsychologen die Verwendung von Intelligenztests ablehnen. Dies ist insbesondere bemerkenswert, da nur 11,7 % sich gegen die Anwendung von Klinischen Testverfahren aussprechen, die in der Regel nicht für den Gebrauch in der Arbeits- und Organisationspsychologie konzipiert sind und für andere Anwendungsfelder normiert wurden. Eine neuere Erhebung stammt von Scheinecker und Wallner (2003). Sie befragten insgesamt 135 Unternehmen in ganz Österreich hinsichtlich der Einsatzhäufigkeit von Personalauswahlinstrumenten (Tabelle 2.3-11).

Tab. 2.3-11: Einsatzhäufigkeit von Personalauswahlinstrumenten
nach Scheinecker & Wallner (2003)

AUSWAHLINSTRUMENT	HÄUFIGKEIT		
	Häufig	manchmal	nie
Interview	50	1	0
Analyse der Bewerbungsunterlagen	47	4	0
Zeugnisse	35	13	3
Assessment Center	8	25	18
Biographische Fragebogen	7	12	32
Referenzen	6	38	7
Leistungstest	6	23	22
Persönlichkeitstest	6	19	26
Arbeitsproben	5	25	21
Intelligenztest	3	12	36
Graphologisches Gutachten	1	1	49

Wie in den Angaben der 51 ausgewerteten Rückläufe erkennbar, werden Intelligenztests, trotz gegebener Gütekriterien und wiederholt belegter prädiktiver Validität von 36 Unternehmen nie verwendet. Auch die standardisierten Leistungstests und Persönlichkeitsfragebogen werden von den befragten Unternehmen seltener angewandt, als bspw. Interviews oder die Analyse von Bewerbungsunterlagen, die beide geringere Validitäten aufweisen können.

Dem praktisch tätigen Eignungsdiagnostiker steht heute darüber hinaus eine Vielzahl unterschiedlicher personaldiagnostischer Instrumente für verschiedene Einsatzgebiete zur Verfügung (siehe auch Kanning & Holling, 2002). Die Autoren fassen zum einen mit dem Berliner Intelligenzstruktur Test (BIS-4,), dem Intelligenz-Struktur-Test (IST 2000 von Amthauer, Brocke, Liepmann & Beauducel, 1997), dem Leistungsprüfsystem (LPS von Horn, 1962), dem Wilde-Intelligenz-Test (WIT von Jäger & Althoff, 1994), Konzentrations-Leistungstest (KLT von Dücker, 1959) oder dem Aufmerksamkeits-Belastungs-Test (d2 von Brickenkamp, 1962) Instrumente zur Messung allgemeiner kognitiver Leistungen zusammen. Nach Hüttemann (2002) konnten Kleine und Jäger (1989) und Wittmann und Matt (1986) für den von Jäger, Süß und Beauducel (1997) auf den Markt gebrachten und besonders in der jüngeren Zeit im Rahmen der Personalauswahl Anwendung findenden BIS-4, eine kriteriumsbezogene Validität für Schulnoten und Leistungen in Hochschuleignungsprüfungen zwischen r. = .40 und r = .60 erschließen. Für den WIT ist im Zusammenhang mit der vorliegenden Untersuchung bemerkenswert, dass das Testverfahren zum einen u.a. an einer Gruppe von Bereitschaftspolizisten (n = 314) normiert wurde und eines der Vorgänger-testverfahren vom aktuell durchgeführten Testverfahren für die Auswahl von Polizeivollzugsbeamten des gehobenen Dienstes war. Hier konnten Wolff und Voullaire (1968 nach Jödden, 2002) bezogen auf die Ausbildungsleistung von Krankenschwestern und Regierungsinspektoren prädiktive Validitätskoeffizienten von r = .59 – r = .76 ermitteln. Zu den standardisierten Instrumenten, die berufsbezogenen Leistungen erfassen sollen, rechnen die Autoren etwa den Berufseignungstest (BET von Schmale & Schmidtke, 1966), die Drahtbiegeprobe (DBP von Imming, 1920 bzw. Lienert, 1967) den Mannheimer Test zur Erfassung physikalisch-technischen Problemlösens (MPT von Conrad, Baumann & Mohr, 1980) oder den Fragebogen zur Analyse belastungsrelevanter Anforderungsbewältigungen (FABA von Richter, Rudolf & Schmidt, 1996). Wie den Tabellen 2.3-10 und 2.3-11 zu entnehmen, spielen auch die Instrumente zur Messung von Persönlichkeitsmerkmalen in der Personalauswahl eine bedeutende Rolle. Hossiep, Paschen und Mühlhaus (2000) geben einen

guten Überblick über die derzeit in der Eignungsdiagnostik verwendeten Verfahren. Beispielhaft sei an dieser Stelle der 16-Persönlichkeitsfaktoren-Test (16 PF von Schneewind, Schröder und Cattell, 1994), das Neo-Fünf-Faktoren-Inventar (NEO-FFI von Borkenau und Ostendorf, 1993) erwähnt. Hossiep et al. (2000) verweisen jedoch auf neuere metaanalytische Ergebnisse, welche die Nützlichkeit von Persönlichkeitsverfahren auch für die Personalauswahl belegen und verweisen bspw. auf Barrick und Mount (1991), die für das Fünf-Faktoren-Modell der Persönlichkeit mit ihrer Metaananalyse differentielle Zusammenhänge zwischen einzelnen Traits und Maßen beruflichen Erfolges feststellen konnten. Insbesondere das Persönlichkeitsmerkmal Gewissenhaftigkeit korrelierte in allen untersuchten Berufsgruppen positiv mit verschiedenen Erfolgskriterien. Auch Salgado (1997) konnte die valide Vorhersage des Berufserfolges bestätigen. Dabei erwiesen sich die Dimensionen Gewissenhaftigkeit und Emotionale Stabilität über verschiedene Berufsgruppen hinweg als valide Prädiktoren für unterschiedliche Maße beruflichen Erfolges.

2.3.2.1.5 Assessment Center Verfahren

Assessment Center (AC) werden heute häufig dann angewendet, wenn die Anzahl der Bewerber nicht zu groß und die Besetzung der Position mit einem geeigneten Bewerber von einer entsprechenden Bedeutung ist. Denn trotz ihrer wiederholt bescheinigten hohen prognostischen Validität handelt es sich um ein aufwendig durchzuführendes und kostspieliges Auswahlverfahren. Nach Jeserich (1991) dient es daneben in erster Linie sowohl der Personalförderung und Personalentwicklung sowie dem Training einzelner Mitarbeiter und zur Abschätzung ihres Potentials. Da das AC für die Auswahl zur Einstellung in den Polizeivollzugsdienst nur eine sehr untergeordnete, weil insbesondere zu teure und zu aufwendig durchzuführende Methode der Personalauswahl darstellt, soll an dieser Stelle nur kurz auf das Verfahren eingegangen werden. Der interessierte Leser sei auf weiterführende Literatur wie Jeserich (1991), Kleinmann (1997), Sarges (1996) und Schuler und Stehle (1992) verwiesen.

Unter einem Assessment-Center wird allgemein ein systematisches Verfahren zur qualifizierten Feststellung von Verhaltensleistungen bzw. Verhaltensdefiziten für mehrere Teilnehmer in Bezug auf vorher festgelegte Anforderungen verstanden (Jeserich, 1991). Dabei ist es das primäre Anliegen, zu erfassende Anforderungssituationen möglichst realistisch in einer Testsituation abzubilden. Häufig nehmen mehrere Beurteilte gleichzeitig

am Verfahren teil und werden durch mehrere unabhängige Beobachter hinsichtlich ihrer Leistungen eingeschätzt. Faustregel ist dabei, dass 12 Teilnehmer von sechs Beobachtern beurteilt werden. Andererseits wird das Vorgehen der individuellen Situation der Teilnehmer sowie des Zweckes seiner Anwendung angepasst und ist in diesem Sinne flexibel. So finden gerade auch in der Führungskräfteauswahl teilweise so genannte Einzel-Assessment`s statt, bei denen ein einzelner Kandidat sich einem Einzelbeurteilungsprogramm stellt.
Häufige Inhalte der individuell den Bedürfnissen der jeweiligen Anforderungsprofile anpassbaren Assessment Center sind Aufgaben wie Rollenspiele, Gruppendiskussionen mit und ohne Rollenvorgabe, Postkörbe, Interviews, Vorträge und Präsentationen, Wirtschaftsspiele oder Organisations-, Analyse- und Entscheidungsaufgaben in einem AC. Exemplarisch erfasste Anforderungsmerkmale sind Administrative Fähigkeiten, Soziale Kompetenz, Kognitive Kompetenz, Leistungsverhalten oder Selbstbild der Teilnehmer. Ziel ist es, Personen herauszufinden, die ein für die Organisation wichtiges Managementpotential besitzen.
Dies gestaltet sich jedoch insofern nicht immer einfach, da Managementpositionen selbst innerhalb eines Unternehmens sehr unterschiedlich sein können (Frieling & Sonntag, 1999).
Die prognostische Validität des Assessment Center konnte durch unterschiedliche Studien belegt werden. So geben Schmidt und Hunter (1998) einen Validitätskoeffizienten von r = .36 an. Vermutlich sind es insbesondere die allgemeinen kognitiven Fähigkeiten, die hinter den eigentlich zu erfassenden Anforderungen stehen und den Erfolg oder Nichterfolg im Auswahlverfahren bestimmen (Klimoski & Brickner, 1987). Bei gegebener inhaltlicher Validität und nachgewiesener prädiktiver Validität, jedoch wiederholt kritisch reflektierter Konstruktvalidität (Höft & Funke, 2001; Kleinmann, 1997; Schuler, 1992b), ist das Assessment Center auch ein augenscheinvalides Instrument, welches aufgrund seiner Kosten jedoch in erster Linie bei solchen Bewerbern zum Einsatz kommen dürfte, denen als potentielle Mitarbeiter eine besondere Wertschätzung zukommt. Damit eignet es sich nicht als Breitbandinstrument in der Eignungsdiagnostik.

2.3.2.1.6 Computergestützte Verfahren

Die Mehrzahl der psychologischen Testverfahren wurde ursprünglich als paper-pencil Verfahren entwickelt und sind heute auch als computergestützte Verfahren erhältlich. In diesem Kontext sind somit als computergestützte Verfahren keine weiteren zusätzlichen Bausteine einer wissenschaftlich geführten psychologischen Personalauswahl gemeint,

sondern das Papier und Bleistiftmedium hat sich im Hinblick auf die Durchführung den veränderten Entwicklungen angepasst.
In fast allen beruflichen Bereichen, wie auch im täglichen Leben, nimmt der Gebrauch von Computern ständig zu. Auch im Bereich der Berufseignungsdiagnostik werden Computer in vielen Bereichen und Situationen eingesetzt. Einen großen Vorteil sieht Schuler (1996) u.a. in der Unterstützung des organisatorischen Ablaufes in Großunternehmen und in der computerspezifischen Durchführung und Auswertung von Arbeitsproben und Simulationen. In der Bundesanstalt für Arbeit und in der Bundeswehr werden beispielsweise häufig wiederkehrende und umfangreiche Testverfahren durchgeführt, welche durch die Unterstützung von Computern effizienter gestaltet werden. Ergebnisse können bspw. unmittelbar zur Verfügung stehen. Durch den Einfluss der Forschung des „Adaptiven Testens“ auf die computergestützte Eignungsdiagnostik ist es möglich, optimale an den jeweiligen Probanden angepasste Untersuchungsschritte zu verwenden (Hornke, 1991). So kann verhindert werden, dass ein Bewerber eine zuvor festgelegte Reihenfolge von vorgegebenen Aufgaben abarbeiten muss. Hier übernimmt der Computer die Aufgabe, für den jeweiligen Probanden jene Aufgaben auszuwählen, die hinsichtlich ihres Schwierigkeitsgrades den jeweiligen Fähigkeiten des Teilnehmers entsprechen. In der Folge stehen die Testlänge und Genauigkeit in einem optimalen Verhältnis und sowohl Unter- wie auch Überforderungen können vermieden werden. Ein Beispiel für ein adaptives Verfahren ist der Frankfurter Konzentrationsleistungs-Test (Fakt) von Moosbrugger und Heyden (1997).

Die computergestützte Messung weist nach Kanning (2002) gegenüber dem klassischen Papier und Bleistift Verfahren mehrere Vorteile auf. Zum einen unterliegen Durchführung und Auswertung in einem weit geringeren Maße dem Einfluss des Diagnostikers, so dass die Objektivität größer wird. Weiter ist die Auswertung in der Regel deutlich schneller und für Organisationen damit auch kostengünstiger als bei herkömmlichen Verfahren. Da insbesondere viele junge Menschen, die häufig die Population der Bewerber um Ausbildungsplätze abbilden, in der heutigen Zeit auch im privaten Bereich mit Computern umgehen, ist zudem anzunehmen, dass dies auch zu einer höheren Akzeptanz der eignungsdiagnostischen Verfahren beitragen kann. Neben der vollen Standardisierung der Durchführung und Auswertung besteht ein weiterer Vorteil in der Erhebung zusätzlicher Daten des Probanden, wie z.B. Latenzzeiten, Fehlerreaktionen und Korrekturen (Schuler, 1996). Nachteile wie z.B. geschlechtsspezifische Unterschiede in der Bearbeitungsgüte oder

generell für Personen ohne Computererfahrung konnten in empirischen Untersuchungen nicht belegt werden (Gittler, 1990; Schmotzer, Kubinger & Maryschka, 1994).

Bisher beschränkt sich der größte Teil des computerunterstützten Testens auf die, auf das Medium Computer übertragene, Vorgabe von Papier-Bleistift-Tests. Eine vermeintliche Äquivalenz beider Verfahren wird jedoch bisher kontrovers diskutiert (vgl. Booth, 1991). Einige Autoren betonen die Notwendigkeit, die testtheoretischen Eigenschaften in jedem Einzelfall zu überprüfen. So konnten Neubauer und Urban (1991) bei einem Vergleich der Computer- und der Standardversion von Ravens APM (1958) signifikante Unterschiede hinsichtlich der Vergleichbarkeit der Normwerte feststellen. Die Äquivalenzproblematik tritt besonders bei kognitiven Fähigkeitstests auf, die unter Zeitdruck bearbeitet werden müssen (Mead & Drasgow, 1993 zit. n. Schuler & Höft, 2001). Die Autoren vermuten, dass die unterschiedlichen motorischen Anforderungen der Verfahrensformen bei den Speed-Varianten besonders stark ins Gewicht fallen und für die Unterschiede verantwortlich sind. Eine Gegenüberstellung von Vor- und Nachteilen anhand unterschiedlicher Testgütekriterien findet sich bei Kubinger (1993).

In der kognitionspsychologisch orientierten Forschung, im Bereich von komplexen, dynamischen Problemstellungen, kann der Computer jedoch einen Beitrag leisten, der durch kein anderes Medium in der Form zu ersetzen ist. Unter diagnostischen Gesichtspunkten werden komplexe Fragestellungen vorgegeben; die Problemlöseprozesse der jeweiligen Probanden können dann durch die Analyse der Fehler, Strategien sowie Lösungsschritte und -zeiten überwacht, gesteuert und ausgewertet werden (Rüppell, 1991).

Trotz der genannten Vorteile konnten die computergestützten Verfahren die herkömmlichen Testverfahren bisher nicht ersetzen. Die Frage nach der Verwendung computergestützter Testverfahren beantworteten in der weiter oben genannten Untersuchung von Steck (1997) nur 8 % der Anwendung positiv. Mögliche Gründe könnten beispielsweise mit der von Booth (1991) angesprochenen fraglichen Äquivalenz der computergestützten Verfahren vs. Papier- und Bleistift-Testens und der damit verbundenen Schwierigkeiten (Itempräsentation) zusammenhängen. Die Möglichkeiten neuer Techniken auch in der Eignungsdiagnostik sollten zudem nicht von der diagnostischen Aufgabe der Gewichtung und Integration einzelner Informationen zu einem Gesamturteil ablenken. Ein korrekter Umgang mit den

Daten der Bewerber erfordert trotz einer objektiveren Durchführung nach wie vor diagnostische Kompetenz und Erfahrung.

2.3.2.1.7 Arbeitsproben

Arbeitsproben sind bei nahezu fast allen Berufsbereichen einsetzbar. Schuler (1996) definiert sie als „standardisierte Aufgaben, welche inhaltlich valide und erkennbar äquivalente Stichproben des erfolgsrelevanten beruflichen Verhaltens provozieren." Die Abgrenzung der Arbeitsproben gegenüber Tests wird nicht immer einheitlich gehandhabt. Einerseits werden nur motorische Aufgaben zu den Arbeitsproben gezählt. Andererseits werden Tests, sofern sie in standardisierter und normierter Form vorliegen, als Arbeitsproben definiert (Schuler, 1996). Mit Hilfe der Arbeitsprobe soll von einer Verhaltensstichprobe auf ähnliches zukünftiges Verhalten geschlossen werden. Beispiele für klassische Arbeitsproben wären z.B. das Halten eines Probeunterrichtes für Pädagogen oder das Fertigen eines Kleidungsstückes für einen Schneider. Auch im Assessment Center finden sich Arbeitsproben als Teilverfahren, wie z.B. Postkorbaufgaben und Planspiele. Die erforderlichen Arbeitstätigkeiten werden hier im Unterschied zu Tests nicht in vermeintliche Personenmerkmale, wie z.B. für die Ausführung erfolgsrelevante Eigenschaften, sondern auf Verhaltensmerkmale übertragen. Ähnlich wie bei der Durchführung von Assessment Center eignet sich die Arbeitsprobe eher für überschaubare Bewerbergruppen und weniger für die Auswahl einer größeren Anzahl von Interessenten.

Nach Schuler (1996) werden für gut gestaltete Arbeitsproben relativ hohe Validitätswerte erreicht. Motorisch ausgeführte Arbeitsproben erbrachten höhere Validitätswerte als hauptsächlich verbal gestaltete (gemessen anhand des Kriteriums der beruflichen Leistung, operationalisiert anhand von Vorgesetztenbeurteilungen). In diesem Zusammenhang bezieht sich Schuler (1996) auf eine Zusammenstellung der Daten von Cascio (1987), welche Validitätswerte zwischen r = .25 und r = .30 aufweisen. In einer Studie von Funke (1986) mit Auszubildenden zum biologisch bzw. chemisch-technischen Assistenten wurde für situative Verfahren (in diesem Fall eine Kombination einer umfangreichen Arbeitsprobe und der Beobachtung der Bewerber in einer Gruppensituation) eine prädiktive Validität von r = .24 gefunden. In einer von Schuler (1996) genannten metaanalytischen Studien von Schmitt (1984) resultieren sogar Validitätswerte von durchschnittlich r = .38, Schmidt & Hunter (1998) kommen sogar mit einem r = .54 zu dem Ergebnis, dass Arbeitsproben nach Intelligenztestverfahren, zu den Prädiktoren mit der größten Validität zählen.

Ein großer Vorteil der Arbeitsprobe ist, ähnlich der Bewertung der Assessment Center durch die Bewerber, ihre hohe Augenscheinvalidität. Sie fördert die soziale Akzeptanz und trägt auch zur Selbstselektion bei (Schuler, 1996). Dagegen stellt ihre Konstruktion, trotz der damit offensichtlich gegebenen Vorzüge des Verfahrens, ein schwieriges Unterfangen dar. Dies besonders unter Berücksichtigung einer anforderungsbezogenen sowie einer theoretisch und methodisch fundierten Gestaltung. In vielen Fällen weisen sie weniger Items als Tests auf, wobei deren Unabhängigkeit oft auch nur begrenzt gegeben ist. Wenn andererseits die Items eine hohe inhaltliche Validität aufweisen, stellt sich die Frage nach deren Generalisierbarkeit im Vergleich zu Fähigkeitstests (Schuler, 1996).
Aufgrund ihrer hohen Akzeptanz seitens der Bewerber und Verwender, ihrer im Prinzip breiten Anwendungsmöglichkeit in nahezu fast allen Berufsbereichen sowie der im Vergleich zu anderen Prädiktoren hohen prädiktiven Validität, stellt die Arbeitsprobe, bei methodisch korrekter Gestaltung, ein geeignetes Verfahren dar, die berufliche Qualifikation verhaltensorientiert zu erfassen. Hierbei sollten jedoch Aufwand und realistische Durchführungsmöglichkeiten miteinander in Beziehung gesetzt werden. Bei der internen Auswahl eines im Dienst befindlichen Polizeibeamten für eine leitende Funktion, z.B. Dienstgruppenleiter, ist die Arbeitsprobe sicherlich ein durch Bewerber und Organisation akzeptiertes und dazu valides Instrument. Für eine große Anzahl von potentiellen Berufanfängern, die sich um einen Ausbildungsplatz für den Polizeivollzugsdienst bewerben, in der vorliegenden Untersuchung n = 735 Personen, wäre sie allerdings kaum realistisch und ökonomisch durchführbar.

Die in den letzten Abschnitten beschriebenen Verfahren, die als Prädiktoren zur Auswahl geeigneter Mitarbeiter in unterschiedlichen beruflichen Bereichen eingesetzt werden, beziehen sich im speziellen Fall nicht auf den in dieser Arbeit untersuchten Kontext der Auswahl von Polizeivollzugsbeamten. Einerseits ist anzumerken, dass es kaum Untersuchungen zur Personalauswahl in diesem Bereich gibt. Andererseits decken sich die Anforderungen im Polizeiberuf in vielerlei Hinsicht mit denen aus anderen Berufen, so dass diese Ergebnisse auf den Bereich der Personalauswahl von Polizisten übertragen werden können.

2.3.2.2 Schulnoten

Schulnoten oder auch Zensuren genannt, sind nach Ingenkamp (1982) ein in Kurzform (Ziffer, Buchstabe oder Adjektiv*)* gefasstes Urteil des Lehrenden über ein Verhalten des Lernenden. Durch die Note wird das beurteilte Verhalten und / oder die Leistung rangmäßig im Vergleich zu dem Verhalten und / oder der Leistung anderer Lernender gebracht, d.h. Schulnoten dienen der Skalierung pädagogisch bedeutsamer Leistungs- und Verhaltensmerkmale. Als Indikatorvariable für die Schulleistung sollen sie interindividuelle Unterschiede in Leistungen und Verhaltensmerkmalen und intraindividuelle Veränderungen der Lernenden deutlich machen. Die Zensuren werden in Zeugnissen zusammengefasst und sollen den jeweiligen Leistungsstand des Schülers darstellen. Im deutschen Sprachgebrauch handelt es sich bei den Schulnotenskalen um inverse Rangskalen, d.h. gute Leistungen entsprechen niedrigen Noten und umgekehrt.

Die Abgabe von Zeugnissen wurde bereits im 16. Jahrhundert vollzogen und mündet in die heutige Zeit. Mittels der Zeugnisvergabe wird das gesellschaftliche Bedürfnis nach Kontrolle und Auslese entsprechend dem Leistungsprinzip vollzogen. Neben der pädagogischen Funktion der Schulnoten, den Schüler mit Leistungsvergleichen vertraut zu machen und ihm Rückmeldungen über den eigenen Leistungsstand zu geben, ermöglichen sie die Zugangsberechtigung zu Ausbildungs- und Arbeitsplätzen gemäß der eigenen Leistung und nicht aufgrund anderer Faktoren (z.B. Herkunft). So ist es auch bei der Auswahl von Auszubildenden für den Beruf des Polizeivollzugsbeamten fast in jedem Bundsland üblich, nur jene Bewerber zu einem weitergehenden Einstellungstestverfahren einzuladen, die über einen bestimmten Notendurchschnitt in vorher festgelegten Fächer verfügen. Somit nehmen Schulnoten einen herausragenden Stellenwert im Leben jedes Einzelnen ein, denn sie beeinflussen den weiteren Lebensweg in einer leistungsorientierten Gesellschaft. Ein Bewerber, der das entsprechende Vorauswahlkriterium „Schulnote 3“ im Fach Deutsch nicht vorzeigen kann, wird keine Einladung zum Auswahlverfahren für seinen Traumberuf erhalten und somit nie erfahren können, ob er nicht in einem möglicherweise valideren Verfahren seine Eignung hätte beweisen können. Und dies, obwohl durch die verschiedenen Schulen, Schultypen und Fächerkombinationen eine mangelnde Vergleichbarkeit der Zeugnisse gegeben ist. Außerdem werden möglicherweise innerhalb einer Schule unterschiedliche Anforderungen, orientiert am jeweiligen Klassennmittel, an die Schüler gestellt, die bei einer vorliegenden Bewerbung in der Phase der Vorauswahl nicht berücksichtigt werden können.

Bewerber A wird mit einer Note 2 im Fach Deutsch angenommen, hat diese jedoch möglicherweise nur deshalb bekommen, weil die Deutschleistungen der Klasse, möglicherweise in einem Stadtteil mit einem hohen Ausländeranteil unter den Schülern, insgesamt sehr schlecht waren. Bewerber B wird dagegen mit einer Note 4 im gleichen Fach abgelehnt, obwohl er objektiv über bessere orthographische Kenntnisse verfügt, jedoch in einem Klassenverband gelernt hat, der sich aus ungewöhnlich vielen guten Schülern zusammensetzt.

Wie bei der diagnostischen Urteilsbildung müssen bei der Bewertung von Schulleistungen Daten gespeichert, auf Richtwerte bezogen und unterschiedlich gewichtet werden. Die Zensurengebung stellt somit eine kognitive Leistung dar, welche den „typischen“ Fehlern eines derartigen Beurteilungsprozesses unterliegt (vgl. auch Abschnitt 2.12). In dem Prozess der Beurteilung der Leistung eines Schülers durch einen Lehrer spielen Faktoren wie selektive Wahrnehmung, Erinnerungsfehler (z.B. serielle Positionseffekte), Urteilstendenzen, implizite Persönlichkeitstheorien, kognitive Schemata, Erwartungs- und Einstellungsartefakte (z.B. Pygmalion-Effekt) sowie Antipathie und Sympathie etc. eine das Ergebnis beeinflussende Rolle.

Ebenso wie bei anderen Prädiktoren interessieren auch bei Schulnoten deren empirische Eigenschaften, d.h. die Gütekriterien der Schulnoten, wie die Objektivität, Reliabilität und Validität.

Objektivität bedeutet in diesem Falle eine möglichst weitgehende Ausschaltung subjektiver Einflüsse auf das Beurteilungsergebnis (Ingenkamp, 1982). Das heißt, verschiedene Lehrer sollten den gleichen Lernerfolg auch gleich beurteilen.

Die ersten Untersuchungen zur Objektivität des Lehrerurteils fanden bereits Anfang des letzten Jahrhunderts statt. Ingenkamp (1982) weist auf eine Arbeit von Ellis 1912/1913 hin, welcher in den USA eine Geometriearbeit mit der Bitte an verschiedene Lehrer schickte, diese zu zensieren. Die eingegangenen Urteile streuten fast über das gesamte Notensystem. Ingenkamp (1982) nennt weiterhin eine Arbeit von Hadley, welcher in seiner Untersuchung zeigen konnte, dass beliebte Schüler im Durchschnitt besser benotet wurden als unbeliebte. Tent (1998) nennt eine Auswertungsobjektivität von r = .8 bis r = .9 bei älteren wie auch bei neuere Studien zur Aufsatzbeurteilung.

Eine grundlegende Schwäche der Zensuren, welche zu einem Mangel an Objektivität führt, ist deren unzulängliche Vergleichbarkeit (Ingenkamp, 1971). Da kein klassenübergreifender Maßstab besteht, anhand derer die Leistungen der Lernenden beurteilt werden können,

müssen die gleichen Noten im selben Fach, auf derselben Stufe, desselben Schultyps nicht die objektiv gleiche Leistung darstellen. Wenn derselbe Lernerfolg somit extrem unterschiedlich zensiert wird, sind die erteilten Schulnoten nicht objektiv. Vollkommene Objektivität der Benotung wäre nur bei solchen Schulleistungen möglich, die in Maß und Zahl zu erfassen sind. Dies könnte für gemessene Leistungen bei einem Langstreckenlauf im Sportunterricht, Lückendiktate oder Klassenarbeiten mit vorgegebener Richtig-Falsch Lösung gelten (Lienert, 1987).

Hinsichtlich der Reliabilität der Schulnoten zeichnet sich ein ebenso schwaches Bild ab. Die Zuverlässigkeit der Schulnoten bedeutet in diesem Falle, dass ein Lehrer dieselbe Arbeit bei verschiedenen Gelegenheiten gleich benotet, das heißt hinsichtlich des gleichen Objektes besteht Urteilskonstanz.
Nach Tent (1998) liegt die Retest-Reliabilität gemittelter Schulnoten in der Grundschule von Jahr zu Jahr bei r> .8 und bleibt mit einer Korrelation von bis zu r = .8 für einen Zeitraum bis zu drei Jahren beachtlich hoch. Bis zum achten Schuljahr hingegen fällt sie bei Einzelnoten und auf der Sekundarstufe niedriger aus und sinkt mit zeitlichem Abstand (r = .2). Für Fremdsprachen und Sport ergeben sich Retest-Reliabilitäten um r = .5, diese Fächer scheinen somit auf Gymnasien relativ stabil benotet zu werden. Eine Untersuchung über die Reliabilität von Mathematiknoten unternahm Dicker 1973. Er ließ 24 deutsche Lehrer dieselbe Schülerarbeit einer 5. Schulstufe im Abstand von 3 Monaten zweimal benoten. Durch die Versuchsanordnung war die Wahrscheinlichkeit, dass sich ein Lehrer bei der zweiten Beurteilung noch an die erste erinnern konnte, gering. Die Auswertung der Unterschiede ergab eine signifikante Abweichung, die jedoch in keinem Fall größer als eine Notenstufe war (Dicker, 1989).
Besonders für die Feststellung der Entwicklung der Leistung, also des Lernzuwachses, ist die Zuverlässigkeit der Messung wichtig. Sonst wäre es nicht klar, ob es sich um eine Steigerung der Leistung handelt oder nur um Zufall. Eine niedrige Reliabilität der Noten würde somit nicht die pädagogische Funktion der Erteilung von Rückmeldungen und der Information über den aktuellen Leistungsstand und dessen Entwicklung erfüllen.

Bezüglich des wichtigsten Gütekriteriums der Validität der Schulnoten lässt sich in Anbetracht der bisher gemachten Erläuterungen feststellen: Ein Prädiktor kann aufgrund einer geringen Objektivität und Reliabilität keine zufrieden stellend hohe Validität aufweisen. Die erschlossene Höhe eines Validitätskoeffizienten hängt wesentlich vom Grad dessen, was an

Gemeinsamkeit durch den Test und das Kriterium erfasst und oft als „Zulänglichkeit“ des Tests bezeichnet wird, sowie der Reliabilität des Tests und der Reliabilität des Kriteriums ab. Zahlreiche Untersuchungen belegen, dass das System der Ziffernnoten nicht den Anforderungen an Validität, Reliabilität und Objektivität in zufrieden stellendem Maße erfüllt (vgl. a. Dumke, 1973; Ingenkamp, 1977). Einen guten Überblick über die gesamte Problematik der Leistungsbeurteilung findet der interessierte Leser bei Weiss (1989 a), Lienert (1987) und Ingenkamp (1982, 1995). Ingenkamp (1982) nennt eine Untersuchung von Pritz, der eine Abiturprüfung in Geographie, die mit der Note 3 beurteilt worden war, absolut wort- und gestenidentisch von derselben Person einmal in 16 und das andere Mal in 21 Minuten sprechen lässt, und die Videoaufnahmen 81 Geographielehrern vorgespielt hat. Die langsame Fassung war bedächtig und ohne Stottern vorgetragen und wurde durchschnittlich mit 3,38 benotet, während die schnellere Fassung mit durchschnittlich 2,51 beurteilt wurde. Ein völlig fachfremdes und irrelevantes Merkmal, nämlich die Sprechgeschwindigkeit, trug demnach wesentlich zur Benotung bei. Der Einfluss artfremder Merkmale und systematischer Urteilsverfälschungen (wie oben angesprochen) können somit in erheblichem Maße die Höhe der Validität von Schulnoten beeinflussen.

Wie absurd die immer wieder auftretende Vorstellung ist, eine Mathematikarbeit sei präziser zu bewerten als eine Sprache oder irgendeine andere Art von Prüfungsarbeiten, stellten Starch und Elliot schon 1913 fest. Die Autoren ließen dieselbe schriftliche Geometriearbeit von 128 amerikanischen Mathematiklehrern mit dem jeweils schulüblichen Punktesystem bewerten. Die kritische Marke zum Bestehen einer Arbeit betrug je nach Schule 70, 75 oder 80 Punkte; die größtmögliche Punktezahl war ebenfalls unterschiedlich. Die größte Streuung war in der Gruppe der Schulen mit kritischem Wert 70 zu verzeichnen: Die 43 Bewertungen reichten von 25 bis 89 Punkten. 49 Lehrer beurteilten jede Frage einzeln und verwendeten dazu eine Skala von 0 bis 12½. Bei einer der zehn Fragen der Arbeit streuten die Bewertungen dabei sogar von 0 bis 12½. Weiss (1989b) legte je eine Rechenarbeit der 4. und eine der 5. Schulstufe 153 bzw. 119 Lehrern zur Beurteilung vor. Die 153 Benotungen der Arbeit aus der 4. Klasse reichten von 1 bis 5, jene der 5. Klasse von 2 bis 5.
Auch Vorurteile haben einen Einfluss auf die Benotung. So erhielt ein Teil der Lehrer die Information, die Arbeit der 4. Schulstufe würde von einem mathematisch begabten Schüler mit einer Neigung zu originellen Lösungen stammen, während es bei jener der 5. Klasse hieß, sie stamme von einem durchschnittlich begabten Schüler und weiter, dass die Arbeit durch unsaubere Form und schlampige Schrift auffiele. Dem anderen Teil der Lehrer wurden

dieselben Informationen vertauscht geboten. Das Ausmaß der Abweichung war dabei in der 4. Klasse „sehr signifikant, in der 5. Klasse nicht signifikant.“ (Weiss, 1989 b).
In einer anderen Studie ließ Weiss erneut 24 Volksschullehrer eine Mathematikarbeit der 4. Klasse und 10 Hauptschullehrer eine der 5. Klasse beurteilen. In beiden Gruppen reichten die Noten ebenfalls von 1 bis 5 (Weiss, 1989b).
Hanisch (1990) legte insgesamt 42 Lehrern zwei verschiedene Schülerarbeiten derselben Schularbeit (5. Klasse) zur Benotung vor. 19 Lehrer bekamen die eine Arbeit (Gruppe A), 23 die der Gruppe B. In Gruppe A streuten die Noten von 2 bis 5, in Gruppe B von 3 bis 5.

Ungeachtet der fehlenden Voraussetzungen an Objektivität, Reliabilität und Validität sind Schulnoten für den eignungsdiagnostisch tätigen Praktiker seit jeher ein beliebtes Kriterium für personelle Entscheidungen. Sie sind einfach zu erheben, somit ökonomisch und preiswert. Durch die in den letzten Jahren steigenden Bewerberzahlen, z.B. für den öffentlichen Dienst, die Privatwirtschaft und natürlich im Hochschulbereich, hat die Frage nach der Verwendbarkeit der Schulnoten bzw. -zeugnisse eine neue Aktualität erhalten, da in den meisten Fällen die Zulassungs- und Einstellungsentscheidungen ausschließlich auf der Grundlage von Schulzeugnissen erstellt werden (Althoff, 1986).

In der Berufseignungsdiagnostik werden Schulnoten neben der Vorauswahl, als so genannte Filter zur Überprüfung bestimmter formaler Voraussetzungen, hauptsächlich zur Vorhersage bestimmter tätigkeitsrelevanter Eigenschaften, wie z.B. „Intelligenz“ und „Lernfähigkeit“ eingesetzt, um von diesen auf den späteren Berufserfolg zu schließen (Sarges, 1995).
Die meist im Rahmen der eingereichten Bewerbungsunterlagen dokumentierten Leistungsbewertungen in Form von Noten liefern den Betrieben ein bekanntes und abgestuftes Bewertungssystem. Anhand dessen lassen sich die Bewerber leicht differenzieren und sie dienen in vielen Fällen als Hauptentscheidungsgrundlage für deren Ablehnung. Häufig, wie in manchen Bundesländern auch bei der Auswahl und Zulassung von Bewerbungen für den Polizeivollzugsdienst üblich, kommt es sogar zu einer gesetzten Notengrenze bei der Annahme der Bewerbungen. Damit wird die Auswahl der Bewerbungsunterlagen zeitlich ökonomischer und erleichtert die Grobsichtung bei großen Anzahlen eingegangener Bewerbungen. Neben der Reduzierung der Bewerberzahlen haben die Noten und Abschlüsse jedoch vielfach eine weitere Funktion. In der Praxis werden hier Rückschlüsse u.a. auf Intelligenz, Lernfähigkeit, Anpassungsfähigkeit, Belastbarkeit u.ä. und daraus wiederum auf den späteren Berufserfolg gezogen (Hollmann & Reitzig, 1995, S.467).

Gute Noten im Sport bedeutet Einsatzfreude, in Singen und Werken Anpassungsfähigkeit und in Fremdsprachen Willenseinsatz (Behrens & Merkel, 1989) und gelten generell als ein Indiz für Engagement und Anpassungsbereitschaft (Berthel, 1995). Aus den gewählten Fachrichtungen werden außerdem auf Interessen sowie Begabungen und Fähigkeiten geschlossen. Besonders dieser Induktionsschluss ist nicht logisch zwingend, da die Wahl der Fächer nicht nur von Fähigkeiten und Interessen abhängig ist, sondern in vielen Fällen auch von den jeweiligen Lehrern, die das Fach anbieten, bestimmt wird (Althoff, 1986). Eine erwartete „mildere" Bewertung durch einen bestimmten Lehrer sowie Sympathien spielen bei der Wahl des Faches oft eine große Rolle. Schulnoten bewerten zudem die Leistungen und Kenntnisse der Vergangenheit und sind nicht zukunftsorientiert, weil sie auf die jeweilige Schulsituation begrenzt sind. Somit sind Schlussfolgerungen aufgrund erreichter Noten für Anforderungen des Arbeitsplatzes zumindest fraglich. Das heißt, ein Lehrer kann nicht die jeweilige Eignung des Schülers für einen bestimmten Beruf bei der Benotung berücksichtigen. Zudem ist bei den erworbenen Schulnoten kaum differenzierbar, ob eine gute Note in Deutsch durch Fleiß oder durch entsprechende intellektuelle Fähigkeiten zustande gekommen ist. Diese Vermutung würde die von Ingenkamp (1969) indirekt bestätigen, der davon ausgeht, dass die Anpassung an geforderte Verhaltensweisen stärker als das eigentliche Leistungspotential in Schulnoten berücksichtigt wird. Folglich ist die eignungsdiagnostische Funktion der Schulnoten unter Berücksichtigung dieses Aspektes nicht auf Leistungs-, sondern auf Anpassungsmerkmale (oder ferner auf andere Verhaltensmerkmale) beschränkt. Dies betrifft insbesondere die Wahl der Fächer in der Oberstufe. Hier werden nach Althoff (1986) die Fächer, die stärker in die Bewertung eingehen, nicht nach Interessen ausgewählt, sondern nach den jeweiligen Lehrern, die das Fach unterrichten. Somit spielt die Vorstellung von geringeren Anforderungen und „milderen" Bewertungen eine größere Rolle als die eigentlichen Interessen. Unter diesen Umständen können die Fächer und somit letztendlich die Schulnoten kein Spiegelbild der Neigungen und Interessen sein.

Schul- und Studiennoten gelten jedoch trotz der o.g. Kritik als die valideste Einzelkomponente der Bewerbungsunterlagen (Schuler, 1996). Hinsichtlich der prognostischen Validität von Schulnoten berichtet Tent (1998) von Werten für den deutschsprachigen Raum von $r < .5$ bei Untersuchung der Noten als Prädiktoren für Ausbildung und Beruf. Nach einer metaanalytischen Neuberechnung von Einzelergebnissen zur prognostischen Validität von Schulnoten von Baron-Boldt, Schuler und Funke (1989) ergab sich eine mittlere korrigierte Validität der Haupt- und Realschulabschlussnoten von $r = .41$ für den Ausbildungserfolg, die der Abiturnote zur Studienerfolgsprognose von $r = .46$. Als

validester Einzelprädiktor erwies sich in beiden Fällen die Mathematiknote. Diese Werte gelten jedoch nur für die Vorhersage des Ausbildungserfolges, für die Prognose des beruflichen Erfolges belaufen sich die berichteten Koeffizienten nach Schuler (1996) auf durchschnittlich r = .15.

Schuler und Funke (1995) fanden in ihrer durchgeführten Meta-Analyse eine Validität von r = .40 für den Schluss von Realschulzeugnissen auf die Leistung beim Abschluss der beruflichen Ausbildung. Dabei wurde jedoch auch deutlich, dass je weiter die Schulleistungen zurückliegen, desto geringer wird ihr Aussagewert.

Meier (2003) untersuchte die Frage, inwieweit Schulnoten mögliche Kriterien sind, die einen Studienplatzbewerber geeignet erscheinen lassen, ein Jurastudium erfolgreich durchzuführen und abzuschließen. Hintergrund für seine Untersuchung war die Annahme, dass hierüber in Juristenkreisen weitgehender Konsens besteht. Gemeinhin wird angenommen, dass das erfolgreiche Jurastudium bestimmte kognitive Fähigkeiten voraussetzt. Namentlich wird die Fähigkeit zu analytischem und logischem Denken als zentral angesehen, ferner die Fähigkeit zum Umgang mit verschiedenartigen, komplexen Informationen und zu sprachlicher Genauigkeit. Von den Studierenden wird erwartet, dass sie in der Lage sind, sich schriftlich und mündlich richtig und überzeugend zu artikulieren. Dabei wird der sprachlichen Ausdrucksfähigkeit in der gegenwärtigen Diskussion eine gesteigerte Bedeutung beigemessen. So steht für das juristische Studium insbesondere die Vorstellung im Raum, dass zwischen dem Erfolg in den Schulfächern Deutsch, Mathematik und Latein (bzw. erste fortgeführte Fremdsprache) und dem Erfolg im Jurastudium ein Zusammenhang besteht.
In seiner Untersuchung fanden insgesamt 224 Fälle in die Stichprobe Eingang, von denen jedoch wegen unzureichender Angaben zu den Schulnoten nur 212 Fälle auswertbar waren. Die ausgewerteten 212 Verfahren bilden knapp ein Viertel (22,6%) des im Jahr 2002 bearbeiteten Gesamtvolumens an Prüfungsfällen im Ersten Juristischen Staatsexamen in Niedersachsen.
Als ein Ergebnis ließ sich feststellen, dass sich zwischen der Abiturnote und dem Bestehen des Examens ein zwar nicht übermäßig starker, aber dennoch signifikanter Zusammenhang nachweisen lässt. Alle diejenigen, die in der Schule eine sehr gute Durchschnittsnote erreicht hatten, bestanden das Erste Juristische Staatsexamen, während von den „nur“ guten und den schwächeren Abiturientinnen und Abiturienten ein etwa gleicher Anteil (13,8 bzw. 17,6%) das Examen nicht bestand. Zwischen Abiturnote und den im Staatsexamen erreichten Punktzahlen besteht sogar ein hochsignifikanter Zusammenhang: Je besser die

Durchschnittsnote im Abitur ist, desto höher ist auch die Gesamtpunktzahl, die im Examen erreicht wurde.
Bei der Betrachtung der einzelnen Kurse der Oberstufen wurden die Auswirkungen des Kurssystems und der damit einhergehenden Wahlmöglichkeit für die Schüler deutlich. Von den als wichtige Prädiktoren zur Diskussion stehenden Fächern Deutsch, Mathematik und Latein (bzw. erste fortgeführte Fremdsprache) zeigte sich, dass nicht alle Abiturienten in den Jahrgangsstufen 12 und 13 in diesen Fächern ununterbrochen Unterricht hatten. Von den Probanden der Stichprobe hatten 94,8% während der vier letzten Halbjahre vor dem Abitur durchgängig Deutschkurse belegt, 80,2% durchgängig Mathematik und nur 19,3% Kurse in Latein sowie 79,7% Kurse in einer anderen fortgeführten Fremdsprache.
Insgesamt ergab sich zusammenfassend folgendes Bild: Die in den Fächern aus dem sprachlich-literarischen und dem gesellschaftswissenschaftlichen Aufgabenfeld erzielten Noten sind zur Vorhersage des Examensergebnisses nur eingeschränkt geeignet; eine bessere Prädiktorwirkung kommt den Noten aus dem mathematisch-naturwissenschaftlichen Aufgabenfeld zu. Im Vergleich zu der ungewichteten Durchschnittsnote im Abitur erweisen sich die in den Oberstufenkursen erzielten Teilleistungen nicht zwingend als das bessere Vorhersageinstrument.

Trotz umfangreicher Studien und teilweise hoher Korrelationen ist die prädiktive Validität der Schulnoten immer wieder umstritten gewesen. So untersuchte Althoff (1986) die Beziehungen zwischen Schulnoten und dem Kriterium Testleistung. Eine Stichprobe bestand aus 120 Abiturienten, die an einer zweitägigen Eignungsprüfung für den gehobenen Dienst bei Kommunalverwaltungen teilnahmen. Eine andere Stichprobe setzte sich aus 96 Realschülern zusammen, welche an einer eintägigen Eignungsprüfung für den mittleren Dienst bei Kommunalverwaltungen teilgenommen haben. In beiden Stichproben fand keine Vorselektion nach Zeugnisnoten statt.
Bei Betrachtung der Korrelationen fällt auf, dass nur ein geringer Teil der Korrelationen (14 von 392) bei der Stichprobe der Abiturienten statistisch signifikant geworden sind. Besonders die für die Teilnahme an Eignungsprüfung oft herangezogene Deutschnote hatte keine statistische Bedeutsamkeit. Ein ähnliches Bild ergab die Betrachtung der Korrelationen in der Stichprobe der Realschüler. Auch in diesem Falle konnten durch die Schulnoten keine Rückschlüsse auf das Abschneiden in der Eignungsprüfung gemacht werden.
In einer weiteren Untersuchung mit 56 Krankenpflegern, die aufgrund des Ergebnisses einer eintägigen Eignungsuntersuchung aus einer Gesamtgruppe nicht nach Zeugnisnoten

selegierten Stichprobe zur Ausbildung zugelassen wurden, sollte die Beziehung zwischen den Schulnoten und möglichen Kriterien des Berufserfolges, also die prognostische Validität, analysiert werden. Auch hier zeigten sich keinerlei signifikante Korrelationen, womit in diesem Falle durch das Zeugnis für die Krankenpflegeausbildung keinerlei Rückschlüsse auf das Ergebnis in der Krankenpflegeprüfung gezogen werden dürfen.
Für die Stichprobe der Bewerber für den gehobenen Dienst einer Landesbehörde und einer Stichprobe von 338 Regierungsanwärtern (Abitur bzw. Fachhochschulreife), die an einer Eignungsprüfung teilgenommen hatten, zeigten sich zumindest für die Mathematiknote eine Reihe signifikanter Korrelationen zu der Zwischenprüfung (Stichprobe der Bewerber für den gehobenen Dienst einer Landesbehörde) und der Inspektorprüfung (Stichprobe der Regierungsanwärter). Erneut zeigten sich keinerlei signifikante Korrelationen zwischen der Leistung in der jeweiligen Prüfung und der Deutschnote. Althoff (1986) zieht aus seinen Ergebnissen das Fazit, dass Schulnoten, abgesehen von der Mathematiknote, kein geeigneter und valider Prädiktor für Intelligenz- und Leistungsmerkmale, im weiteren Sinne, für die berufliche Bewährung, darstellen.

Die Schulzeugnisse von Bewerbern für ein Auswahlverfahren zur Einstellung in den Polizeivollzugsdienst verglich Petersen (2002). Insgesamt nahmen 689 Bewerber, die sich für eine Einstellung in den gehobenen Vollzugsdienst beworben hatten sowie 156 Bewerber für den mittleren Dienst, an der Untersuchung teil. Im Vorfeld fand keine Vorauswahl aufgrund der Zeugnisnoten oder anderer Kriterien statt, so dass eine Vorselektion vermieden werden konnte.
Bei den Bewerbern für den mittleren Polizeivollzugsdienst zeigte sich, dass weder die einzelnen Zeugnisnoten in den verschiedenen erhobenen Fächern signifikant mit dem erreichten Gesamtergebnis im Auswahlverfahren korrelierten, noch hinsichtlich des berechneten Notendurchschnittes der Einzelfächer. Bei den Bewerbern für den gehobenen Polizeivollzugsdienst korrelierte die Deutschnote signifikant ($r = .21$) mit dem Gesamtergebnis. Die von Althoff (1986) und anderen Autoren als validester Einzelprädiktor beschriebene Mathematiknote korrelierte weder mit den Ergebnissen der Intelligenz- und Leistungstests oder anderer Testteile noch mit dem Endergebnis signifikant. Auf der Einzelprädiktorebene konnten signifikante Korrelationen einzig zwischen der zweiten Fremdsprache und dem Ergebnis im Diktat ($r = .44$) sowie der Sportnote und dem Ergebnis in der Sportprüfung ($r = .32$) nachgewiesen werden.

Interessant ist jedoch das Ergebnis, welches die oben beschriebenen Ausführungen von Althoff (1986) und Ingenkamp (1969) bestätigen könnten, dass eine Vielzahl von signifikanten Korrelationen zwischen Schulnoten und Verhaltensratings von Psychologen in einem Rund- und Einzelgespräch gefunden wurden. Möglicherweise spielen bestimmte Persönlichkeits- und somit Verhaltensmerkmale (wie z.B. Anpassung, Sozialverhalten) bei der Zensurengebung eine größere Rolle als sie eigentlich innehaben sollten.

Mehr Übereinstimmung scheint es hinsichtlich der Prädiktion beruflicher Kriterien zu geben. Korrelationen zu beruflichen Leistungskriterien liegen nach Drumm (1992) nur noch zwischen r = .05 und r = .25. Trost und Kirchenkamp (1993) haben weder für Schulzeugnisnoten noch Examensnoten von Hochschulen Koeffizienten von größer als r = .10 im Vergleich mit betrieblichen Leistungskriterien finden können.

Zeugnisse vermitteln durch ihre vorrangige Bewertungsform der Note eine schnell zugängliche Information, die jedoch inhaltlich nicht immer eindeutig zu interpretieren ist. Probleme bestehen in der mangelnden notenbezogenen Vergleichbarkeit der Schulen und Hochschulen und der offenen Interpretation des Inhalts einer Note. Trotzdem oder gerade deswegen sind Noten, wie die genannten Analogieschlüsse zeigen, beliebte lndikatoren für Eignungsdiagnosen. So liegt die Gefahr nahe, die Bedeutung von Noten aufgrund ihres objektiven Anscheins zu überschätzen.
Wenn die Leistung eines Schülers nicht mit der eines Schülers derselben Stufe desgleichen Faches objektiv vergleichbar ist, ist nicht nur die pädagogische Funktion der Schulnoten als Indikatorvariable des Schulerfolges nicht zu erfüllen, sondern auch deren Prädiktorwert fragwürdig. Der schulische „Lebenslauf“ eines Kindes ist in höherem Maße abhängig von dem in der Klasse herrschenden Leistungsstandard und der diagnostischen Kompetenz des Lehrers und nicht in dem erforderlichen Maße von der eigenen Leistung.

2.3.2.3 Kriterien

Die Frage, wie geeignet eignungsdiagnostische Instrumente hinsichtlich ihrer Vorhersagegüte sind, steht in direkter Abhängigkeit von dem gewählten Kriterium. So ergeben sich bspw. teilweise erhebliche Unterschiede bei den mittleren Validitätskoeffizienten, je nach dem ob das Kriterium der Ausbildungs- oder der Berufserfolg ist (Schuler & Funke, 1989). Die Unterschiede sind in Tabelle 2.3-12 dargestellt.

Tab. 2.3-12: Mittlerer Validitätskoeffizient für zwei unterschiedliche Kriterien nach Schuler und Funke (1989)

	KRITERIUM	
PRÄDIKTOR	AUSBILDUNGSERFOLG	BERUFSERFOLG
Vorstellungsgespräch	.10	.14
Persönlichkeitstest	-	.15
Schulnoten	-	.15
Bewerbungsunterlagen	-	.18
Biographische Fragebogen	.30	.37
Assessment Center	-	.37
Arbeitsproben	-	.30
Kognitive Fähigkeitstests	.54	.45

Bei Betrachtung der Tabelle wird deutlich, das Schulnoten mit r. = 15 nur eine relativ geringe prädiktive Validität hinsichtlich des Kriteriums Berufserfolg besitzen und im Vergleich zu anderen Prädiktoren wie Kognitive Fähigkeitstests oder Arbeitsproben eine deutlich untergeordnete Rolle spielen. Betrachtet man dagegen ihre Vorhersagegüte hinsichtlich des Kriteriums Ausbildungserfolg, ergeben sie ein vollkommen anderes Bild und gewinnen trotz ihrer zwischenzeitlich nicht mehr unumstrittenen Vorhersagekraft an Bedeutung. So berichten wie schon erwähnt Baron-Boldt, Funke und Schuler (1989) über eine prädiktive Validität von r = .41 für den Ausbildungs- und r = .46 für den Studienerfolg.

Somit leistet die Wahl des Kriteriums schon in dieser Hinsicht einen entscheidenden Beitrag bei der Bewertung eines Prädiktors. Was darüber hinaus letztendlich als Erfolg in der Ausbildung bzw. im Beruf angesehen wird, d.h. die konkrete Operationalisierung des Kriteriums selbst, ist eine weitere Einflussgröße, die sowohl bei der Bewertung vorliegender Ergebnisse unterschiedlicher Studien berücksichtigt werden sollte, wie auch im Kontext der Auswahl von geeigneten Prädiktoren in der praktischen Eignungsdiagnostik.
Wie schon in Kapitel 2.1 beschrieben, steht vor der Konstruktion und Auswahl von geeigneten Prädiktoren und Kriterien die Arbeitsplatzanalyse mit dem Ziel, für beide eine empirische Basis zu erstellen (Kannheiser & Frieling, 1992). Mit den Prädiktoren werden Personenmerkmale erhoben, die späteres Verhalten, insbesondere der Erfolg eines Teilnehmers am Auswahlverfahren, weiter in der Ausbildung oder am Arbeitsplatz, vorhersagen können sollen. Die Prädiktoren sind eignungsdiagnostische Verfahren (vgl. Kapitel 2.3.2.1), die Persönlichkeitsmerkmale beliebiger Art sowie weitere Variablen zur Prognose des Berufserfolges operationalisieren und für die Vorhersage des Kriteriums

eingesetzt werden (Hossiep (1995); Maukisch, 1978). Kriterien dagegen bedeuten die Übersetzung der Zielvorstellung eines Personalauswahlverfahrens in Verhaltensdimensionen. Für sie gelten die gleichen methodischen Anforderungen wie für andere eignungsdiagnostische Instrumente auch, d.h. Objektivität, Reliabilität und Diskriminationsfähigkeit. Dabei ist nach Hossiep (1995) zu bedenken, dass die Validität von Kriterien nicht überprüft werden kann, da es sich hier nicht, wie bei den eignungsdiagnostischen Verfahren, um Konstruktionen handelt, die an der Wirklichkeit geprüft werden, sondern um die Wirklichkeit selbst. Wichtiger ist an dieser Stelle die Frage, ob die herausgesuchten Kriterien für den ausgesuchten Zweck wichtig und bedeutsam sind und für die Tätigkeit eine entsprechende Relevanz besitzen.

Die Frage, welches Kriterium den Ausbildungserfolg repräsentieren kann und für diesen bedeutsam und wichtig ist, wird häufig durch die Berücksichtigung der schulischen und praktischen Leistung der Auszubildenden beantwortet. So werden die Ergebnisse der internen oder externen Ausbildungsprüfungen herangezogen und können zu den entsprechenden Prädiktoren (z.B. Schulnoten oder Ergebnisse in den Auswahlverfahren) in Beziehung gesetzt werden. Für den Bereich des Polizeivollzuges ist es somit eine Möglichkeit, als Kriterium für den Ausbildungserfolg, die Ergebnisse der Abschlussprüfungen für die jeweiligen Laufbahnen im mittleren oder gehobenen Vollzugsdienst heranzuziehen. So können die erreichten Noten in den Zwischen- oder Abschlussprüfungen ein repräsentatives und bedeutsames Kriterium für den Prädiktor Gesamtergebnis im Auswahlverfahren darstellen. Es kann aber auch sinnvoll sein, einzelne Auswahlinstrumente einer eignungsdiagnostischen Batterie auszuwählen, um durch den Korrelationsschluss zu erfassen, inwieweit bspw. eine Teilleistung Deutschprüfung geeignet ist, die für den Ausbildungszweck notwendige Vorhersage der späteren Deutschleistung zu erschließen.

Neben der von Hossiep (1995) thematisierten Schwierigkeit, die Validität des Kriterium zu erschließen, darf auch die Frage der Reliabilität des Kriteriums nicht außer acht gelassen werden, da sie eine grundsätzliche Voraussetzung für die Validität darstellt. In diesem Zusammenhang dürften die Schwierigkeiten, die zu einer geringen Reliabilität von Schulnoten beitragen (s.a. Kapitel 2.3.2.2) u.a. auch für die Erhebung des Kriteriums Ausbildungserfolg gelten. Analog den Ergebnissen auf allgemeinen Schulen werden die Ausbildungsleistungen der zukünftigen Polizeivollzugsbeamten durch den Unterricht im Klassenverband oder Kursen und den damit einhergehenden schriftlichen und mündlichen

Prüfungen erschlossen. Ähnlich wie die Lehrer an allgemein bildenden Schulen unterliegen somit die Lehrenden in der Polizeiausbildung gleichen Urteils- und Bewertungsverzerrungen, welche die Reliabilität des Kriteriums negativ beeinflussen. Standardisierte Ankreuzverfahren (multiple choice) finden nur selten Anwendung, eine Operationalisierung der Beurteilung in einer mündlichen Prüfung, vergleichbar einer standardisierten Durchführung wie bei standardisierten Interviews, sind häufig nicht gebräuchlich. Vergleichbare Annahmen gelten hinsichtlich der Objektivität des Kriteriums Ausbildungserfolg.

Scheint es noch relativ einfach sich auf die Operationalisierung des Kriteriums Ausbildungserfolg zu einigen, da in der Regel angenommen werden kann, dass die Ausbildungsleistung durch den schulischen und praktischen Erfolg in der Ausbildung abgebildet werden kann – auch wenn hier schon Einschränkungen bei der unterschiedlichen Bewertung und Gewichtung der schulischen und praktischen Anteile und deren Wertigkeit für den jeweiligen Ausbildungserfolg bzw. spätere Relevanz für den Berufserfolg deutlich werden können – wird es inhaltlich offensichtlich schwieriger festzulegen, was das Kriterium Berufserfolg ausmacht und wie ein solcher operationalisiert werden kann.

Allein die Abhängigkeit beruflichen Erfolges von einer Vielzahl Faktoren macht eine Definition schwierig. So spielt zum einen familiäre und soziale Hintergründe sowie Einstellungen und Erwartungen, aber auch Ausbildungs- und Weiterbildungsmöglichkeiten eine entscheidende Rolle. Auch die gesellschaftlichen Rahmenbedingungen, die sich verändernde Situation auf dem Arbeitsmarkt und die Chancen, in beruflichen Feldern überhaupt erfolgreich tätig werden zu können, erschweren eine Festlegung. Eine Unterscheidung zwischen so genannten „harten“ und „weichen“ Kriterien (Rösler, 1992) ist nicht immer so eindeutig wie sie auf den ersten Blick scheint. Als harte Kriterien werden dabei die der direkten Beobachtung zugänglichen Verhaltensweisen wie Fehlzeiten, Kündigungen, Unfälle etc. bezeichnet. Im Bereich des Polizeivollzugsdienstes können auch Faktoren wie erreichte Beförderungen innerhalb eines gewissen Zeitrahmens oder Ermittlungserfolge, die sich im Rahmen der Kriminalitätsstatistik nachweisen und belegen lassen, dazu zählen. Weiche Kriterien sind bspw. Beurteilungen des Verhaltens einzelner Arbeitnehmer durch Vorgesetzte oder Kollegen.
Dabei sollte nicht außer acht bleiben, dass auch so genannte harte Kriterien subjektiven Wertungen unterliegen, denn bspw. wäre bei den Fehlzeiten auch zu bewerten, ob es sich um vermeidbares vs. nicht vermeidbares oder entschuldigtes vs. unentschuldigtes Fehlen handelt.

Ein Polizeibeamter kann bspw. für die Funktion eines Dienstgruppenleiters die persönliche, intellektuelle und soziale Kompetenz als voraussetzende Eignung mitbringen, wenn jedoch aus politischen Sparzwängen heraus keine Beförderungsstellen angeboten werden dürfen, wird er nicht in der Lage sein, diese belegen zu können. Zieht man in diesem Fall in einer Untersuchung zum Berufserfolg eines Polizeibeamten als Kriterium die normalerweise üblichen Beförderungen in einem bestimmten Zeitrahmen heran, so zählt dieser Proband in der Untersuchung trotz potentieller Fähigkeiten und Voraussetzungen für den Berufserfolg unter Umständen zu jenen Personen, die weniger erfolgreich waren. Darüber hinaus kann eine derartige von außen vorgegeben Blockade eigener Karrierepläne zu Frustrationen, Neid und Berufsunzufriedenheit und in der Folge zu insgesamt schlechteren Leistungen in der bisherigen Position führen. Eignungsrelevante Personenmerkmale können hier zur individuellen Berufserfolgsprognose nur insoweit beitragen, wie die Eignungsdiagnostik diese Merkmale erfassen kann und diese zwischen dem Zeitpunkt der Prognose und der Zeitspanne auf die sich diese wiederum bezieht, unverändert bleiben bzw. die Veränderungen regelhaft und damit vorhersagbar vonstatten gehen (Hossiep, 1995). Dies ist bei derartigen Rahmenbedingungen häufig nicht der Fall. Denn die Berufseignungsdiagnostik ist eine Statusdiagnostik, zur Stützung jeweils vorgefundener Merkmale oder Merkmalskombinationen und keine Prozessdiagnostik, die einen Veränderungsprozess begleitet.

Weiter sind die für den Erfolg einer Tätigkeit relevanten Merkmale so umfassend, dass sie kaum vollständig durch eine noch so gründliche Anforderungsanalyse erfasst werden können oder gar bekannt sein dürften. Erschwerend kommt hinzu, dass diese Merkmale schon gar nicht zum Zeitpunkt einer Entscheidung über Annahme oder Ablehnung eines Bewerbers vorliegen. Dies dürfte insbesondere für jene Berufe bedeutsam sein, deren Anforderungsprofil sich laufend verändert, weil es sich aktuellen Entwicklungen anpassen muss. So war es vor einigen Jahren für einen erfolgreichen Polizeibeamten eher Hobby, sich mit PC oder elektronischer Datenverarbeitung zu beschäftigen. In der heutigen Zeit ist eine auf elektronische Daten gestützte Ermittlungsarbeit nicht mehr aus dem Alltag eines Polizeibeamten weg zu denken und im Rahmen der Strafverfolgung im Bereich der Internetkriminalität werden sogar sehr spezielle Kenntnisse erwartet.

Als vergleichsweise stabil geltende Merkmale, die sich relativ gut durch psychologische Auswahlinstrumente erfassen lassen, gelten in erster Linie Intelligenz, allgemeine

Leistungsfähigkeit, soziale Kompetenz und Selbstvertrauen (Schuler & Funke, 1993). Im Management, dass im Rahmen des Polizeivollzugsdienstes den Führungskräften des gehobenen Dienstes und generell dem höheren Dienst entspricht, sind es nach Sarges und Weinert (1991) überdurchschnittliche intellektuelle Fähigkeiten, hohe soziale Kompetenz und einen prägnanten Leistungs- und Gestaltungswillen.

2.3.2.4 Zusammenwirken von Prädiktoren und Kriterien

Hossiep (1995) weist darauf hin, dass bei der Bewertung der Zusammenhänge zwischen Prädiktoren und Kriterien sich die Frage erhebt, wie die Relation zwischen beiden kausal interpretiert werden kann. Hier zeigt sich eine Bandbreite vom Konzept der strikten Kausalität bis hin zu einer Generalisierung einer empirisch ermittelten korrelativen Beziehung. Krapp (1979) konstatiert, dass im Rahmen einer Prognosestellung nur funktionale Abhängigkeiten interessieren, dabei sei es irrelevant, wie diese funktionalen Abhängigkeiten kausal erklärt werden können. Wichtig und Ziel der Prognose sei es, die im Entscheidungsvorgang beobachteten treatment-Effekte möglichst exakt vorher zu sagen. Hossiep (1995) stellt in Anlehnung an Krapp (1979) verschiedene typische Modellannahmen über den Zusammenhang von Prädiktor- und Kriteriumsvariablen dar.
Zum einen ist hier die Annahme einer strikt linearen Beziehung zwischen der Ausprägung des Prädiktors und der entsprechenden Kriteriumsausprägung (Abb. 2.3-4) zu nennen.

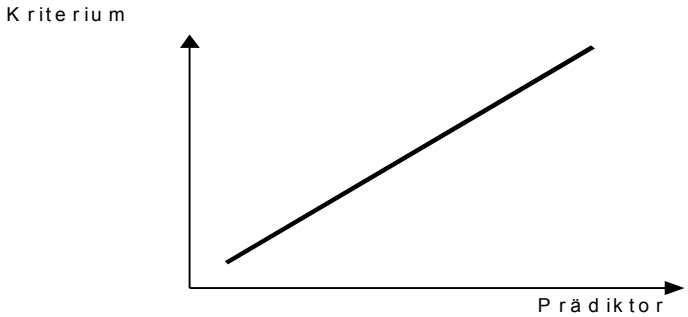


Abb. 2.3-4: Strikt lineare Beziehung zwischen Prädiktor und Kriterium

Die eingezeichnete Grade steht hierbei für einen linearen Zusammenhang zwischen bspw. dem Prädiktor „Intellektuelle Leistungsfähigkeit" und dem Kriterium „Berufserfolg". Je größer das durch den Probanden erreichte Ergebnis hinsichtlich des erfassten Prädiktors ist, desto größer ist das zu erwartende Niveau seiner Kriteriumsleistung, d.h. desto größer ist sein beruflicher Erfolg. Erreicht ein Bewerber im IST 2000 einen Stanine-Wert von 9, so ist zu

erwarten, dass sein zukünftiger Berufserfolg größer sein wird als der des Mitbewerbers mit einem Stanine Wert von 7. Eine mögliche Leistungsobergrenze findet hier keine Berücksichtigung.

Die in Abbildung 2.3-4 dargestellte lineare Beziehung scheint nicht sehr wahrscheinlich. Es ist eher anzunehmen, dass keine strikt lineare Beziehung zu erwarten ist, sondern ein Zusammenhang zwischen Prädiktor und Kriterium, der sich kurvenlinear darstellt (Abbildung 2.3-5).

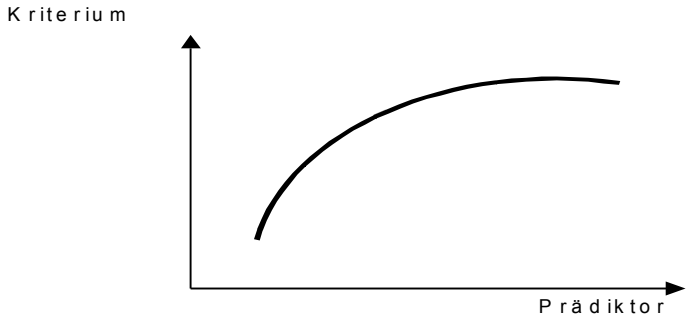


Abb. 2.3-5: Kurvenlineare Beziehung zwischen Prädiktor und Kriterium

In dieser Abbildung wird deutlich, dass die zunehmende Höhe hinsichtlich des erfassten Prädiktors nicht gleichsam ein immer größer werdendes Niveau der Kriteriumsleistung mit sich bringt, somit Prädiktorwert und Kriteriumswert nicht im gleichen Maße ansteigen und damit auch eine potentielle Leistungsgrenze Berücksichtigung findet. In der Praxis würde dies bedeuten, dass es relativ unerheblich für die Ausprägung des Berufserfolges ist, hier also kein Zuwachs mehr zu erkennen ist, wenn eine gewisse Grundhöhe erreicht worden ist. So wäre bspw. zu erwarten, dass bis zu einem Stanine-Wert von 7 ein deutlicher Zusammenhang zwischen den erreichten Leistungen der Bewerber im Intelligenztest gegeben ist, eine anschließende darüber liegende Differenzierung jedoch keinen Sinn macht. Althoff (1984) verweist darauf, dass kurvenlineare Beziehungen im oberen Bereich der Standardwerte von Leistungstests keine Seltenheit sind. Sie lassen sich möglicherweise durch Unterforderung erklären bzw. dadurch, dass in vielen Berufen über die nachweisbaren Zusammenhänge zwischen kognitiven Leistungsfähigkeiten und Berufserfolg hinaus, eine Vielzahl anderer Faktoren – Motivation, Einstellung, sonstige Kenntnisse, Vorgesetzte – eine erfolgreiche Berufsausübung bestimmen.

Eine weitere Alternative über die Bewertung der Beziehung zwischen dem erreichten Prädiktorwert und dem Kriterium ist probabilistisch. Hier steht die Annahme im Hintergrund, dass bei gleicher Prädiktorausprägung eines Bewerbers dessen Kriteriumsausprägung durchaus unterschiedlich sein kann. Dabei sind diese unterschiedlichen Ausprägungen jedoch mit Wahrscheinlichkeiten behaftet. Auch hier gäbe es wieder unterschiedliche Möglichkeiten, um Wahrscheinlichkeitsfunktionen darzustellen (Abb. 2.3-6).

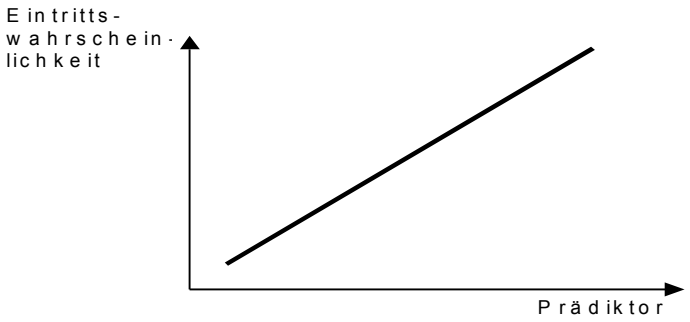


Abb. 2.3-6: Wahrscheinlichkeitsfunktion für Prädiktor und Kriterium

Eine lineare Darstellung zeigt, dass mit ansteigender Größe des Prädiktors die Erfolgswahrscheinlichkeit kontinuierlich ansteigt.

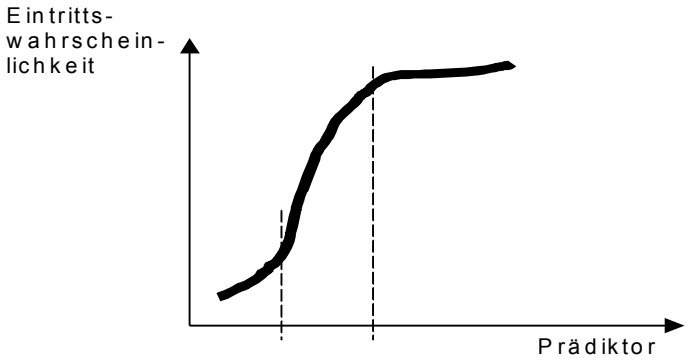


Abb. 2.3-7: Wahrscheinlichkeitsfunktion für Prädiktor und Kriterium

In Abbildung 2.3-7 wird beispielhaft das so genannte Schwellen-Wert-Konzept dargestellt. Hier sind die Auswirkungen der unterschiedlichen Prädiktorausprägungen auf das Kriterium im unteren und oberen Bereich annähernd vergleichbar. Im unteren Bereich liegt die Erfolgswahrscheinlichkeit ungefähr gleich Null, im oberen Bereich annähernd bei 1,0. Dagegen gibt es eine deutliche Veränderung der Ausprägung zwischen beiden Bereichen der Ausprägung. Übertragen auf die Praxis könnte dies wie folgt aussehen: Ein im IST 2000

durch einen Stellenbewerber erworbener Stanine-Wert von 1-2 bedeutet in etwa eine gleich niedrige Erfolgswahrscheinlichkeit für den prognostizierten Beruferfolg. Analog verhält es sich in den Bereichen 8-9, die für eine vergleichbar hohe berufliche Erfolgsaussicht sprechen. Dagegen verändert sich die Erfolgswahrscheinlich in den Bereichen 3-7 erheblich und die Bedeutung der Diskriminierung einzelner Stanine-Werte ist für die Selektion oder Platzierung der Bewerber entscheidend.
Über die formulierten Beispiele hinaus finden sich in der Fachliteratur weitere diesen Rahmen übersteigende Beispiele zum Zusammenwirken zwischen Prädiktor und Kriterium. Genannt seien in diesem Zusammenhang nur beispielhaft multiple additive Beziehungen, die Interaktionsbeziehungen oder das multiple Schwellenwertkonzept.

3. Ableitung der Fragestellung

3.1 Überblick und Einführung

Im vorausgehenden Kapitel 2 wurde dargestellt, dass in einer Vielzahl von Studien in unterschiedlichen Kontexten der grundsätzliche Nutzen eignungsdiagnostischer Verfahren und damit deren prädiktive Validität belegt werden konnte. So besitzen bspw. häufig Anwendung findende Verfahren wie kognitive Leistungstests (r = .51), Arbeitsproben (r = .54), strukturiert geführte Interviews (r = .51) oder Assessment Center (r = .36) eine zufrieden stellend hohe prädiktive Validität, um das Ziel ihrer Anwendung zu gewährleisten, d.h. für eine bestimmte Funktion ausreichend geeignete Bewerber auszuwählen und nicht geeignete zurückzuweisen. Wie in Kapitel 2.3 beschrieben, ist die grundlegende Basis hierfür, die Auswahl geeigneter Prädiktoren (z.B. o.g. Auswahlinstrumente). Zum anderen ist jedoch die Bestimmung geeigneter Kriterien, an denen der Erfolg festgemacht werden kann, genau so wichtig. Die erschlossene Höhe des ermittelten Validitätskoeffizienten hängt allerdings wesentlich von verschiedenen Faktoren ab. Einerseits der Grad dessen, was an Gemeinsamkeit zwischen Prädiktor (hier Testergebnis im Auswahlverfahren) und Kriterium (hier Ausbildungsergebnis nach bestandener Auswahl) erfasst wird und in der Regel als die Zulänglichkeit eines Tests bezeichnet wird. Andererseits sind die Reliabilitäten von Test und Kriterium ebenso von entscheidender Bedeutung und dies trotz teilweise methodisch bedingter Unvereinbarkeit von Reliabilität und Validität (homogene Aufgaben in einem Testverfahren führen zu einer höheren Reliabilität, heterogene Aufgaben dagegen zu einer höheren Validität). Die Vorhersageleistung eines unreliablen Verfahrens ist auf alle Fälle

gering, die eines hochreliablen steht jedoch in Abhängigkeit von der Reliabilität des Kriteriums (Lienert, 1989). Die Höhe der Kriteriumsvalidität variiert allerdings in unterschiedlichen Untersuchungen. Dies steht zum einen in Abhängigkeit vom gewählten Kriterium sowie der jeweiligen Reliabilitäten von Prädiktor und Kriterium im Zusammenhang. Andererseits sind bei der Bewertung statistische Artefakte und Messfehler (z.B. kleine Stichprobengrößen) als Erklärung heran zu ziehen.

Neben der Maximierung der prädiktiven Validität ist es zentrales Ziel der eignungsdiagnostischen Praxis, aus einer Vielzahl diagnostischer Instrumente diejenigen auszuwählen, die mit einem möglichst kleinen Aufwand und geringen Kosten eine größtmögliche Enge des Zusammenhanges zwischen gewählten Prädiktoren und Kriterien aufzeigen. Damit ist eine wichtige Frage im Rahmen der Beurteilung von Auswahlverfahren, inwieweit die aktuell durchgeführten Testbausteine (siehe Kapitel 2.2) diesem Anspruch gerecht werden können. Außerdem sollte weiter erfasst werden, ob die zusammengestellte Testbatterie auch Aspekte der inkrementellen Validität berücksichtigt.

Wie erwähnt, ist der zentrale Punkt für alle Überlegungen zur Bewertung eignungsdiagnostischer Testverfahren und Methoden und für die sich daraus ableitenden Entscheidungen hinsichtlich der Auswahl, der Zusammenstellung zu Testbatterien und des Einsatzes dieser, die Enge des Zusammenhanges zwischen Prädiktoren und Kriterien, folglich die prognostische Validität. Die Validierung kann somit nach Hossiep (1995) als ein allgemeines Prinzip angesehen werden, welches den Erkenntniswert einer diagnostischen Vorgehensweise erforscht. Anhand der Analyse der prädiktiven Validität können beispielsweise unter einer Vielzahl diagnostischer Indikatoren diejenigen ausgewählt werden, welche bei einem Minimum an Aufwand und Kosten eine größtmögliche Enge des Zusammenhanges zwischen Prädiktoren (Schulnoten oder Testverfahren) und dem Kriterium (Erfolg im Auswahlverfahren oder Ausbildung / Studium) versprechen. Jedoch handelt es sich nach Melahmed (1992) um die Ermittelung von Validitäten, die dynamisch sind. Die Anforderungen, wie bspw. in diesem Fall an ein Studium an der Fachhochschule für Öffentliche Verwaltung im Rahmen der Ausbildung zum gehobenen Polizeivollzugsdienst sowie jene zur Ausübung des Polizeiberufes, unterliegen mit der Zeit sich verändernden Erwartungen. Gleiches gilt für den Kreis der Interessenten, die sich für ein solches Studium entscheiden. Insofern ist es wichtig, generell regelmäßig zu überprüfen, inwieweit ein Auswahlverfahren seinem Zweck gerecht wird, d.h. die passenden Bewerber im Vorfeld auszuwählen und dabei die nicht geeigneten Bewerber zurückzuweisen. Einerseits ist ein

Ausbildungsinstitut wie die Fachhochschule interessiert, ein valides Auswahlinstrument vorzufinden, mit dessen Hilfe potentiell erfolgreiche Studenten erfasst werden, die auch den späteren umfangreichen Aufgaben im Studium gerecht werden können. Außerdem gebietet es andererseits die Testfairness, dass diejenigen Bewerber, welche den Anforderungen von Studium und Polizeiausbildung am ehesten gerecht werden, auch wirklich die Auswahl bestehen, d.h. nicht irrtümlich zurückgewiesen werden. Dabei ist es genau so wichtig, jenen Bewerbern, die aufgrund ihres erreichten Ergebnisses im Auswahlverfahren und eines somit möglicherweise zu antizipierenden schlechteren Abschneidens oder Scheiterns in der Ausbildung bzw. im späteren Beruf, dieses zeitig deutlich zu machen und ihnen somit Enttäuschungen zu ersparen.
Vordringlicher Zweck der vorliegenden Arbeit war es, die prognostische Validität des aktuellen Auswahlverfahrens für die Einstellung in den gehobenen Polizeivollzugsdienst, d.h. für ein Studium an der Fachhochschule für Öffentliche Verwaltung – Fachbereich Polizei - zu erschließen. Auch hier gilt es weiter, wie grundsätzlich für alle Unternehmen, dessen Wirtschaftlichkeit im Sinne eines dynamischen Konzeptes regelmäßig zu überprüfen und somit die aktuelle Güte eines Auswahlverfahrens zu evaluieren. Ziel dieser Evaluation, d.h. der Beurteilung und in der Folge möglicherweise auch die Verbesserung von praktischen Maßnahmen, ist es, nicht ausreichend valide Testbestandteile aus dem Auswahlverfahren zu entfernen oder durch Testverfahren mit höherer prognostischer Validität zu ersetzen. Denn nur mit einem prognostisch validen Testverfahren werden diejenigen Bewerber ausgewählt, die für das Unternehmen einen „Gewinn“ bringen, „längere Zeit“ im Unternehmen bleiben und den Anforderungen der Berufsausübung gerecht werden können. Gleichzeitig können jene Bewerber früher selektiert werden, die keine ausreichenden Leistungen erbringen, das Unternehmen möglicherweise kurzfristig, trotz hoher Ausbildungskosten, wieder verlassen oder den Anforderungen der Berufsausübung nicht gerecht werden.

Die Beurteilung eines Auswahlverfahrens muss aber auch der sich mit der Zeit verändernden Rahmenbedingungen für die beruflichen Anforderungen Rechnung tragen, so dass die Güte der gewählten Prädiktoren und Kriterien aktuellen Entwicklungen regelmäßig angepasst werden können. Weiter bedarf eine grundlegende Erweiterung der Ziel-Population der Bewerber, wenn bspw. wie im untersuchten Verfahren erreicht werden soll, mehr Bewerber einer nicht deutschen Nationalität zu gewinnen, einer erneuten Überprüfung der Güte des Auswahlverfahrens.

Steigende Personalkosten und Bedarf an geeigneten Bewerbern unterstreichen die Notwendigkeit einer wissenschaftlich fundierten Personalauswahl. Aufgrund betriebswirtschaftlicher Gesichtspunkte bekommt der Aufwand, der notwendig ist, um geeignete Interessenten für eine vakante Position auszuwählen, eine besondere ökonomische Bedeutung zu. Angesichts der zunehmenden Komplexität in einer heute hoch technisierten und anspruchsvollen Berufswelt und der damit einhergehenden Anforderungen an die zukünftigen Mitarbeiter einer Organisation sowie der Erfahrungen, dass kostenintensiv ausgebildete Mitarbeiter ein Berufsleben lang weitere Investitionen für die Organisation rechtfertigen können müssen – Stichworte sind hier Human Ressource Management und strategische Personalentwicklung – ist eine professionelle Personalauswahl im betriebswirtschaftlichen Interesse der Organisation aber auch des Bewerbers.

Wenn sich ein Auswahlverfahren hinsichtlich der Kosten-Nutzen Aspekte rentieren soll, ist eine wichtige Voraussetzung, dass es keine anderen Datenträger gibt, mittels derer mit einem geringeren Aufwand eine gleich hohe Güte in der Voraussage erzielt werden kann. In diesem Zusammenhang wird immer wieder die Bedeutung von Schulnoten kontrovers diskutiert. In unterschiedlichen Bundesländern werden Bewerbungen für die Ausbildung im Polizeivollzugsdienst nur angenommen, wenn die Interessenten in bestimmten Fächerkombinationen (häufig sind es die Noten in Deutsch, Mathematik und Sport) einen geforderten Mindestdurchschnitt vorweisen können. So erhebt sich die Frage, ob Schulnoten als valide Prädiktoren des beruflichen Erfolges bzw. des Erfolges im Rahmen eines Studiums / einer Ausbildung angesehen werden und somit aufwendige und kostspielige Auswahlverfahren ersetzen oder als Vorauswahlinstrument einen geeigneten Beitrag zu einer erfolgreichen Vorselektion bieten können. Die in verschiedenen Untersuchungen (vgl. a. Althoff, 1986; Ingenkamp, 1969; Petersen, 2002; Schuler, 1998) aufgetretenen Zweifel begründen sich möglicherweise u.a. durch unterschiedliche nationale Schulsysteme mit ungleichen Anforderungen an die Schüler sowie allgemeine Probleme bei der vergleichenden Leistungsbeurteilung der Schüler durch Lehrer. Wenn es aber einzelne Schulnoten, wie z.B. die Note in Mathematik, oder Schulnotenkombinationen geben sollte, welche eine hohe prognostische Validität besitzen, bliebe zu hinterfragen, inwieweit dies auf den hiesigen Kontext – Auswahl für den gehobenen Polizeivollzugsdienst – übertragbar wäre und falls eine ausreichend hohe prognostische Validität erreicht werden würde, inwieweit diese zu konkreten Veränderungen im Auswahlverfahren führen könnte.

Ist es das politische Ziel, vermehrt Bürger ausländischer Herkunft für die Aufgaben des Polizeivollzugsdienstes zu gewinnen und führt diese Vorgabe zu Veränderungen des Auswahlverfahrens - Einführung eines kulturfairen Leistungstests - erhebt sich die Frage, ob diese Veränderung wirklich dazu beitragen kann, das gestellte Ziel zu erreichen. Bei der Zusammenstellung der einzelnen Testverfahren wurde deshalb berücksichtigt, dass ein sprachgebundener kognitiver Leistungstest, der Fertigkeiten und Wissen der deutschen Sprache und Kultur erfordert, unter Umständen bei nicht deutschen Bewerbern zu niedrigeren Leistungen aufgrund anderer Herkunft, Muttersprache und Sozialisation führen kann. Damit dieser Aspekt keine Fehlinterpretationen im Sinne einer geringeren Leistungsfähigkeit und eines geringeren Intelligenzniveaus mit sich bringt, wurde ein kulturfairer kognitiver Leistungstest eingeführt. Insofern ist die Beantwortung der Frage sinnvoll, inwieweit sich der Einsatz des Verfahrens im genannten Sinne rechtfertigt. Eine Beantwortung erfolgt durch den Vergleich der im Auswahlverfahren erzielten Ergebnisse beider Gruppen sowohl im kulturfairen wie auch im nicht kulturfairen kognitiven Leistungstest.

3.2 Konkretisierung der allgemeinen Fragestellungen

Aus den Darstellungen im einleitenden Kapitel ergeben sich für die Untersuchung im Rahmen der Beurteilung eines Auswahlverfahrens insgesamt drei relevante Fragestellungen:

1. Ermittlung der prädiktiven Validität von Schulnoten hinsichtlich ihrer prognostischen Bedeutung für das Auswahlverfahren sowie das sich anschließende Studium. Können Schulnoten möglicherweise die derzeit Verwendung findenden eignungsdiagnostischen Instrumente ergänzen oder gar ersetzen? Haben Schulnoten eine höhere Güte als die derzeit verwendeten eignungsdiagnostischen Instrumente? Haben Schulnoten eine ausreichend gute Kosten-Nutzen Relation?

2. Ermittlung der prädiktiven Validität des Auswahlverfahrens hinsichtlich seiner prognostischen Bedeutung für den Erfolg im Studium. Besitzen die derzeit Verwendung findenden Instrumente des Auswahlverfahrens eine zufrieden stellend hohe prädiktive Validität? Ist die Zusammenstellung des Auswahlverfahrens unter Berücksichtigung der inkrementellen Validität seiner Instrumente sinnvoll?

3. Vergleich der Ergebnisse deutscher und nicht deutscher Teilnehmer des Auswahlverfahrens in einem kulturfairen und einem nicht kulturfairen kognitiven Leistungstest. Unterscheiden sich deutsche und nicht deutsche Teilnehmer hinsichtlich ihrer Ergebnisse in den beiden angewandten Instrumenten? Kann mit einem eingeführten kulturfairen kognitiven Leistungstest das politisch vorgegebene Ziel erreicht werden, mehr Bewerber nicht deutscher Abstammung für den Polizeidienst zu gewinnen?

In der ersten Fragestellung soll die prognostische Validität der vorliegenden Schulnoten von den Bewerbern um einen Ausbildungsplatz für den gehobenen Polizeivollzugsdienst erfasst werden. Dazu werden zum einen die Noten einzelner Schulfächer, zum anderen die errechnete Durchschnittsnote dieser Fächer aus den verfügbaren Zeugnissen entnommen. Neben diesem Prädiktor wird als Kriterium zur Beantwortung der Fragestellung der Erfolg der Bewerber im Auswahlverfahren herangezogen. Weiter werden im zweiten Abschnitt die Schulnoten jener Bewerber erneut als Prädiktor verwandt, die sich aufgrund ihrer erreichten Ergebnisse im Auswahlverfahren für die Einstellung in den gehobenen Polizeivollzugsdienst haben qualifizieren können und ein Jahr später das Studium an der Fachhochschule aufgenommen haben. Als Kriterium dient in diesem Fall die Gesamtnote der Zwischenprüfung im Studium. Abschließend wird die Kosten-Nutzen-Relation des Prädiktors Schulnoten anhand der in Abschnitt 2.1 beschriebenen Payoff-Funktion des Brodgen-Cronbach-Gleser-Modells berechnet.

Zur Erschließung der prädiktiven Validität des Auswahlverfahrens bilden die Ergebnisse der Bewerber in den einzelnen Instrumenten des Auswahlverfahrens sowie das erreichte Gesamtergebnis die Prädiktoren ab. Dabei werden die Daten jener Bewerber berücksichtigt, die aufgrund ihrer erreichten Ergebnisse im Rahmen der sequentiellen Auswahlstrategie an allen Testbausteinen haben teilnehmen können. Als Kriterium werden erneut die Ergebnisse jener Bewerber, die sich aufgrund ihres Gesamtergebnisses für die Einstellung qualifizieren konnten und das Studium aufgenommen haben, herangezogen. Kriteriumsvariable ist dabei der Studienerfolg der Studenten in der Zwischenprüfung in Form der erreichten Gesamtpunktzahl. Zudem werden in einem weiteren Schritt die einzelnen Instrumente des Auswahlverfahrens hinsichtlich ihrer inkrementellen Validität untersucht.

Zum Vergleich der Ergebnisse der Bewerber deutscher und nicht deutscher Herkunft werden deren im Auswahlverfahren erreichten Ergebnisse in den beiden Anwendung findenden

kognitiven Leistungstests (kulturfair sowie nicht kulturfair) gegenübergestellt und die Frage erhoben, ob es hier signifikante Leistungsunterschiede gibt. Es soll die Frage beantwortet werden, inwieweit die Einführung eines kulturfairen Testverfahrens dazu beiträgt, evtl. vorhandene kulturell bedingte Benachteiligungen ausländischer Bewerber auszugleichen.

3.3 Formulierung der Hypothesen

3.3.1 Hypothesen zur Fragestellung 1:

Analyse der prognostischen Validität von Schulnoten

3.3.1.1 Hypothesen zur Analyse der prognostischen Validität einzelner Schulnoten bezüglich des Gesamtergebnisses im Auswahlverfahren

Hypothese 1

Zwischen der vorliegenden Deutschnote im Zeugnis der Bewerber und dem Gesamtergebnis im Auswahlverfahren besteht ein statistisch signifikanter negativer Zusammenhang, d.h. eine niedrige Deutschnote steht in Beziehung mit einer höheren Gesamtpunktzahl im Auswahlverfahren et vice versa.

Hypothese 2

Zwischen der vorliegenden Mathematiknote im Zeugnis der Bewerber und ihrem Gesamtergebnis im Auswahlverfahren besteht ein statistisch signifikanter negativer Zusammenhang, d.h. eine niedrige Mathematiknote steht in Beziehung mit einer höheren Gesamtpunktzahl im Auswahlverfahren et vice versa.

Hypothese 3

Zwischen der vorliegenden Note im Fach Englisch im Zeugnis der Bewerber und ihrem Gesamtergebnis im Auswahlverfahren besteht ein statistisch signifikanter negativer Zusammenhang, d.h. eine niedrige Englischnote steht in Beziehung mit einer höheren Gesamtpunktzahl im Auswahlverfahren et vice versa.

Hypothese 4

Zwischen der vorliegenden Note im Fach Geschichte, Wirtschaft und Politik im Zeugnis der Bewerber und ihrem Gesamtergebnis im Auswahlverfahren besteht ein statistisch signifikanter negativer Zusammenhang, d.h. eine niedrige Note in Geschichte, Wirtschaft und Politik steht in Beziehung mit einer höheren Gesamtpunktzahl im Auswahlverfahren et vice versa.

Hypothese 5

Zwischen der vorliegenden Note im Fach Sport im Zeugnis der Bewerber und ihrem Gesamtergebnis im Auswahlverfahren besteht ein statistisch signifikanter negativer Zusammenhang, d.h. eine niedrige Sportnote steht in Beziehung mit einer höheren Gesamtpunktzahl im Auswahlverfahren et vice versa.

Hypothese 6

Zwischen dem errechneten Durchschnittswert der vorliegenden Noten im Zeugnis der Bewerber und ihrem Gesamtergebnis im Auswahlverfahren besteht ein statistisch signifikanter negativer Zusammenhang, d.h. eine niedrige Durchschnittsnote steht in Beziehung mit einer höheren Gesamtpunktzahl im Auswahlverfahren et vice versa.

3.3.1.2 Hypothesen zur Analyse der prognostischen Validität einzelner Schulnoten bezüglich der Gesamtnote in der Zwischenprüfung

Hypothese 1

Zwischen der vorliegenden Deutschnote im Zeugnis der Bewerber und der Gesamtnote in der Zwischenprüfung besteht ein statistisch signifikanter positiver Zusammenhang, d.h. eine niedrige Deutschnote steht in Beziehung mit einer niedrigen Gesamtnote in der Zwischenprüfung im Studium et vice versa.

Hypothese 2

Zwischen der vorliegenden Mathematiknote im Zeugnis der Bewerber und der Gesamtnote in der Zwischenprüfung besteht ein statistisch positiver signifikanter Zusammenhang, d.h. eine niedrige Mathematiknote steht in Beziehung mit einer niedrigen Gesamtnote in der Zwischenprüfung im Studium et vice versa.

Hypothese 3

Zwischen der vorliegenden Note im Fach Englisch im Zeugnis der Bewerber und der Gesamtnote in der Zwischenprüfung besteht ein statistisch signifikanter positiver Zusammenhang, d.h. eine niedrige Englischnote steht in Beziehung mit einer niedrigen Gesamtnote in der Zwischenprüfung im Studium et vice versa.

Hypothese 4

Zwischen der vorliegenden Note im Fach Geschichte, Wirtschaft und Politik im Zeugnis der Bewerber und der Gesamtnote in der Zwischenprüfung besteht ein statistisch signifikanter positiver Zusammenhang, d.h. eine niedrige Note in Geschichte, Wirtschaft und Politik steht

in Beziehung mit einer niedrigen Gesamtnote in der Zwischenprüfung im Studium et vice versa.

Hypothese 5

Zwischen der vorliegenden Note im Fach Sport im Zeugnis der Bewerber und der Gesamtnote in der Zwischenprüfung besteht ein statistisch signifikanter positiver Zusammenhang, d.h. eine niedrige Sportnote steht in Beziehung mit einer niedrigen Gesamtnote in der Zwischenprüfung im Studium et vice versa.

Hypothese 6

Zwischen der errechneten Durchschnittsnote im Zeugnis der Bewerber und der Gesamtnote in der Zwischenprüfung besteht ein statistisch signifikanter positiver Zusammenhang, d.h. eine niedrige Durchschnittsnote steht in Beziehung mit einer niedrigen Gesamtnote in der Zwischenprüfung im Studium et vice versa.

3.3.2 Hypothesen zur Fragestellung 2:

Analyse der prognostischen Validität des Auswahlverfahrens

3.3.2.1 Hypothesen zur Analyse der prognostischen Validität einzelner Testbausteine des Auswahlverfahrens bezüglich der Gesamtpunktzahl in der Zwischenprüfung

Hypothese 1

Zwischen dem Ergebnis im Testbaustein „Lückendiktat“ des Auswahlverfahrens und der Gesamtpunktzahl in der Zwischenprüfung besteht ein statistisch signifikanter positiver Zusammenhang. Eine höhere Punktzahl im Lückendiktat steht in Beziehung zu einer höheren Gesamtpunktzahl in der Zwischenprüfung et vice versa.

Hypothese 2

Zwischen dem Ergebnis im Testbaustein „Bericht-Deutschleistung“ des Auswahlverfahrens und der Gesamtpunktzahl in der Zwischenprüfung besteht ein statistisch signifikanter positiver Zusammenhang. Eine höhere Punktzahl für die Deutschleistung im Bericht steht in Beziehung zu einer höheren Gesamtpunktzahl in der Zwischenprüfung et vice versa.

Hypothese 3

Zwischen dem Ergebnis im Testbaustein „Bericht-Inhalt“ des Auswahlverfahrens und der Gesamtpunktzahl in der Zwischenprüfung besteht ein statistisch signifikanter positiver Zusammenhang. Eine höhere Punktzahl für den Inhalt im Bericht steht in Beziehung zu einer höheren Gesamtpunktzahl in der Zwischenprüfung et vice versa.

Hypothese 4

Zwischen dem Ergebnis im Testbaustein „Kulturfairer kognitiver Leistungstest“ des Auswahlverfahrens und der Gesamtpunktzahl in der Zwischenprüfung besteht ein statistisch signifikanter positiver Zusammenhang. Eine höhere Punktzahl im kulturfairen Testverfahren steht in Beziehung zu einer höheren Gesamtpunktzahl in der Zwischenprüfung et vice versa.

Hypothese 5

Zwischen dem Ergebnis im Testbaustein „Sporttest“ des Auswahlverfahrens und der Gesamtpunktzahl in der Zwischenprüfung besteht ein statistisch signifikanter positiver Zusammenhang. Eine höhere Punktzahl im Sporttest steht in Beziehung zu einer höheren Gesamtpunktzahl in der Zwischenprüfung et vice versa.

Hypothese 6

Zwischen dem Ergebnis im Testbaustein „Nicht kulturfairer kognitiver Leistungstest“ des Auswahlverfahrens und der Gesamtpunktzahl in der Zwischenprüfung besteht ein statistisch signifikanter positiver Zusammenhang. Eine höhere Punktzahl im kognitiven Leistungstest steht in Beziehung zu einer höheren Gesamtpunktzahl in der Zwischenprüfung et vice versa.

Hypothese 7

Zwischen dem Ergebnis im Testbaustein „Vorstellungsgespräch“ des Auswahlverfahrens und der Gesamtpunktzahl in der Zwischenprüfung besteht ein statistisch signifikanter positiver Zusammenhang. Eine höhere Punktzahl im Vorstellungsgespräch steht in Beziehung zu einer höheren Gesamtpunktzahl in der Zwischenprüfung et vice versa.

Hypothese 8

Zwischen dem Ergebnis im Testbaustein „Rundgespräch“ des Auswahlverfahrens und der Gesamtpunktzahl in der Zwischenprüfung besteht ein statistisch signifikanter positiver Zusammenhang. Eine höhere Punktzahl im Rundgespräch steht in Beziehung zu einer höheren Gesamtpunktzahl in der Zwischenprüfung et vice versa.

3.3.2.2 Hypothese zur Analyse der prognostischen Validität des Gesamtergebnisses im Auswahlverfahren bezüglich der Gesamtpunktzahl in der Zwischenprüfung

Hypothese 1

Zwischen dem Gesamtergebnis im Auswahlverfahren und der Gesamtpunktzahl in der Zwischenprüfung besteht ein statistisch signifikanter positiver Zusammenhang. Ein höhere Gesamtpunktzahl im Auswahlverfahren lässt eine höhere Gesamtpunktzahl in der Zwischenprüfung erwarten et vice versa.

3.3.2.3 Hypothese zur Analyse der inkrementellen Validität des Auswahlverfahrens

Hypothese 1

Die Vorhersagekraft aller Teile des Auswahlverfahrens zusammengenommen ist statistisch signifikant höher als die Vorhersagekraft einzelner Teile, d.h. die Kombination der einzelnen Testteile zu einem Gesamtauswahlverfahren findet seine Rechtfertigung und hat als Ganzes eine zufrieden stellend hohe Validität.

3.3.3 Hypothesen zur Fragestellung 3:

Vergleich der Ergebnisse deutscher und nicht deutscher Teilnehmer am Auswahlverfahren

3.3.3.1 Hypothese zur Analyse der Leistungsunterschiede zwischen den deutschen und den nicht deutschen Teilnehmern in einem kulturfairen kognitiven Leistungstest

Hypothese 1

Es gibt keinen statistisch signifikanten Unterschied zwischen den Ergebnissen deutscher und nicht deutscher Teilnehmer am Auswahlverfahren hinsichtlich ihrer Ergebnisse in einem kulturfairen Testverfahren.

3.3.3.2 Hypothese zur Analyse der Leistungsunterschiede zwischen den deutschen und den nicht deutschen Teilnehmern in einem nicht kulturfairen kognitiven Leistungstest

Hypothese 1

Es gibt einen statistisch signifikanten Unterschied zwischen den Ergebnissen deutscher und nicht deutscher Teilnehmer am Auswahlverfahren hinsichtlich ihrer Ergebnisse in einem nicht kulturfairen Testverfahren.

Neben der zentralen Frage nach der prognostischen Validität eines Auswahlverfahrens, ist ein weiterer wichtiger Faktor die Kosten-Nutzen-Relation. Wie bereits in Abschnitt 2.1 dargestellt, spielen Faktoren wie bspw. die Kosten des Auswahlverfahrens pro Bewerber und die geschätzte Verweildauer der Bewerber in einer Institution eine bedeutende Rolle. Anhand der so genannten Payoff-Funktion des Brodgen-Cronbach-Gleser-Modells (zit. n. Höft, 2001,

Kersting, 2004) lässt sich der Nutzenzuwachs, der durch ein Testverfahren entsteht, rechnerisch ermitteln. Mittels dieser Nutzenfunktion soll im Rahmen der Beurteilung des Auswahlverfahrens erfasst werden, welchen Nutzen die gegebenen Prädiktoren (in U1 = Schulnoten; in U2 = Auswahlverfahren), bei Schätzung der Grundquote, in Relation zu den entstehenden Kosten erbringen.

3.4 Versuchsplan

Zusammenfassend ergibt sich, wie in Tabelle 3.4-1 dargestellt, aus den drei beschriebenen Fragestellungen folgender Versuchsplan:

Tab. 3.4-1: Untersuchungsplan für die Fragestellungen 1 - 3

<table>
<tr><td colspan="3">FRAGESTELLUNG 1: ANALYSE DER PROGNOSTISCHEN VALIDITÄT VON SCHULNOTEN</td></tr>
<tr><td>Gesamtstichprobe</td><td>Prädiktoren</td><td>Kriterien</td></tr>
<tr><td>Teilnehmer am Auswahlverfahren für die Einstellung in den gehobenen Polizeivollzugsdienst (Einstellung Oktober 1999)</td><td>• Verschiedene Schulnoten
• Durchschnittsnote</td><td>• Gesamttestergebnis im Auswahlverfahren</td></tr>
<tr><td>Teilstichprobe</td><td>Prädiktoren</td><td>Kriterien</td></tr>
<tr><td>nach Ergebnis im Auswahlverfahren geeignete und eingestellte Auszubildende des gehobenen Polizeivollzugsdienstes</td><td>• Verschiedene Schulnoten
• Durchschnittsnote</td><td>• Gesamtnote in der Zwischenprüfung im Studium</td></tr>
<tr><td colspan="3">Berechnung der Kosten-Nutzen-Relation von Schulnoten anhand der Payoff-Funktion des Prädiktors Schulnoten hinsichtlich des Kriteriums Zwischennote</td></tr>
<tr><td colspan="3">FRAGESTELLUNG 2: ANALYSE DER PROGNOSTISCHEN VALIDITÄT DES AUSWAHLVERFAHRENS</td></tr>
<tr><td>Teilstichprobe</td><td>Prädiktoren</td><td>Kriterien</td></tr>
<tr><td>nach Ergebnis im Auswahlverfahren geeignete und eingestellte Auszubildende des gehobenen Polizeivollzugsdienstes</td><td>Punktzahl in den einzelnen Bausteinen des Auswahlverfahrens:
• Lückendiktat
• Bericht: Deutschleistung
• Bericht: Inhalt
• Kulturfairer Leistungstest
• Sport
• Nicht Kulturfairer Leistungstest
• Vorstellungsgespräch
• Rundgespräch
• Gesamtpunktzahl</td><td>• Gesamtpunkte der Zwischenprüfung im Studium</td></tr>
<tr><td colspan="3">Berechnung der Kosten-Nutzen-Relation von Schulnoten anhand der Payoff-Funktion des Prädiktors Ergebnis im Auswahlverfahren hinsichtlich des Kriteriums Zwischennote</td></tr>
<tr><td colspan="3">FRAGESTELLUNG 3: VERGLEICH DER ERGEBNISSE DEUTSCHER UND NICHT-DEUTSCHER TEILNEHMER</td></tr>
<tr><td>Teilstichprobe</td><td colspan="2">Fragestellung</td></tr>
<tr><td>Teilnehmer am kulturfairen und nicht kulturfairen kognitiven Leistungstest des Auswahlverfahrens für die Einstellung in den gehobenen Polizeivollzugsdienst (Einstellung Oktober 1999)</td><td colspan="2">Vergleich der Testergebnisse deutscher und nicht deutscher Bewerber in einem kulturfairen und einem nicht kulturfairen kognitiven Leistungstest</td></tr>
</table>

4. Fragestellung 1: Analyse der prognostischen Validität von Schulnoten

4.1 Einführung in die Fragestellung 1

Die sich aus dem Abschnitt 3 ergebenden Hypothesen für die Fragestellung 1 wurden in einer prospektiven Längsschnittstudie untersucht. Zu ihrer Beantwortung wurde die prognostische Validität der Schulnoten der Bewerber um einen Ausbildungsplatz für den gehobenen Polizeivollzugsdienst erfasst. Dabei wurden, wie im Abschnitt 3 dargestellt, zum einem einzelne Schulnoten, zum anderen die errechnete Durchschnittsnote der vorliegenden Zeugnisse herangezogen. Der Prädiktor Schulnoten bildeten die Noten des letzten aktuellen und durch die Probanden im Rahmen ihrer Bewerbung eingereichten Schulzeugnisses ab. Zur Erfassung des Erfolges im Auswahlverfahren (Kriterium) wurde das Gesamtergebnis der Teilnehmer im Auswahlverfahren verwendet. Im zweiten Schritt der ersten Untersuchung wurden ausschließlich die Daten der Bewerber herangezogen, die sich aufgrund ihrer erreichten Ergebnisse im Auswahlverfahren für die Einstellung in den gehobenen Polizeivollzugsdienst haben qualifizieren können und ein Jahr später das Studium an der Fachhochschule aufnahmen. Als weiteres Kriterium zur Erschließung der prognostischen Validität der Schulnoten wurde in diesem Fall die Gesamtnote der Zwischenprüfung im Studium erhoben.
Neben der Analyse zur prädiktiven Validität der Schulnoten sollte zusätzlich die Kosten-Nutzen-Relation des Prädiktors anhand der in Abschnitt 2.1 beschriebenen Payoff-Funktion des Brodgen-Cronbach-Gleser-Modells berechnet werden.

4.2 Beschreibung der erhobenen Prädiktoren

Wie in Kapitel 2.2 dargestellt, ist die grundlegende Voraussetzung für die Einstellung in den gehobenen Polizeivollzugsdienst der Nachweis der Fachhochschulreife oder des Abiturs als letzten erreichten Schulabschluss. Somit dienten die von den Bewerbern eingereichten Zeugnisse als Basis für die Erhebung potentiell geeigneter Schulnoten (Prädiktoren) und Bestimmung ihrer prognostischen Validität. Da ein größerer Teil der Probanden zum

Zeitpunkt ihrer Bewerbung noch zur Schule ging und dementsprechend über kein Abschlusszeugnis verfügte, wurde hier, gemäß den in Kapitel 2.2 beschriebenen Voraussetzungen für die Einstellung, das letzte aktuelle Zwischenzeugnis herangezogen. Allerdings zeigte sich in der Praxis, dass neben wenigen, in den meisten Lehrplänen vorgeschriebenen Grundlagenfächern, wie z.B. Deutsch und Mathematik, die Schüler der Oberstufen in vielen Schulen der Bundesrepublik Deutschland ihre Fächerwahl in gewissen Grenzen individuell gestalten können. Beispielhaft sei hier nur die Möglichkeit der Wahl von Leistungs- und Prüfkursen genannt. In der Folge fanden sich vielfältige Kombinationen von Fächern und somit Schulnoten in den vorliegenden Zeugnissen der Bewerber. Um eine größtmögliche Anzahl von Bewerbungen berücksichtigen zu können, war eine Standardisierung bei der Auswahl der Prädiktoren (Schulnoten) kaum leistbar und hätte zu einer sehr eingeschränkten Erfassung von Schulnoten geführt, die nur kleine Fallzahlen erlaubt und einen größeren Teil der Bewerber nicht hätte berücksichtigen können.

Aufgrund der oben beschriebenen Unterschiedlichkeit der Schulsysteme von Bundesland zu Bundesland, der differierenden Vorgaben für Pflichtfächer innerhalb einzelner Bundesländer an verschiedenen Schularten (bspw. Allgemeinbildendes Gymnasium, Waldorf Schule, Fachgymnasium, Kolleg) sowie der beschriebenen Möglichkeit zur teilweise individuellen Fächerwahl auf ein und derselben Schule wurden, um möglichst viele Noten in die Untersuchung aufnehmen zu können, diejenigen Fächer ausgewählt, welche im Großteil der Schulen Bestandteil des vorgeschriebenen Lehrplanes sind. Hierzu wurden die vorliegenden Zeugnisse gesichtet und diejenigen Fächer einbezogen, die insbesondere für die Analyse der prognostischen Validität von Schulnoten hinsichtlich des Kriteriums „Erfolg im Auswahlverfahren“ (zweiter Teil der Fragestellung 1) in den Extremgruppen in ausreichend großer Anzahl ($n \geq 30$) vorkamen. Aufgrund bestehender unterschiedlicher Angebote an den Schulen waren nicht alle angebotenen Fächer identisch benannt, erfassten jedoch vergleichbare Lehrinhalte. Nach Abstimmung mit erfahrenen Pädagogen der Ausbildungsinstitution wurden die Fächer Geschichte, Wirtschaft und Politik zusammen erhoben. In der Regel wurde in diesen Fächern nur eine Note im Zeugnis angegeben. Lagen im Ausnahmefall zwei Noten vor, so wurde, analog wie in der Praxis bei der Annahme von Bewerbungen, jenes Fach bewertet, in dem die bessere Note vorlag. Da nicht alle vorliegenden Zeugnisse der Bewerber über eine Durchschnittsnote verfügten, waren zur Berechnung dieser einige Vorüberlegungen notwendig. Eine gleichberechtigte Aufnahme aller Noten in ein Gesamtergebnis (Mittelwert) würde bedeuten, dass alle aufgenommenen

Schulnoten mit dem gleichen Gewichtungsbeitrag in die Durchschnittsnote eingehen würden. Wie aus der Literatur bekannt (siehe Kapitel 2), besitzen die einzelnen Noten jedoch unterschiedliche Gewichte und sind zudem interkorreliert. Aus diesem Grund wurde die Durchschnittsnote aufgrund des Ergebnisses einer einfachen linearen Regression berechnet. Wie der Darstellung im Anhang zu entnehmen, leisten alle erhobenen Noten einen die Varianz aufklärenden Beitrag und so wurde für die Analyse der prognostischen Validität die errechnete Durchschnittsnote als weiterer Prädiktor aufgenommen.
In Tabelle 4.2-1 sind die als Einzelprädiktoren erhobenen Schulnoten aufgeführt.

Tab. 4.2-1: Erfasste Schulnoten als Einzelprädiktoren

1	Deutsch	4	Geschichte, Wirtschaft, Politik
2	Mathematik	5	Sport
3	Englisch	6	Durchschnittsnote

4.3 Beschreibung der erhobenen Kriterien

4.3.1 Kriterien zur Erfassung des Erfolges im Auswahlverfahren

Der Erfolg bzw. Nicht-Erfolg eines Bewerbers im Auswahlverfahren für die Einstellung in den gehobenen Polizeivollzugsdienst lässt sich anhand der Ergebnisse in den einzelnen Testteilen sowie am erreichten Gesamttestergebnis ablesen. In der vorliegenden Untersuchung wurde aus Gründen der Vergleichbarkeit mit dem zweiten Teil der Untersuchung zur Bestimmung der prognostischen Validität von Schulnoten hinsichtlich des Kriteriums „Abschlussnote in der Zwischenprüfung“ entschieden, auch im ersten Teil als Kriterium das erreichte Gesamtergebnis im Auswahlverfahren (siehe Tabelle 4.3-1) zu wählen.

Tab. 4.3-1: Kriterium zur Bestimmung der prädiktiven Validität der Schulnoten hinsichtlich des Ergebnisses im Auswahlverfahren

Kriterium	Gesamtpunktzahl setzt sich zusammen aus den durch die Teilnehmer erreichten Einzelergebnissen (Punkte) in den Testbausteinen:
Gesamtpunktzahl im Auswahlverfahren	- Lückendiktat - Bericht Deutschleistung - Bericht Inhalt - Kulturfairer Leistungstest - Sporttest - Nicht kulturfairer Leistungstest - Vorstellungsgespräch - Rundgespräch

Im Auswahlverfahren wurden die Leistungen der Teilnehmer in den unterschiedlichen Testbausteinen anhand einer neunstufigen Skala bewertet. Ein Punktwert von eins bildete die schlechteste und ein Punktwert von neun die beste Bewertung ab. Die Transformation der einzelnen Bewertungen (Rohwerte) auf eine neunstufige Skala erfolgte in Abhängigkeit des jeweiligen Testverfahrens (siehe hierzu auch Kapitel 2.2.3.2). Zu beachten ist hierbei, dass das Auswahlverfahren grundsätzlich im Sinne einer sequentiellen pre-reject Auswahlstrategie durchgeführt wurde. Somit wurde der erste Test allen am Tag eingeladenen Probanden gemeinsam vorgegeben. Diejenigen Teilnehmer, welche nicht einen Mindestwert von vier Punkten erreichten, schieden aus dem weiteren Verfahren aus und wurden verabschiedet. Die verbliebenen Bewerber bekamen den nächsten Test zur Bearbeitung präsentiert. Diese Prozedur wiederholte sich für alle folgenden Komponenten des Auswahlverfahrens. Eine differenzierte Beschreibung einzelner Testbestandteile des Auswahlverfahrens findet sich unter Abschnitt 2.2.

4.3.2 Kriterien zur Erfassung des Studienerfolges

Für die Operationalisierung des Studienerfolges boten sich aufgrund der Studienstruktur grundsätzlich zwei Alternativen an: einerseits das Ergebnis in der Zwischenprüfung und andererseits das Ergebnis in der Laufbahnprüfung II. Aufgrund begrenzter zeitlicher Ressourcen, der Einschätzung der Praktikabilität der Untersuchung und des Erhebungsaufwandes für die weiteren Fragestellungen wurden die Ergebnisse der Zwischenprüfung im ersten Studienabschnitt als Operationalisierungen für den Studienerfolg herangezogen. Um den besonderen praktischen Aspekt der Berufstätigkeit Rechnung zu tragen, wurde darüber hinaus die Leistung der sich an die theoretische Zwischenprüfung und zu diesem gehörenden anschließenden Berufspraktikum in die Datenerhebung und Analyse aufgenommen. In der Tabelle 4.3-2 werden die Prüfungsfächer dargestellt, die die Grundlage für das Gesamtergebnis im zweiten Teil der ersten Untersuchung zur Bestimmung der prädiktiven Validität der Schulnoten bildeten. Außer in den Fächern „Einsatzlehre“ und „Kriminalistik“, bei denen zusätzlich Einzelleistungen (Hausarbeit oder Referat) mit in die Bewertung eingehen, wurden die Leistungen in Form von Klausuren bewertet (vgl. a. Kapitel 5.5.2.2).

Tab. 4.3-2: Kriterium zur Bestimmung der prädiktiven Validität der Schulnoten hinsichtlich des Studienerfolges

Kriterium	Gesamtnote (durch die Teilnehmer erreichte Einzelergebnissen in den Prüfungsfächern)
Gesamtnote im Auswahlverfahren	- Einsatzlehre - Kriminalistik - Allgemeines Verwaltungsrecht/ Polizeirecht - Strafrecht/ Ordnungswidrigkeitenrecht / Bürgerliches Recht - Kriminologie - Psychologie - Führungslehre - Staatsrecht - Berufspraktische Studienzeit

Die Bewertung der Leistungen der Studenten erfolgte anhand eines 15-Punkte-Bewertungssystems. Die erreichten Leistungen im Praktikum (berufspraktische Studienzeit) wurden dagegen in Form eines Beurteilungsbogens (Rating) durch Beamte der jeweiligen Dienststelle bewertet. Die Punktzahlen wurden dann in das in der Bundesrepublik Deutschland an den meisten Schulen und Fachhochschulen gebräuchliche Notensystem von Note 1 bis Note 6 transformiert (siehe Tabelle 4.3-3).

Tab. 4.3-3: Transformation der Punkte in den Prüfungsfächern der Zwischenprüfung in Noten

Punkte	Note
14-15	1 (sehr gut)
11-13	2 (gut)
8-10	3 (befriedigend)
5-7	4 (ausreichend)
2-4	5 (mangelhaft)
0-1	6 (ungenügend)

Aus datenschutzrechtlichen Gründen war es nur möglich, das Gesamtergebnis (Gesamtpunktzahl bzw. Gesamtnote) der Studenten in der Zwischenprüfung zu erheben. Für die Analyse der prognostischen Validität der Schulnoten wurde somit das Gesamtergebnis, welches sich aus der theoretischen sowie der praktischen Zwischenprüfung ergibt, als Kriteriumsvariable verwendet.

4.4 Versuchsplan für Fragestellung 1

Zur Beantwortung der Fragestellung 1 (Analyse der prognostischen Validität von Schulnoten) und damit der unter Kapitel 3.3 dargestellten Hypothesen, ergab sich, wie der Tabelle 4.4-1 zu

entnehmen, folgender Versuchsplan. Der Versuchsplan gliederte sich in drei Abschnitte: Den ersten Teil bildete die Erschließung der prognostischen Validität der einzelnen Schulnoten bzw. ihres errechneten Durchschnittwertes (Prädiktoren) hinsichtlich des erreichten Gesamtpunktergebnisses im Auswahlverfahren (Kriterium) ab. Im zweiten Teil wurde die prognostische Validität der einzelnen Schulnoten bzw. deren Durchschnittswert (Prädiktoren) hinsichtlich des Kriteriums „Gesamtnote in der Zwischenprüfung" im Studium untersucht. Im dritten Teil zur Analyse der prognostischen Validität von Schulnoten wurde mittels der Payoff-Funktion (Höft, 2001) in Anlehnung an Kersting (2004), die Kosten-Nutzen-Relation von Schulnoten berechnet.

Tab. 4.4-1: Versuchsplan für die Fragestellung 1

Fragestellung 1: Analyse der prognostischen Validität von Schulnoten		
Teil 1		
Gesamtstichprobe	Prädiktoren	Kriterium
Teilnehmer am Auswahlverfahren für die Einstellung in den gehobenen Polizeivollzugsdienst (Einstellung Oktober 1999)	a) Schulnote in: - Deutsch - Mathematik - Englisch - Geschichte /Wirtschaft / Politik - Sport b) Durchschnittsnote	Gesamtpunktzahl im Auswahlverfahren
Teil 2		
Teilstichprobe	Prädiktoren	Kriterium
Nach Ergebnis im Auswahlverfahren geeignete und eingestellte Auszubildende des gehobenen Polizeivollzugsdienstes (Einstellung Oktober 1999)	a) Schulnote in: - Deutsch - Mathematik - Englisch - Geschichte / Wirtschaft / Politik - Sport b) Durchschnittsnote	Gesamtnote der Zwischenprüfung im Studium
Teil 3		
Berechnung der Kosten-Nutzen-Relation von Schulnoten anhand der Payoff-Funktion		

4.5 Durchführung

Die Durchführung der Untersuchung fand während einer Hospitation im Zeitraum September 1998 bis April 2001 an der Landespolizeischule, Personal-Auswahl-Center, in Hamburg statt. Dabei war zu beachten, dass das Auswahlverfahren im Personal-Auswahl-Center sowie die Prüfung an der Fachhochschule durch einen festgelegten Ablauf determiniert waren (siehe auch Kapitel 2.2). Die Erhebung der Schulnoten fand ohne weitere Einschränkung der Population statt. Jede vorliegende Bewerbung wurde angenommen, d.h. es fand bei der

Annahme der Bewerbungen um einen Ausbildungsplatz keine Vorselektion hinsichtlich gegebener Vorauswahlbedingungen, z.B. einen geforderten Mindestdurchschnitt bei den Schulnoten, statt. Für die Erhebung der Daten wurden im ersten Schritt ab September – Oktober 1998 die Schulnoten (Prädiktoren) aus den Bewerbungsunterlagen aller Teilnehmer am Auswahlverfahren für die geplante Einstellung in den gehobenen Polizeivollzugsdienst im Oktober 1999 erfasst. Die erhobenen Prädiktoren dienten sowohl der Analyse ihrer prognostischen Validität hinsichtlich der Kriterien „Erfolg im Auswahlverfahren“ wie auch „Erfolg in der Zwischenprüfung“ für die Teilnehmer, welche erfolgreich an der Einstellungsprüfung teilnahmen und im Oktober 1999 in den gehobenen Polizeivollzugsdienst eingestellt wurden.

Für den zweiten Teil der Untersuchung wurden die Schulnoten der im Auswahlverfahren erfolgreichen und später in den gehobenen Polizeivollzugsdienst eingestellten Bewerber als Prädiktoren erhoben. Als Kriterium wurde, aus in Kapitel 3 dargestellten Gründen, das Gesamtergebnis der Studenten in der Zwischenprüfung der Ausbildung zum gehobenen Polizeivollzugsdienst an der Fachhochschule herangezogen.

Abschließend wurde im dritten Teil der Untersuchung zusätzlich die Kosten-Nutzen-Relation der Prädiktoren Schulnoten anhand der in Abschnitt 2.1 beschriebenen Payoff-Funktion des Brodgen-Cronbach-Gleser-Modells (zitiert nach Höft, 2001) in Anlehnung an Kersting (2004) berechnet.

4.5.1 Beschreibung der Stichprobe

4.5.1.1 Gesamtstichprobe zur Erhebung des Prädiktors „Schulnoten“ und des Kriteriums „Erfolg im Auswahlverfahren“

Am Auswahlverfahren für die Einstellung in den gehobenen Polizeivollzugsdienst im Oktober 1999 nahmen insgesamt 735 Personen teil. Anhand der eingereichten Bewerbungsunterlagen dieser Stichprobe wurden die Schulnoten des letzten aktuellen Zeugnisses entnommen. Wie aus den Tabellen 4.5-1 – 4.5-5 zu entnehmen, gab es 52,8% (388) männliche und 47,2% (347) weibliche Bewerber. Außerdem setzte sich die Bewerberstichprobe zu 94% (691) aus deutschen und 6% (44) nicht deutschen Bewerbern zusammen. Die Altersstruktur der Bewerber lag zwischen dem 17. und 41. Lebensjahr, dabei hatten 93 % der Bewerber das 30.

Lebensjahr noch nicht vollendet, insgesamt waren 83 % jünger als 25 Jahre. Der Altersmittelwert lag bei 20,86 Jahren. Bei den Bewerbern handelte es sich somit überwiegend um Schüler, die unmittelbar vor ihrem jeweiligen Schulabschluss standen und sich um einen Ausbildungsplatz bzw. Studienplatz bemühten. 51,6% davon besitzen die Fachhochschulreife und 32% die Hochschulreife, weitere 121 Teilnehmer (16,4%) hatten einen anderen, die Einstellungsvoraussetzungen erfüllenden, gleichwertigen Schulabschluss. Für die Einstellung in den Bereich der Schutzpolizei bewarben sich 510 (69,4%), für den Dienstzweig der Kriminalpolizei 171 (23,3%) und für den der Wasserschutzpolizei 54 (7,3%) Personen.
Die Verteilung der Herkunft nach einzelnen Bundesländern sah wie folgt aus: Der größte Anteil der Bewerber (171 Personen oder 23,3%) rekrutierte sich aus der Hansestadt Hamburg. 163 (22,2%) der Bewerber kamen aus Niedersachsen, 114 (15,5%) aus Schleswig – Holstein und 112 (15,2%) aus Mecklenburg-Vorpommern. Die anderen 23,8 % (175) setzten sich aus Bewerbern der übrigen Bundesländer zusammen. Nur aus dem Saarland lagen keine Bewerbungen vor.

Tab. 4.5-1: Zusammensetzung der Gesamtstichprobe: Geschlecht

Geschlecht n = 735	Häufigkeit	Prozent
männlich	388	52,8
weiblich	347	47,2

Tab. 4.5-2: Zusammensetzung der Gesamtstichprobe: Nationalität

Nationalität n = 735	Häufigkeit	Prozent
deutsch	691	94,0
nicht deutsch	44	6,0

Tab. 4.5-3: Zusammensetzung der Gesamtstichprobe: Schulbildung

Schulbildung n = 735	Häufigkeit	Prozent
Fachhochschulreife	379	51,6
Abitur	235	32,0
sonstige *	121	16,4

* sonstiger, als gleichwertig anerkannter Bildungsabschluss

Tab. 4.5-4: Zusammensetzung der Gesamtstichprobe: Alter der Bewerber

Alter n = 735	Häufigkeit	Prozent
17 – 20 Jahre	488	66,4
21 – 25 Jahre	134	18,2
26 – 30 Jahre	78	10,6
älter als 30 Jahre	35	4,8

Tab. 4.5-5: Zusammensetzung der Gesamtstichprobe: Bewerbung für Dienstzweig

Dienstzweig n = 735	Häufigkeit	Prozent
Schutzpolizei	510	69,4
Kriminalpolizei	171	23,3
Wasserschutzpolizei	54	7,3

4.5.1.2 Teilstichprobe zur Erhebung des Kriteriums „Erfolg im Studium“

Um die prädiktive Validität der Schulnoten hinsichtlich des Kriteriums Ausbildungserfolg bzw. Studienerfolg erschließen zu können, wurden für die Stichprobe jene Teilnehmer herangezogen, die am Auswahlverfahren für die Einstellung in den gehobenen Polizeivollzugsdienst erfolgreich teilgenommen und im Oktober 1999 die Ausbildung an der Fachhochschule begonnen hatten. Eingangs nahmen 134 Personen, die aufgrund ihres erfolgreichen Abschneidens im Auswahlverfahren eingestellt werden konnten, an dieser Erhebung teil. Allerdings schieden vor der Datenerhebung, 1 ½ Jahre nach Einstellung, sieben Probanden aus unterschiedlichen Gründen aus (u.a. nicht ausreichende Leistungen, Berufswechsel) und nahmen nicht an der Zwischenprüfung teil. Grundsätzlich wäre es wünschenswert gewesen, in diesen Fällen unterscheiden zu können, aus welchen Gründen die Eingestellten nicht an der Zwischenprüfung teilgenommen haben. Unter Umständen können nämlich jene darunter sein, die fälschlicherweise eingestellt wurden, obwohl sie nicht geeignet sind. Aufgrund der überlassenen Daten durch die Ausbildungsinstitution war dies jedoch nicht möglich.

Wie den Tabellen 4.5-6 – 4.5-10 zu entnehmen, verblieb somit eine Stichprobengröße von 127 Probanden. Es gab 55,1% (70) männliche und 44,9% (57) weibliche Studenten. 96,9% (123) waren deutscher und 3,1% (4) nicht deutscher Abstammung bzw. Nationalität. Die Altersstruktur der Bewerber lag zwischen dem 18. und 33. Lebensjahr, dabei hatten 97,0 % der Bewerber das 30. Lebensjahr noch nicht vollendet. Der Altersmittelwert lag bei 21,01 Jahren. 57 Teilnehmer (44,9%) verfügten über die Fachhochschulreife, 55 (43,3%) über die allgemeine Hochschulreife. Weitere 15 Teilnehmer (11,8%) hatten einen anderen, die Einstellungsvoraussetzungen erfüllenden, vergleichbaren Schulabschluss. Eingestellt in die Schutzpolizei waren 84 (66,1%) sowie in den Dienstzweig der Kriminalpolizei 31 (24,4%) und den der Wasserschutzpolizei 12 (9,4%) Probanden. Die Verteilung nach Herkunft aus den einzelnen Bundesländern sah wie folgt aus: 30 Personen (23,6%) kamen aus Hamburg, 24

(18,9%) aus Schleswig-Holstein, 29 (22,8%) aus Niedersachsen und 17 (13,4%) aus Mecklenburg-Vorpommern. Die verbliebenen Teilnehmer 27 (21,3%) setzten sich aus Bewerbern anderer Bundesländer (10) zusammen.

Tab. 4.5-6: Zusammensetzung der Teilstichprobe: Geschlecht

Geschlecht n = 127	Häufigkeit	Prozent
Männlich	70	55,1
Weiblich	57	44,9

Tab. 4.5-7: Zusammensetzung der Teilstichprobe: Nationalität

Nationalität n = 127	Häufigkeit	Prozent
Deutsch	123	96,9
nicht deutsch	4	3,1

Tab. 4.5-8: Zusammensetzung der Teilstichprobe: Schulbildung

Schulbildung n = 127	Häufigkeit	Prozent
Fachhochschulreife	57	44,9
Abitur	55	43,3
sonstige *	15	11,8

* sonstiger, als gleichwertig anerkannter Bildungsabschluss

Tab. 4.5-9: Zusammensetzung der Teilstichprobe: Alter der Eingestellten

Alter n = 127	Häufigkeit	Prozent
18 – 20 Jahre	76	59,7
21 – 25 Jahre	29	22,9
26 – 30 Jahre	20	15,7
älter als 30 Jahre	2	1,6

Tab. 4.5-10: Zusammensetzung der Teilstichprobe: Verteilung der Dienstzweige

Dienstzweig n = 127	Häufigkeit	Prozent
Schutzpolizei	84	66,1
Kriminalpolizei	31	24,4
Wasserschutzpolizei	12	9,4

4.5.2 Ablauf der Untersuchung

4.5.2.1 Erhebung des Prädiktors „Schulnoten“

Die Erhebung des Prädiktors Schulnoten wurde von September bis Oktober 1998 durchgeführt. Dazu wurden aus allen eingegangenen Bewerbungsunterlagen die im Abschnitt 4.2 aufgeführten Schulnoten des letzten aktuellen Schulzeugnisses erhoben. Wie dort beschrieben, wurden aufgrund der uneinheitlichen Schulsysteme, der unterschiedlichen Vorgaben für Pflichtfächer an verschiedenen Schularten sowie der Möglichkeit zur teilweise individuellen Fächerwahl, diejenigen Fächer ausgewählt, um möglichst viele Noten in die Untersuchung aufnehmen zu können, welche im Großteil der vorliegenden Zeugnisse am häufigsten vorkamen. Für die in den Bewerbungsunterlagen aufgeführten Schulnoten des jeweiligen letzten Abschlusszeugnisses konnten, gemäß dem in Kapitel 4.4 aufgeführten Versuchsplan, folgende Kennwerte (Tab. 4.5-11) für die Prädiktoren erhoben werden:

Tab. 4.5-11: Kennwerte für die erhobenen Schulnoten als Prädiktoren

Schulnoten n = 735	N	Mx	Sx
Durchschnittsnote*	689	2,81	.54
Deutsch	682	2,97	.76
Mathe	676	3,22	.98
Englisch	671	3,16	.84
Geschichte / Wirtschaft / Politik	677	2,71	.77
Sport	628	1,89	.75

*berechnete Durchschnittsnote aus den vorliegenden Fächern (siehe auch Kapitel 4.2)

Wie bereits beschrieben, resultieren die unterschiedlichen Fallzahlen in den oben aufgeführten Fächern u.a. auch daher, dass, abgesehen von einigen im Lehrplan vorgeschriebenen Grundlagenfächern, die Schüler in der Oberstufe in einem Großteil der Schulen der Bundesrepublik Deutschland ihre Fächer individuell wählen können. Somit liegen nicht von allen 735 Teilnehmern am Auswahlverfahren in jedem der einzelnen Fächer Daten vor. In der Zusammenstellung der Kennwerte wird weiter deutlich, dass von den insgesamt 689 Teilnehmern des Auswahlverfahrens die Noten in den zur Berechnung der Durchschnittsnote herangezogenen Fächern erhoben werden konnten. Von den restlichen 46 Probanden lagen keine Schulnoten vor. Die Daten wurden mittels des Statistikprogrammes SPSS, Version 10.0, eingegeben, zur Verrechnung vorbereitet sowie analysiert. Aus datenschutzrechtlichen

Gründen wurde dabei darauf geachtet, dass die Namen der einzelnen Personen durch Identifikationsnummern ersetzt wurden.

In der Tabelle 4.5-12 sind die Schulnoten noch einmal nach jenen Teilnehmern geordnet dargestellt, die das Gesamtverfahren erfolgreich haben bestehen können und jene, welche aufgrund unterschiedlicher Gründe (Tab. 4.5-18), aus dem Auswahlverfahren haben ausscheiden müssen.

Tab 4.5-12: Kennwerte für die Prädiktoren „Schulnoten“ im Auswahlverfahren erfolgreicher und nicht erfolgreicher Teilnehmer

N = 735	Im Auswahlverfahren erfolgreich N = 159			Im Auswahlverfahren ausgeschieden N = 576		
Schulnoten	N	Mx	Sx	N	Mx	Sx
Durchschnittsnote*	150	2,66	.56	539	2,85	.53
Deutsch	148	2,83	.69	534	3,01	.77
Mathe	147	3,05	1,08	529	3,27	.94
Englisch	147	2,99	.89	524	3,21	.82
Geschichte / Wirtschaft / Politik	149	2,67	.79	528	2,73	.76
Sport	49	2,58	.95	333	2,96	.76

*berechnete Durchschnittsnote aus den vorliegenden Fächern (siehe auch Kapitel 4.2)

Für den zweiten Teil der Untersuchung 1 wurden die Unterlagen der Bewerber, die erfolgreich am Auswahlverfahren teilnahmen und zum Oktober 1999 in den Polizeidienst eingestellt werden konnten, erfasst und bis zur Erhebung des Kriteriums aufbewahrt. Hier konnten insgesamt Daten von 127 Teilnehmern zur Analyse herangezogen werden. Zwar wurden im Oktober 1999 134 Teilnehmer am Bewerbungsverfahren eingestellt, letztendlich blieben für die Datenerhebung zum zweiten Teil der Untersuchung aufgrund unterschiedlicher Gründe (s.a. Kapitel 5.5.1) nur 127 Teilnehmer übrig. Für die in den aus den Bewerbungsunterlagen entnommenen Schulnoten des jeweiligen letzten Schulzeugnisses, die den später eingestellten Teilnehmern wieder zugeordnet wurden, lagen folgende Kennwerte (Tab. 4.5-13) als Prädiktoren vor:

Tab 4.5-13: Kennwerte für die Prädiktoren „Schulnoten“ bei den eingestellten Bewerbern

Schulnoten n = 127	N	Mx	Sx
Durchschnittsnote*	121	2,72	.52
Deutsch	119	2,87	.71
Mathe	119	3,13	1,03
Englisch	119	3,07	.89
Geschichte / Wirtschaft / Politik	121	2,73	.79
Sport	114	1,74	.75

*berechnete Durchschnittsnote aus den vorliegenden Fächern (siehe auch Kapitel 4.2)

4.5.2.2 Erhebung des Kriteriums „Erfolg im Auswahlverfahren“

Zur Erhebung des Kriteriums „Erfolg im Auswahlverfahren“ wurden im zweiten Schritt zur Beantwortung der Fragestellung 1 die Ergebnisse in den einzelnen Testbausteinen der zum Auswahlverfahren eingeladenen Teilnehmer, die sich für die Einstellung in den gehobenen Polizeivollzugsdienst im Oktober 1999 beworben hatten, erhoben. Die Durchführung fand im Zeitraum von Dezember 1998 – Mai 1999 statt. Die Bedingungen der Durchführung sind standardisiert, so dass von einer angemessenen Durchführungsobjektivität ausgegangen werden kann. Bei erfolgreichem Abschneiden in den Testverfahren des ersten Prüfungstages stellen sich die Bewerber am nächsten Tag dem Personalärztlichen Dienst zur Untersuchung der gesundheitlichen Tauglichkeit vor. Wenn diese den Bewerbern positiv bescheinigt wird, nehmen sie am dritten Prüfungstag teil. In Tabelle 4.5-14 wird schematisch der zeitliche Ablauf der Datenerhebung für die einzelnen Probanden dargestellt.

Die erreichten Leistungen der Teilnehmer in den unterschiedlichen Testverfahren wurden wieder überwiegend anhand einer neunstufigen Skala bewertet. Analog wie im ersten Teil bildete ein Punktwert von eins die schlechteste und ein Punktwert von neun die beste Bewertung ab. Die Transformation der einzelnen Bewertungen (Rohwerte) auf die neunstufige Skala erfolgte in Abhängigkeit des jeweiligen Testverfahrens. Eine Ausnahme gab es wieder bei den Bewertungen im Vorstellungs- und Rundgespräch. Hier wurde jeweils eine 15-Punkte-Skala beim Rating durch die Interviewer zu Grunde gelegt. Weiter war bei der Aufnahme der Ergebnisse die sequentielle pre-reject Auswahlstrategie zu beachten. Alle Teilnehmer, welche einen Mindestwert von vier Punkten nicht erreichten, schieden aus dem weiteren Verfahren aus.

Tab. 4.5-14: Ablauf des Auswahlverfahrens für den gehobenen Polizeivollzugsdienst

1. Prüfungstag	2. Prüfungstag	3. Prüfungstag
- Lückendiktat - Bericht Deutschleistung - Bericht Inhalt - Kulturfairer Leistungstest - Sporttest - Nicht kulturfairer Leistungstest	- Prüfung der gesundheitlichen Tauglichkeit durch den ärztlichen Dienst	- Vorstellungsgespräch - Rundgespräch

Zur Berechnung des Endergebnisses wurden die einzelnen Testteile gewichtet und anschließend aufaddiert. Die errechnete Summe bildete dann das Abschlussergebnis des Bewerbers. Die Gewichtung richtete sich weniger nach empirisch belegbaren Hintergründen

(bspw. der Höhe eines empirisch ermittelten prädiktiven Zusammenhanges zwischen Testteil und Berufserfolg), sondern nach inhaltlichen Überlegungen der durchführenden Organisation. In der Tabelle 4.5-15 findet sich eine Darstellung zur Gewichtung der einzelnen Testteile und in Tabelle 4.5-16 eine Übersicht der erhobenen Kriterien und deren Kennwerte für den ersten Teil zur Bestimmung der prädiktiven Validität von Schulnoten. Zur veranschaulichenden Darstellung ist in Tabelle 4.5-17 das erreichte Ergebnis eines anonymisierten Teilnehmers am Auswahlverfahren beispielhaft zusammengefasst.

Wie aus der Tabelle 4.5-16 zu entnehmen, nahmen von den 735 Teilnehmern des ersten Testbausteins nur 207 (28,2%) am letzten teil. In der Regel wurden die anderen Teilnehmer aufgrund der sequentiellen Auswahlstrategie aufgrund nicht ausreichender Leistungen vom weiteren Verfahren ausgeschlossen. Diese und weitere Ausscheidungsgründe der Teilnehmer sind zur Übersicht in Tabelle 4.5-18 aufgezeigt.

Tab. 4.5-15: Gewichtung der einzelnen Testbausteine im Auswahlverfahren

Testbausteine	Gewichtung
Lückendiktat	Ergebnis (Punkte 1 – 9) x 1
Bericht Deutschleistung Bericht Inhalt	Die Ergebnisse (Punkte 1 – 9) der beiden Berichtsleistungen werden addiert und fließen durch zwei dividiert in das Gesamtergebnis ein
Kulturfairer Leistungstest	(Punkte 1 – 9) x 2
Sporttest	Die Ergebnisse der 5 Untertests (Punkte 1 – 9) werden addiert und durch ihre Anzahl dividiert
Nicht kulturfairer Leistungstest	Die jeweiligen vier Untertests werden addiert (Punkte 1 – 9) und fließen gemeinsam in das Gesamtergebnis ein
Vorstellungsgespräch Rundgespräch	Die Ergebnisse der beiden Gesprächsteile (Punkte 1 – 15) werden addiert und fließen mit zwei multipliziert in das Gesamtergebnis ein
Gesamttestergebnis	Addition der gewichteten Ergebnisse

Tab. 4.5-16: Kennwerte der Testergebnisse im Auswahlverfahren

Testbausteine n = 735	N	Mx	Sx
Lückendiktat	735	5,4	1,64
Bericht Deutschleistung	734	5,24	1,69
Bericht Inhalt	734	5,9	2,19
Kulturfairer Leistungstest	721	7,12	1,62
Sporttest	662	4,71	1,43
Nicht kulturfairer Leistungstest	435	18,94	4,72
Vorstellungsgespräch	207	7,29	2,65
Rundgespräch	207	6,98	2,63
Gesamttestergebnis	735	48,49	25,64

Tab. 4.5-17: Veranschaulichende Darstellung eines erreichten Gesamtergebnisses im Auswahlverfahren (Kriterium)

Proband männlich, 23 Jahre		
Testbausteine	Gewichtung	Ergebnis
Lückendiktat	Ergebnis x 1	erreicht: = 8 Punkte
Bericht Deutschleistung Bericht Inhalt	Die Ergebnisse der beiden Berichtsleistungen werden addiert und fließen durch zwei dividiert in das Gesamtergebnis ein	erreicht: 7 und 8 Punkte / 2 = 7,5 Punkte
Kulturfairer Leistungstest	Ergebnis x 2	erreicht 8 Punkte * 2 =16 Punkte
Sporttest	Die Ergebnisse der 5 Untertests werden addiert und durch ihre Anzahl dividiert	erreicht: 5, 5, 5, 7 und 4 Punkte / 5 = 5,2 Punkte
Nicht kulturfairer Leistungstest	Die jeweiligen vier Untertests werden addiert und fließen gemeinsam in das Gesamtergebnis ein	erreicht: 7 + 8 + 7 + 9 Punkte = 31 Punkte
Vorstellungsgespräch * und Rundgespräch *	Die Ergebnisse der beiden Gespräche werden addiert und fließen mit zwei multipliziert in das Gesamtergebnis ein	erreicht (10 + 8 Punkte) * 2 = 36 Punkte
Gesamttestergebnis	Addition der gewichteten Ergebnisse	103,7 Punkte

* 15er Skala

Tab. 4.5-18: Ausscheidungsgründe im Auswahlverfahren

Ausscheidungsgründe n = 735	Häufigkeit	Prozent
Lückendiktat	38	5,2
Bericht (Deutschleistung und Inhalt)	30	4,1
Kulturfairer Leistungstest	9	1,2
Sporttest	192	26,1
Nicht kulturfairer Leistungstest	207	28,2
Vorstellungs- und Rundgespräch	36	4,9
Ärztlich nicht tauglich	33	4,5
Strafrechtliche Ermittlungen	4	0,5
Mit Angaben von Gründen verzichtet	19	2,6
Ohne Angaben von Gründen verzichtet	8	1,1
Keine Ausscheidungsgründe, d.h. im Gesamtverfahren erfolgreich	159	21,6
Gesamtteilnehmer	735	100

Auf der Grundlage ihres erreichten Ergebnisses im Auswahlverfahren waren insgesamt 207 Bewerber erfolgreich. Allerdings konnten nur 159 Teilnehmer, neben ihres erfolgreichen Abschneidens im Auswahlverfahrens, auch die sich anschließenden ärztlichen Untersuchungen sowie die polizeilichen Ermittlungen über das Vorliegen etwaiger Straftaten o.ä. bestehen und standen im Rahmen der Bestenauslese für die Einstellung im Oktober 1999 zur Verfügung. Nachdem die Kriteriumsdaten vorlagen, wurde zur Ermittlung der prognostischen Validität die Produkt-Moment-Korrelation zwischen einer Prädiktor- und einer Kriteriumsvariable berechnet (Lienert & Raatz, 1994). Weiter wurde mittels

regressionsanalytischer Untersuchungen die prognostische Validität der einzelnen Schulnoten und die der Durchschnittsnote des letzten aktuellen Schulzeugnisses erschlossen.

4.5.2.3 Erhebung des Kriteriums „Erfolg im Studium“

Um die mögliche Aussagekraft der Schulnoten als Prädiktor für den Erfolg in einem Auswahlverfahren bewerten zu können, sollte im zweiten Teil der Untersuchung 1 die Frage beantwortet werden, inwieweit Schulnoten hinsichtlich des Ausbildungserfolges eine hinreichend valide Vorhersage erlauben. Hierzu wurden von den Studenten der Ausbildung im gehobenen Polizeivollzugsdienst, die im Oktober 1999 eingestellt worden waren, die Ergebnisse ihrer Zwischenprüfung des Berufspraktikums im Frühjahr 2000 erfasst (s.a. Kapitel 4.5.2.3). Wie schon in Kapitel 4 beschrieben, hätte sich für die Operationalisierung des Studienerfolges auch das Ergebnis der abschließenden Laufbahnprüfung nach drei Jahren Ausbildung angeboten. Aufgrund begrenzter zeitlicher Ressourcen (vorgeschriebene Dauer der Hospitationszeit) sowie des zu erwartenden Zugewinns durch eine zeitliche Verlängerung, wurde entschieden, die Ergebnisse der Zwischenprüfung im ersten Studienabschnitt als Operationalisierungen für den Studienerfolg zu wählen. Um den praktischen Aspekten der Berufsausbildung und der späteren Berufstätigkeit Rechnung zu tragen, wurde die praktische Teilleistung der Zwischenprüfung, die Note der sich der theoretischen Prüfung anschließenden berufspraktischen Zeit, in das Gesamtergebnis integriert. Die Bewertung der Leistungen der Studenten erfolgte anhand eines 15-Punkte-Bewertungssystems (zur genauen Berechnung des Zwischenprüfungsergebnisses siehe auch Kapitel 5.2). Anschließend wurden die Ergebnisse zur Verrechnung mit den Schulnoten aufgrund des an der Ausbildungsinstitution gegebenen Transformationssystems in das in der Bundesrepublik Deutschland an den meisten Schulen gebräuchliche Notensystem von 1 bis Note 6 umgesetzt (s.a. Tab. 4.3-3).

Zur Ermittlung der prognostischen Validität wurde nach Beendigung der Zwischenprüfung und Vorliegen der Kriteriumsdaten erneut die Produkt-Moment-Korrelation zwischen einer Prädiktor- und einer Kriteriumsvariable berechnet (Lienert & Raatz, 1994). Weiter wurde auch mittels multipler Regressionsanalysen die Prädiktionskraft der Schulnoten hinsichtlich ihrer Vorhersageleistung bezüglich des Studienerfolges untersucht.

4.5.2.4 Berechnung der Kosten-Nutzen-Relation

Wie in den vorangegangenen Kapiteln bereits beschrieben und dargestellt, wurde neben der prädiktiven Validität der Schulnoten auch deren Nutzen und Kosten bezüglich des ereichten Ergebnisses im Auswahlverfahren anhand der Payoff-Funktion des Brodgen-Cronbach-Gleser-Modells (Höft, 2001) in Anlehnung an Kersting (2004) bestimmt. Das Zusammenspiel der Faktoren Grundquote, Selektionsquote und Prädiktorvalidität zeigt, dass der Einsatz eines validen Verfahrens umso wichtiger ist, je geringer der Anteil Geeigneter unter den unausgelesenen Bewerbern und je größer die Selektionsquote ist. Andererseits ist die Prädiktorvalidität weniger wichtig, wenn die Grundrate hoch und die Selektionsquote niedrig ist. Darüber hinaus ist es für ein Unternehmen wichtig einschätzen zu können, welchen monetären Nutzen ein Auswahlverfahren bietet (s.a. Kapitel 2.1). Die Berechnung der Payoff-Funktion des Brodgen-Cronbach-Gleser-Modells erfolgte, wie in Abbildung 4.5-1, nach folgender Formel:

$$\Delta U = NE \cdot T \cdot SDY \cdot r_{xy} \cdot \bar{z}_x - C \cdot NB$$

Abb. 4.5-1: Payoff-Funktion des Brodgen-Cronbach-Gleser-Modells (Höft, 2001)

Die inhaltliche Bedeutung der Parameter ist in der Tabelle 4.5-19 zusammengefasst.

Zur Berechnung der Payoff-Funktion wurden die verfügbaren Daten der Gesamtstichprobe und Teilstichprobe herangezogen sowie durch die erschlossenen Ergebnisse zur Bestimmung der prädiktiven Validität von Schulnoten ergänzt. Da nicht alle für die Formel notwendigen Daten vorlagen, wurden verschiedene Parameter aufgrund gegebener Erfahrungen geschätzt bzw. interpoliert.

Für die Berechnung der Payoff-Funktion wird der unkorrigierte Validitätskoeffizient in die Analyse aufgenommen. Bei den in der Eignungsdiagnostik verbreiteten Artefaktkorrekturen muss die Gefahr einer Überkorrektur berücksichtigt werden. Während eine Minderungskorrektur für Kriterienunreliabilität angemessen ist, darf eine Streuungseinschränkung in der Validierungsstichprobe, wie in dieser Untersuchung der Fall, nur so weit korrigiert werden, bis sie der Modellannahme einer durch Vorauswahl selegierten Bewerberstichprobe entspricht (Funke & Barthel, 1995). Aufgrund dessen wird

konservativerweise, um eine Überkorrektur zu verhindern, der unkorrigierte Validitätskoeffizient in der Kosten-Nutzen-Analyse verrechnet.

Tab. 4.5-19: Inhaltliche Bedeutung der Parameter der Payoff-Funktion des Brodgen-Cronbach-Gleser-Modells am Beispiel gehobener Polizeivollzugsdienst

Δ U	Der Nutzenzuwachs in Geldeinheiten, der durch das Testverfahren entsteht	
NE	Anzahl der Eingestellten	n = 134
T	Anzahl der berücksichtigten Zeiteinheiten (in Jahren), welche die Bewerber durchschnittlich in einer Institution verbringen	10 Jahre
SDY	Standardabweichung der Berufsleistung in Geldeinheiten 40 % des jährlichen Entgeltes x 2	Besoldungsstufen A 10 – A12 (Differenz zwischen den besten Mitarbeitern und durchschnittlichen Mitarbeitern) Durchschnittliches Jahresentgelt in Euro: A 10: 3052,69 * 12 = 36632,28 A 10: 3452,85 * 12 = 41434,20 A 10: 3843,33 * 12 = 46119,96 = (36632,28+41434,20+46119,96)/3 = 41395,48 41395,48 * 2 = 82,790,96 40 % = 33116,38
r xy	Validitätskoeffizient Zwischen der errechneten Durchschnittsnote aus den Fächern Deutsch, Mathematik, Englisch, Geschichte et al und Sport (Prädiktor) und dem Kriterium Ergebnis in der Zwischenprüfung	zu berechnen
$\bar{z}$ x	Durchschnittlicher standardisierter Testwert (Prädiktorwert) der Ausgewählten.	zu berechnen
C	Kosten für das Testverfahren pro Anwendung	Geschätzte Kosten in Euro pro Bewerber für: Mitarbeiter Annahme Bewerbungen Mitarbeiter Aufnahme Noten Mitarbeiter ärztliche Untersuchung Mitarbeiter polizeiliche Überprüfung Mitarbeiter Einstellung (siehe auch Kapitel 2.2.2)
NB	Anzahl der Bewerber	N = 735

4.6 Ergebnisdarstellung

Bei der Berechnung der prognostischen Validität der Noten hinsichtlich des Kriteriums „Endergebnis in der Zwischenprüfung“ muss berücksichtigt werden, dass es sich bei der Teilstichprobe, der nach dem Ergebnis im Auswahlverfahren geeigneten und eingestellten Bewerber des gehobenen Polizeivollzugsdienstes, um eine einseitig selegierte Stichprobe handelt. Da nur die Daten der Probanden, die im Auswahlverfahren erfolgreich waren, eingestellt und zur Zwischenprüfung zugelassen wurden, einbezogen werden konnten, fehlte der linke Auslauf der Testpunktwerteverteilung. Der statistische Zusammenhang gegenüber der Gesamtgruppe der Bewerber wird aufgrund dessen eher unterschätzt. Nach Lienert (1989) kann jedoch trotzdem mit einiger Vorsicht ein Repräsentativschluss auf die Validität des

Testes für die gesamte Bewerberpopulation vollzogen werden, indem eine Selektionskorrektur durchgeführt wird (vgl. Lienert, 1989, S. 307).
Die angegebenen Korrelationskoeffizienten sind somit für die Varianzeinschränkung korrigiert, ergänzend werden die unkorrigierten Werte jeweils darunter in Klammern mit der Signifikanzangabe ausgewiesen. Wenn die Korrelation auf dem Niveau 0,05 % zweiseitig signifikant wird, erfolgt die Angabe des Signifikanzniveaus mit „* “, bei einer Korrelation auf dem 0,01 % -Niveau erfolgt die Angabe „** “.

4.6.1 Prognostische Validität von Schulnoten hinsichtlich des Kriteriums Auswahlverfahren

Gemäß den unter Kapitel 3.3 dargestellten Hypothesen, ergaben sich zu deren Beantwortung zur Analyse der prognostischen Validität von Schulnoten (Kriterium Auswahlverfahren) die nachfolgenden Ergebnisse.
In Tabelle 4.6-1 sind die Einzelkorrelationen nach Pearson der erhobenen Schulnoten (Prädiktoren) mit dem Kriterium „Endergebnis im Auswahlverfahren“ (Gesamtpunkte) dargestellt.

Tab. 4.6-1: Einzelkorrelationen Schulnoten (Prädiktoren) mit dem Kriterium „Endergebnis im Auswahlverfahren“

Anzahl der Teilnehmer	Schulnoten	Endergebnis im Auswahlverfahren
689	Durchschnittsnote***	- .212**
682	Deutsch	- .124**
676	Mathe	- .158**
671	Englisch	- ,189**
677	Geschichte / Wirtschaft / Politik	- .037
628	Sport	- .130**

*berechnete Durchschnittsnote aus den vorliegenden Fächern (siehe auch Kapitel 4.2)

In Tabelle 4.6-2 – 4.6-4 sind die Ergebnisse der Regression (Einschluss) der Prädiktoren Schulnoten hinsichtlich des Kriteriums „Endergebnis im Auswahlverfahren“ (Gesamtpunkte) dargestellt.

Tab. 4.6-2: Regression: Prädiktor „Schulnoten“, Kriterium „Endergebnis im Auswahlverfahren“; Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	.251a	.063	.055	24,662

a. Einflussvariablen: (Konstante) Schulnoten Sport, Englisch, Mathematik, Geschichte / Wirtschaft / Politik, Deutsch

Tab. 4.6-3: Regression: Prädiktor „Schulnoten“, Kriterium „Endergebnis im Auswahlverfahren“; ANOVA b

Modell		Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
1	Regression	24110,494	5	4822,099	7,928	.000 a
	Residuen	358240,53	589	608,218		
	Gesamt	382351,03	594			

a. Einflussvariablen: (Konstante) Schulnoten Sport, Englisch, Mathematik, Geschichte / Wirtschaft / Politik, Deutsch
b. Abhängige Variable: Endergebnis Auswahlverfahren

Tab. 4.6-4: Regression: Prädiktor „Schulnoten“, Kriterium „Endergebnis im Auswahlverfahren“; Koeffizienten a

Modell		Nicht standardisierte Koeffizienten		standardisierte Koeffizienten	T	Signifikanz
		B	Standardfehler	Beta		
1	(Konstante)	75,063	5,383		13,945	.000
	Deutsch	-1,078	1,619	-.032	-.666	.506
	Mathematik	-2,528	1,114	-.098	-2,269	.024
	Englisch	-5,485	1,399	-.182	-3,920	.000
	Geschichte et al.	3,383	1,512	.105	2,238	.026
	Sport	-3,158	1,371	-.094	-2,304	.022

a. Abhängige Variable: Endergebnis Auswahlverfahren

4.6.2 Prognostische Validität von Schulnoten hinsichtlich des Kriteriums Zwischenprüfung im Studium

Gemäß den unter Kapitel 3.3 dargestellten Hypothesen, ergaben sich zu deren Beantwortung zur Analyse der prognostischen Validität von Schulnoten (Kriterium Endnote Zwischenprüfung) nachfolgende Ergebnisse.

In Tabelle 4.6-5 sind die Einzelkorrelationen nach Pearson der erhobenen Schulnoten (Prädiktoren) mit dem Kriterium „Endergebnis in der Zwischenprüfung“ des Studiums (Gesamtnote) dargestellt.

Tab. 4.6-5: Ergebnis Einzelkorrelation zwischen den Schulnoten (Prädiktoren) und dem Endergebnis Zwischenprüfung (Kriterium)

Anzahl der Teilnehmer	Schulnoten	Endergebnis Zwischenprüfung
121	Durchschnittsnote***	.297 (.288**)
119	Deutsch	.215 (.200**)
119	Mathe	.249 (.261**)
119	Englisch	.097 (.103)
121	Geschichte / Wirtschaft / Politik	.315 .(323**)
114	Sport	-.056 (- .052)

***berechnete Durchschnittsnote aus den vorliegenden Fächern (siehe auch Kapitel 4.2)

In den Tabellen 4.6-6 – 4.6-8 sind die Ergebnisse der Regression (Einschluss) der Prädiktoren „Schulnoten“ hinsichtlich des Kriteriums „Endergebnis in der Zwischenprüfung“ (Gesamtnote) dargestellt.

Tab. 4.6-6: Regression: Prädiktor „Schulnoten“, Kriterium „Endergebnis in der Zwischenprüfung“; Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	.406a	.165	.125	.4592

a. Einflussvariablen: (Konstante) Schulnoten Sport, Englisch, Mathematik, Geschichte / Wirtschaft / Politik, Deutsch

Tab. 4.6-7: Regression: Prädiktor „Schulnoten“, Kriterium „Endergebnis in der Zwischenprüfung“; ANOVA b

Modell		Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
1	Regression	4,338	5	.868	4,114	.002 a
	Residuen	21,931	104	.211		
	Gesamt	26,269	109			

a. Einflussvariablen: (Konstante) Schulnoten Sport, Englisch, Mathematik, Geschichte / Wirtschaft / Politik, Deutsch
b. Abhängige Variable: Endergebnis Zwischenprüfung

Tab. 4.6-8: Regression: Prädiktor „Schulnoten“, Kriterium „Endergebnis in der Zwischenprüfung“; Koeffizienten a

Modell	Nicht standardisierte Koeffizienten		standardisierte Koeffizienten	T	Signifikanz
	B	Standardfehler	Beta		
1 (Konstante)	2,245	.231		9,730	.000
Deutsch	9,973E-02	.074	.144	1.340	.183
Mathematik	8,201E-02	.046	.170	1,780	.078
Englisch	-2,467E-02	.056	-.045	-.442	.659
Geschichte et al.	.163	.061	.269	2,689	.008
Sport	-.102	.061	-.157	-1,669	.098

a. Abhängige Variable: Endergebnis Zwischenprüfung

4.6.3 Kosten-Nutzen Relation

Wie unter Kapitel 4.5.2.4 erläutert, erfolgte die Berechnung der Payoff-Funktion des Brodgen-Cronbach-Gleser-Modells für die Note mit der höchsten Prädiktionskraft nach der Formel: $\Delta U = NE \cdot T \cdot SDY \cdot rxy \cdot zx - C \cdot NB$. Die Darstellung und das Ergebnis der Berechung sind in Tabelle 4.6-9 wiedergegeben.

Die Anwendung der Schulnoten als Auswahlinstrument ergibt im Vergleich zu einer bloßen Zufallsauswahl (rxy= 0) für die nächsten ca. 10 Jahre einen Nutzenzuwachs von knapp 3,1 Millionen Euro.

Tab. 4.6-9: Berechnung der Payoff-Funktion von Schulnoten

Δ U	Der Nutzenzuwachs in Geldeinheiten, der durch das Testverfahren entsteht:	
NE	Anzahl der Eingestellten	134
T	Anzahl der berücksichtigten Zeiteinheiten (in Jahren), welche die Bewerber durchschnittlich in einer Institution verbringen	10 Jahre
SDY	Standardabweichung der Berufsleistung in Geldeinheiten 40 % des jährlichen Entgeltes x 2	33116,38 Euro
r xy	Validitätskoeffizient der Durchschnittsnote	r = .288
zx	Durchschnittlicher standardisierter Testwert (Prädiktorwert)	.27
C	Kosten für das Testverfahren pro Anwendung	500 Euro
NB	Anzahl der Bewerber	735
Nutzenzuwachs in Geldeinheiten:		
= (134* 10* 33116,38 * .288 * .27) – (500 * 735) = 3.083.173,81 Euro		

4.6.4 Beantwortung der Hypothesen

Die in Kapitel 3.3.1.1 formulierten Hypothesen zur Analyse der prognostischen Validität der einzelnen Schulnoten und der errechneten Durchschnittsnote bezüglich des Kriteriums „Gesamtergebnis Auswahlverfahren“ können wie folgt beantwortet werden, wobei die korrigierten Korrelationskoeffizienten ausgewiesen werden:

Wie aus der Tabelle 4.6-1 ersichtlich, ergaben sich zwischen den Schulnoten Deutsch (r = -.124), Mathematik (r = -.158), Englisch (r = -.189) und Sport (r = -.130) auf dem 1% Niveau empirische signifikante nachweisbare Zusammenhänge. Auch die errechnete Durchschnittsnote korrelierte auf dem 1% Niveau signifikant mit dem Kriterium (r = -.212).
Wie aus den Tabellen 4.6-2 und 4.6-3 zu entnehmen, ergibt sich ein korrigiertes R2 von .055. Die Regressionsgleichung ist mit F = 7,928 signifikant. Dabei leisten die Prädiktoren Mathematik, Englisch und Geschichte et al. einen signifikanten Beitrag zur Regressionsgleichung (Tabelle 4.6-4).

Die in Kapitel 3.3.1.2 formulierten Hypothesen zur Analyse der prognostischen Validität der einzelnen Schulnoten und der errechneten Durchschnittsnote bezüglich des Kriteriums „Gesamtnote Zwischenprüfung“ können wie folgt beantwortet werden:

Wie aus der Tabelle 4.6-5 ersichtlich, ergaben sich zwischen den Schulnoten Deutsch (r = .215), Mathematik (r = .249), Geschichte et al. (r = .315) und der Durchschnittsnote (r =

.297) auf dem 1% Niveau empirische signifikante nachweisbare Zusammenhänge. Wie aus den Tabellen 4.6-6 und 4.6-7 zu entnehmen, ergibt sich ein korrigiertes R2 von .125. Die Regressionsgleichung ist mit F = 4,144 signifikant. Dabei leistet nur der Prädiktor Geschichte et al. einen signifikanten Beitrag zur Regressionsgleichung (Tabelle 4.6-8).

4.7 Interpretation der Ergebnisse

Ziel der Untersuchung 1 war die Analyse der prognostischen Validität der Schulnoten von Bewerbern um einen Ausbildungsplatz für den gehobenen Polizeivollzugsdienst. Hierzu wurden in einem ersten Schritt aus den eingegangenen Bewerbungsunterlagen die Schulnoten Deutsch, Englisch, Mathematik, Sport und die vorliegende Note in einem der Fächer Geschichte, Politik oder Wirtschaft als Prädiktoren erhoben. Bei der Erhebung der Prädiktoren wurden jene Fächer einbezogen, die in ausreichend großer Anzahl ($n \geq 30$) vorkamen. Nach Abstimmung mit erfahrenen Pädagogen der Ausbildungsinstitution wurden einzelne Fächer zusammen erhoben. Da nicht alle vorliegenden Zeugnisse der Bewerber über eine Durchschnittsnote verfügen, waren zur Berechnung dieser einige Vorarbeiten notwendig. Eine gleichberechtigte Aufnahme aller Noten in ein Gesamtergebnis (Mittelwert) würde bedeuten, dass alle aufgenommenen Schulnoten mit dem gleichen Gewichtungsbeitrag in die Durchschnittsnote eingehen. Wie aus der Literatur bekannt (siehe Kapitel 2), besitzen die einzelnen Noten jedoch unterschiedliche Gewichte und sind zudem interkorreliert. Aus diesem Grund wurde die Durchschnittsnote aufgrund des Ergebnisses einer einfachen linearen Regression berechnet und ebenfalls als Prädiktor zur Analyse der prognostischen Validität verwendet. Als Kriterium wurde das Gesamtergebnis der Teilnehmer im Auswahlverfahren verwendet.

Der zweite Schritt der Untersuchung umfasst die Analyse der Daten derjenigen Bewerber, die sich aufgrund ihrer erreichten Ergebnisse im Auswahlverfahren für die Einstellung in den gehobenen Polizeivollzugsdienst haben qualifizieren können und ein Jahr später das Studium an der Fachhochschule aufnahmen. Ziel war es, die Vorhersageleistung von Schulnoten auch hinsichtlich eines weiteren Kriteriums zu erschließen. So wurde hier die Gesamtnote der Zwischenprüfung im Studium als Kriterium erhoben. In diesem Teil der Untersuchung wurden die erhobenen Noten aus o.g. Fächern um jene bereinigt, die von Teilnehmern am Testverfahren stammten, die sich nicht für eine Einstellung haben qualifizieren können bzw. aus anderen Gründen aus dem Testverfahren ausschieden.

Bei der Erstellung eines Auswahlverfahrens und im Weiteren der Überprüfung seiner Güte ist es neben der Maximierung der prognostischen Validität ebenso das Ziel, diejenigen Testverfahren und Instrumente mit den geringsten Kosten und dem größtmöglichen Nutzen für eine Organisation auszuwählen. Somit stellte sich die Frage, ob Schulnoten als Prädiktoren des beruflichen Erfolges bzw. des Ausbildungserfolges mit einem relativ geringen zeitlichen Erhebungsaufwand und geringen Kosten andere teurere und zeitlich aufwendigere Prädiktoren ersetzen können. In zahlreichen Untersuchungen zur Bedeutung des Prädiktors Schulnoten wird diese Frage eher kontrovers diskutiert (vgl. a. Althoff, 1986; Ingenkamp, 1969; Petersen, 2002, Schuler, 1998). Falls jedoch Schulnoten eine ausreichend hohe prognostische Validität zur Vorhersage des Gesamtergebnisses im Auswahlverfahren oder / und des Erfolges in der Zwischenprüfung aufweisen, wäre unter Berücksichtigung der Kosten-Nutzen Relation zu überlegen, ob das in dieser Form bestehende Auswahlverfahren für den gehobenen Polizeivollzugsdienst noch in gegebener Form sinnvoll ist bzw. es nicht zumindest mit modifizierter Struktur durchgeführt werden müsste. Aus diesem Grund wurde die Kosten-Nutzen Relation anhand der in Abschnitt 2.1 beschriebenen Payoff-Funktion des Brodgen-Cronbach-Gleser-Modells berechnet.

Nach Schuler (1996) liegt die Höhe eines durchschnittlichen Korrelationskoeffizienten in der Berufseignungsdiagnostik in einem Bereich von r = .30. Eine Prognose des beruflichen Erfolges mit einem Validitätskoeffizienten von r = .50 ist bereits als sehr gut zu bezeichnen, (Kapitel 2.3), jedoch nur selten bei eingesetzten eignungsdiagnostischen Prädiktoren zu finden. Gemäß der allgemeinen Richtlinien an die erforderliche Höhe von Validitätskoeffizienten von Lienert (1989) ist deshalb ein Test in einem solchen Umfang valide, wenn seine Anwendung eine bessere Vorhersage ermöglicht als seine Unterlassung. Demnach wäre bereits ein Validitätskoeffizient zufrieden stellend, der sich auf dem 1%-Niveau absichern lässt. Tests mit Validitätskoeffizienten unter r = .3 sind allerdings nach Lienert (1989), auch wenn sie statistisch abgesichert sind, für die alleinige Verwendung nahezu nutzlos. Sie können allerdings im Rahmen einer Testbatterie gute Dienste leisten, insbesondere dann, wenn sie mit den anderen Tests niedrig korrelieren. Unter diesen Gesichtspunkten sind die errechneten bivariaten Korrelationskoeffizienten der Schulnoten Deutsch (r = -.12), Mathematik (r = -.16), Englisch (r = -.19), Sport (r = -.04) und der Durchschnittsnote (r = -.21) im Hinblick auf die Vorhersage des Gesamtergebnisses im Auswahlverfahren zwar signifikant, die Zusammenhänge jedoch nur sehr gering. Im Vergleich dazu, wurde in einer Studie von Althoff (1986) hinsichtlich der Beziehungen

zwischen Schulnoten und dem Kriterium Testleistung, nur ein geringer Teil der 392 Korrelationen, nämlich genau 14, überhaupt signifikant.

Validitätskoeffizienten anderer Studien, wie z.B. Baron - Boldt, Funke und Schuler (1989) zeigen allerdings mit einer prädiktiven Validität der Schulnoten von r = .41 für den Ausbildungserfolg und r = .46 für den Studienerfolg deutlich höhere Validitätskoeffizienten. Insofern erscheinen die vorliegenden prognostischen Validitäten der genannten Noten unzureichend. Auch wenn es im ersten Schritt der Untersuchung 1 nicht um die Vorhersage des Ausbildungserfolges ging, haben Schulnoten als alleiniges Auswahlinstrument für den gehobenen Polizeivollzugsdienst in Anlehnung an die oben genannte Aussage von Lienert (1989) aufgrund dieser Ergebnisse kaum einen praktischen Nutzen.

Von Bedeutung können sie allerdings im Rahmen einer anderen Funktion sein. Wie in Kapitel 2.3 dargestellt, muss die prognostische Validität der Bewerbungsunterlagen insgesamt mit r = .18 (Reilly & Chao, 1982) als ungenügend erachtet werden. Hier leistet zumindest die errechnete Durchschnittsnote mit einer prognostischen Validität von r = -.21 einen Beitrag zur Vorhersage des Endergebnisses im Auswahlverfahren und könnte als Vorauswahlkriterium einen validen Beitrag für die Aussagekraft der Bewerbungsunterlagen liefern. Ähnliches gilt unter Umständen für die Englischnote mit einem Validitätskoeffizienten von r = -.19. Die weiteren Schulnoten liegen im Vergleich dazu mit Korrelationen in einem Bereich von r = -.13 (Sport) und niedriger und bleiben ohne praktischen Nutzen. Gerade auch unter Berücksichtigung der in den letzten Jahrzehnten deutlich geringer werdenden mittleren Korrelation für die Noten von bspw. Realschulzeugnissen mit einer prognostischen Validität von r = .26 (Schuler, 1998), spricht die prognostische Validität der errechneten Durchschnittsnote zumindest für eine gewisse Nützlichkeit der Verwendung als Einzelprädiktor im Rahmen der Vorauswahl. Schuler (1998) verweist darüber hinaus auf eine abnehmende Tendenz auch im Bereich der allgemeinen Hochschulreife, wie sie in der Regel die Zugangsvoraussetzung für die Aufnahme eines Studiums im gehobenen Polizeivollzugsdienst ist.

Trotz diese durchaus nützlichen Funktion im Rahmen der Vorauswahl stellen sich die erhobenen Schulnoten, orientiert an den von Lienert (1989) gestellten Anforderungen für die Höhe von Validitätskoeffizienten, als alleiniger Prädiktor zur Auswahl von Polizeivollzugsbeamten als wenig nützlich dar. Im gleichen Sinne sind sie aber im Rahmen

einer Testbatterie einsetzbar und haben als Vorauswahlfilter die Funktion einer Vorselektion und bilden durchaus eine nützliche erste Stufe im Auswahlprozess.

Die regressionsanalytischen Berechnungen zeigen ein ähnliches Bild: Trotz eines signifikanten F-Bruches (overall) liegt die multiple Korrelation mit R = .251 und mit einem Determinationskoeffizienten von .063 erstaunlich niedrig, d.h. nur ca. 6%, der Varianz des Kriteriums werden durch die Schulnoten erklärt. Dieser Wert an aufgeklärter Varianz ist derart gering, dass andere Einflüsse für das erfolgreiche Abschneiden im Auswahlverfahren eine ebenso große oder noch größere Rolle zu spielen scheinen als die Schulnoten.
Die Englischnote, welche nach der Durchschnittsnote die höchste bivariate Korrelation mit dem Kriterium Endergebnis im Auswahlverfahren besitzt, leistet auch regressionsanalytisch den höchsten Beitrag zur Vorhersage des Kriteriums. Sprachliche Fertigkeiten, wie z.B. sprachliche Ausdrucksfähigkeit und sprachliches Verständnis, wie sie im Fach Englisch erforderlich sind, scheinen für das erfolgreiche Abschneiden im Auswahlverfahren behilflich zu sein. Viele der Testbausteine im Auswahlverfahren (vgl. Kapitel 2.2) weisen sprachliche Komponenten auf, bei denen die genannten sprachlichen Fähigkeiten Voraussetzung für das erfolgreiche Abschneiden sind. Zu nennen wären hier u.a. das Lückendiktat, das Verfassen eines Berichtes sowie die sprachliche Darstellung im Vorstellungs- und Rundgespräch. Das hier sich bestätigende Ergebnis spricht neben der Durchschnittsnote für die zusätzliche Verwendung der Englischnote als Vorauswahlkriterium.

Analog der bivariaten Korrelation, stellt die Mathematiknote den Prädiktor mit dem dritthöchsten ß-Gewicht dar. Allerdings konnte hier nicht das Ergebnis anderer Studien bestätigt werden, nach denen sich die Mathematiknote als validester Einzelprädiktor im Kontext der Vorhersage des Ausbildungserfolges erwiesen hatte (z.B. Althoff, 1986). Sie erreicht hier eine ähnlich geringe Aussagekraft (r = .15), wie dies für die Prognose des beruflichen Erfolges bekannt ist (Schuler, 1996).
Bei Betrachtung der Ergebnisse der regressionsanalytischen Berechnungen zeigt sich, dass die ß-Gewichte aller Schulnoten, außer dem des Prädiktors „Schulnote Deutsch“, signifikant werden. Das unbedeutende Gewicht der „Deutschnote“ ist insofern überraschend, als die bivariate Korrelation dieses Prädiktors mit dem Kriterium „Endergebnis im Auswahlverfahren“ auf dem 1 %- Niveau mit r = -.12 signifikant wurde. Die Erklärung für diesen Widerspruch findet sich allerdings in den Korrelationen der Schulnoten Mathematik, Englisch, Geschichte et al. und Sport. Dort zeigen sich durchweg signifikante

Interkorrelationen, d.h. in der Schulnote „Deutsch“ stecken Anteile der anderen beschriebenen Noten und vice versa. Diejenigen Eigenschaften und Fähigkeitsanteile, welche die Leistung in dem Fach „Deutsch“ bestimmen, sind ebenso für die Leistung in den anderen genannten Fächern erforderlich. In diesen Kontext passt ebenso das Ergebnis des Prädiktors „Schulnote in Geschichte et al“. Die bivariate Korrelation dieses Prädiktors mit dem Kriterium ist sehr gering und nicht signifikant, während sein Gewicht in der Regressionsanalyse den zweithöchsten Beitrag für die Vorhersage des Kriteriums Endergebnis liefert.

Für die Vorhersage des Kriteriums „Gesamtnote in der Zwischenprüfung“ ergeben sich signifikante bivariate korrigierte Korrelationen der Noten Deutsch (r = .22), Mathematik (r = .25) und Geschichte et al. (r = .32), die im Vergleich zu den Korrelationskoeffizienten für die Vorhersage des Endergebnisses im Auswahlverfahren zufrieden stellend höher liegen. Im Vergleich aller Prädiktoren weist die Schulnote in Geschichte et al. sogar mit r = .32 den höchsten Validitätskoeffizienten auf. Bezüglich der von Lienert (1989) gestellten Anforderungen an die Höhe der Validitätskoeffizienten ist der Zusammenhang zwischen der Note in Geschichte und der Gesamtnote in der Zwischenprüfung eher gering, lässt jedoch tendenziell einen Einfluss vermuten. Wie schon dargestellt, können nach Lienert (1989) Prädiktoren mit Validitätskoeffizienten im Bereich von r = .3 und höher von praktischer Bedeutung sein. Insbesondere als Ergänzung der oben beschriebenen Noten könnte die Geschichtsnote auch im Rahmen der Vorauswahl von Bewerbern für den gehobenen Polizeivollzugsdienst eine nützliche und praktische Verwendung finden. Das gleiche Bild zeigt sich in der Analyse der multiplen linearen Regression. Neben dem signifikanten F-Bruch zur Überprüfung der Hypothese, dass der multiple Korrelationskoeffizient größer ist als Null, besitzt der Prädiktor „Schulnote in Geschichte et al.“ einen bedeutsamen Einfluss auf das Endergebnis in der Zwischenprüfung (höchstes und einzig signifikantes ß-Gewicht). Dieses Ergebnis stützt die Verwendung dieser Note im Kontext der Vorauswahl der Bewerber für den gehobenen Polizeivollzugsdienst.

Auch die Durchschnittsnote weist in Relation zu den anderen Noten einen vergleichsweise hohen Validitätskoeffizient von r = .3 auf, bleibt jedoch genauso unter der in praxi geforderten Höhe für Validitätskoeffizienten von r = .5 (Schuler, 1996). Auch wenn die Durchschnittsnote in dieser Untersuchung eine signifikante Korrelation mit der Gesamtnote in der Zwischenprüfung aufweist und sie generell auch in unterschiedlichen anderen Studien (vgl. z.B. Baron-Boldt, 1989) als der beste Einzelprädiktor für die Vorhersage des Studien-

und Berufserfolges gilt, reicht dieses Ergebnis nicht an andere bereits in diesem Zusammenhang gefundene Höhen von Validitätskoeffizienten heran. So ermittelte Baron-Boldt (1989) bspw. eine mittlere korrigierte prognostischen Validität von r = .46 bzw. r = .41.

In Relation zu prädiktiven Validitäten der bisherigen Prädiktoren zur Vorhersage des Ausbildungserfolges (insbesondere die anzunehmende prädiktive Validität der kognitiven Leistungstests) und aufgrund der nur geringen Verbreitung der Geschichtsnote in den vorliegenden Zeugnissen der Bewerber, bieten sich beide Noten trotz dieser zumindest annähernd zufrieden stellenden Validitäten nicht als Ersatz- bzw. zusätzlicher Baustein im Auswahlverfahren an.

Die Noten in den Fächern Deutsch, Englisch, Mathematik und Sport haben aufgrund ihrer nicht signifikanten ß-Gewichte keinen signifikanten Einfluss auf das Endergebnis in der Zwischenprüfung. Die bivariaten Korrelationen der Fächer Deutsch und Mathematik zeigen jedoch signifikante Korrelationen mit dem Kriterium. Die bereits beschriebene Mulitkollinearität und deren Konsequenz in Suppressoreffekten zeigen, dass die einzelnen Fähigkeiten, die sich in den unterschiedlichen Noten manifestierten, sich nicht unbedingt immer diskriminieren lassen.

Für den Einsatz in der Praxis spielt somit nicht nur der Validitätsnachweis eine Rolle, sondern es müssen ebenso die Kosten der Verwendung des Verfahrens und dessen Nutzen berücksichtigt werden. So ergibt sich bei Berücksichtigung der Ergebnisse der Kosten-Nutzen Analyse im Vergleich zu einer bloßen Zufallsauswahl (rxy = 0) ein anderes Bild. Für die Durchschnittsnote als Einzelprädiktor für die nächsten 10 Jahre ergäbe sich ein Nutzenzuwachs von fast 3,1 Millionen Euro. Im Rahmen der Vorauswahl von Bewerbern für den gehobenen Polizeivollzugsdienst müssen die Schulnoten lediglich den Bewerbungsunterlagen als Daten entnommen werden. Dies bedeutet im Prinzip weder zeitlich noch hinsichtlich der Kosten einen großen zusätzlichen Erhebungsaufwand. Trotz der in dieser Untersuchung eher geringen Validitätskoeffizienten besitzen die Noten zumindest gegenüber einer Zufallsauswahl einen vergleichsweise hohen Nutzen. Selbst bei der Kosten-Nutzen Analyse der Mathematiknote mit einem eher niedrigen Validitätskoeffizient von r = .22 ergibt sich ein Nutzenzuwachs von ca. 2,4 Millionen Euro (T = 10 Jahre). In Anlehnung an das von Kersting (2004) verwendete Beispiel (Nettonutzen = 300.500 Euro) und dessen Parameter ergibt sich zur Beurteilung der Aussagekraft der vorliegenden

Ergebnisse für die Durchschnittsnote (zwei eingestellte und 14 eingeladene Bewerber) bei einer angenommenen Verweildauer von zehn Jahren, ein zusätzlicher Nettonutzen, im Vergleich zu einer Zufallsauswahl, von 46.112,05 Euro.

Die berechneten Validitätskoeffizienten für die Schulnoten sind weder hinsichtlich des Kriteriums „Gesamtpunkte im Auswahlverfahren“ noch des Kriteriums „Abschlussnote in der Zwischenprüfung“ zufrieden stellend. Mit einem vergleichsweise hohen Validitätskoeffizienten von r = .30 bzw. r = .32 erlauben hinsichtlich des Kriteriums „Abschlussnote in der Zwischenprüfung“ nur die Durchschnittsnote und die Note in dem Fach Geschichte et al. eine zufrieden stellend hohe Vorhersage des Kriteriums und besitzen einen praktischen Wert.

Insgesamt tragen die in dieser Untersuchung gefundenen Validitätskoeffizienten der Noten nicht zu einem einheitlicheren Bild bei. So konnte bspw. die hohe prognostische Validität der Mathematiknote nicht belegt werden. In Anbetracht der fehlenden Güte, sind auch Kosten und Nutzen Überlegungen in Relation zur Zufallsauswahl kein überzeugendes Argument für die Verwendung von Schulnoten als alleinige Prädiktoren zur Vorhersage des Ergebnisses im Auswahlverfahren bzw. des Studienerfolges.
Zudem geht eine Vorhersage des Studienerfolges auf Grundlage dieser Resultate, trotz der regressionsanalytischen signifikanten Ergebnisse, mit einem bestimmten Fehleranteil einher. Die durch die Prädiktoren Schulnoten aufgeklärte Varianz des Kriteriums „Endergebnis in der Zwischenprüfung“ ergibt nur ein von R2 = .125, die restlichen fast 80% an Kriteriumsvarianz werden durch andere Faktoren beeinflusst. Neben der starken Multikollinearität der Prädiktoren beeinflusst auch die kleine Stichprobe diese eher niedrigen Ergebnisse.

Die meisten psychologischen Variablen, die in empirischen Studien erhoben werden, sind in der Regel interkorreliert. Mit steigender Korrelation unter den Prädiktoren nimmt die Genauigkeit der Parametereinschätzung jedoch ab (Rochel, 1983). Die Multikollinearität kann zu Verzerrungen der Teststatistik führen und erschwert die Interpretation der ß-Gewichte. Anhand der durchgängig signifikanten Interkorrelationen zwischen den Prädiktoren „Schulnoten“ (außer Geschichte et al.) ist bereits zu erkennen, dass die ß-Gewichte in der späteren Regressionsgleichung nicht allein die Produkt-Moment-Korrelation jedes Prädiktors mit dem Kriterium widerspiegeln, sondern auch mit den Einflüssen der anderen Prädiktoren konfundiert sind. Es handelt sich in diesem Falle um einen so genannten Suppressoreffekt.

Welche Schulnoten (und damit welche Merkmale und Fertigkeiten) das erfolgreiche Abschneiden im Auswahlverfahren am besten determinieren, erscheint aufgrund des beschriebenen Suppressoreffektes nicht ganz eindeutig.

Ein weiterer Aspekt der die Interpretation der Ergebnisse beeinflusst, betrifft die Transformation der einzelnen erhobenen Ergebnisse des Auswahlverfahrens. Zulässige Transformationen bezeichnen die Möglichkeit, gegebene Skalenwerte durch andere zu ersetzen, welche die Informationen in gleicher Weise ausdrücken (Borg & Staufenbiel, 1997). Es gehen demnach keine Informationen verloren und keine neuen Aspekte kommen hinzu. Das Skalenniveau der vorhandenen Daten findet man dadurch heraus, indem untersucht wird, ob die Relationen der Zahlen gewisse empirische Relationen darstellen können. Guttman (1972) befindet jedoch, dass die Zuweisung von Merkmalen zu einzelnen Stufen der Skala auf Hypothesen beruht, wie die Werte mit anderen Beobachtungen zusammenhängen. Der Prozess der Festlegung des Skalenniveaus und jener der Transformation unterliegt demnach den gleichen Prozessen und somit Fehlern, die der diagnostischen Urteilsbildung unterliegen. Die erreichten Rohwerte in den einzelnen Testbausteinen des Auswahlverfahrens wurden mittels einer 9er- und 15er-Skala transformiert. Da es sich in diesem Falle um Ratings handelt, unterliegen sie den in Kapitel 2.1 genannten Beurteilerfehlern. Die Art der Gewichtung der Ergebnisse beruht - bis auf die Ausnahme der beiden standardisierten kognitiven Leistungstests - auf den Erfahrungen der auswertenden Mitarbeiter und nicht auf empirisch belegbaren Ergebnissen. Daher kann es zu einer Verzerrung der Ergebnisse kommen. Um diesen Einfluss auf die Ergebnisse aufzufangen, wäre es notwendig gewesen, die Rohwerte der einzelnen Verfahren als Datenbasis zur Analyse zu verwenden. Da jedoch nur die bereits transformierten Ergebnisse zugänglich waren, konnte diese Kompensation leider nicht durchgeführt werden.

Eine weitere Komponente, deren Einfluss nicht unberücksichtigt bleiben sollte, ist die fragliche Reliabilität und Validität der Prädiktoren sowie die Reliabilität der Kriterien. Wie schon unter Kapitel 3,1 dargestellt, ist die Vorhersageleistung eines unreliablen Verfahrens auf alle Fälle gering, die eines hochreliablen steht jedoch in Abhängigkeit von der Reliabilität des Kriteriums (Lienert, 1989). Die unzulänglichen Gütekriterien der Schulnoten (Ojektivität, Reliabilität, Validität) wurde schon in Kapitel 2.3.2.2 dargestellt. Darüber hinaus kann hinsichtlich der einzelnen Komponenten des Testverfahrens, aus deren Gesamtergebnis sich das Kriterium des ersten Teils der Fragestellung zusammensetzt, allenfalls für die

standardisierten Testteile (kulturfairer und nicht kulturfairer kognitiver Leistungstest) eine ausreichend hohe Reliabilität belegt werden. Dagegen wurde die Reliabilität der weiteren Verfahrensteile (Diktat, Berichtsfertigung, Sporttest, Vorstellungs- und Rundgespräch) nicht erschlossen. Die nicht ausreichend hohe Reliabilität etc. der Schulnoten und einzelner Testbausteinen dürfte zu einer Unterschätzung des wahren Zusammenhanges zwischen Prädiktor und Kriterium beitragen. Weiter stützt sich die Zusammenstellung der einzelnen Testteile zu einem Gesamtverfahren nicht auf eine wissenschaftlich begründete Konzeption eines Personalauswahlverfahrens (vgl. Kapitel 2.1) und ist nicht einer vorausgehenden Anforderungsanalyse entsprungen.

Die Regressionsgewichte werden in der Regressionsanalyse so ermittelt, dass die vorhergesagten y-Werte als gewichtete Linearkombinationen der Prädiktoren in der jeweiligen Stichprobe maximal mit den empirischen y-Werten korrelieren. Es stellt sich daher die Frage, ob die gewonnene Regressionsgleichung in anderen Stichproben zur Vorhersage von Kriteriumswerten eingesetzt werden kann oder ob man nicht in einer anderen Stichprobe zu anderen Ergebnissen kommt. Zur Feststellung der Stabilität der Koeffizienten wäre daher die Durchführung einer Kreuzvalidierung von großer Bedeutung gewesen. In der vorliegenden Untersuchung war dies aus zeitlichen Gründen, aufgrund der selektiven Stichprobe und damit einhergehenden Verjüngung der Stichprobengröße jedoch nicht umsetzbar. Die gewonnene Regressionsgleichung ist somit nicht an einer anderen Teilstichprobe überprüft worden, so dass eine Übertragbarkeit der Regressionsgleichung auf andere Teilstichproben und damit eine Vorhersage der Endergebnisse einer anderen Stichprobe als nicht unproblematisch und mit einem mehr oder weniger hohen Fehleranteil angesehen werden muss.

Als Fazit der Untersuchung 1 lässt sich festhalten, dass die oben genannten Schulnoten der Organisation im Vergleich zur bloßen Zufallsauswahl einen finanziellen Gewinn erbringen. Sie scheinen allerdings als Ersatz für bestehenden eignungsdiagnostische selbst wenig geeignet, können jedoch im Rahmen der Vorauswahl, als Bausstein einer Testbatterie und damit einhergehender Vorselektion gewinnbringend verwandt werden. Gerade die Bewertung der Kosten und der Nutzen der Prädiktoren belegt, dass ihre Verwendung auch unter ökonomischen Gesichtspunkten für die Organisation von Bedeutung ist.

5. Fragestellung 2: Analyse der prognostischen Validität des Auswahlverfahrens

5.1 Einführung in die Fragestellung 2

Die sich aus dem Abschnitt 3 ergebenden Hypothesen für die Fragestellung 2 werden ebenfalls in einer prospektiven Längsschnittstudie untersucht. Als Prädiktorvariablen stehen die einzelnen Instrumente des Auswahlverfahrens sowie dessen Gesamtergebnis zur Verfügung. Hierbei wurden die im Testverfahren erreichten Ergebnisse aller Bewerber, die sich für eine Einstellung in den gehobenen Polizeivollzugsdienst bewarben, erhoben. Die Ergebnisse jener Bewerber, die im Rahmen der sequentiellen Auswahlstrategie aufgrund ihrer guten Ergebnisse an allen Testbausteinen teilnahmen und welche am Ende aufgrund ihres Gesamtergebnisses für die Einstellung vorgesehen werden konnten, bildeten die Datenbasis für die Erhebung des Prädiktors. Als Kriteriumsvariable sollte der Studienerfolg in Form der erreichten Gesamtpunktzahl der Studenten in der Zwischenprüfung des Studiums herangezogen werden. Deshalb wurden die Daten des gleichen Probandenkreises 1 ½ Jahre nach deren Einstellung erhoben.

5.2 Beschreibung der erhobenen Prädiktoren

Zur Analyse der prädiktiven Validität des Auswahlverfahrens wurden die von den Testteilnehmern am Auswahlverfahren erreichten Leistungen in den unterschiedlichen Testteilen sowie ihr erreichtes Gesamttestergebnis als Prädiktoren verwendet. Eine ausreichend valide Vorhersage ist nur dann möglich, wenn die Testteile beruflichen Erfolg bzw. Misserfolg mit hinreichend großer Wahrscheinlichkeit vorhersagen können, demnach eine genügend große prognostische Validität aufweisen. Gemäß den in Kapitel 2.2 dargestellten Grundlagen für die Einstellung in den gehobenen Polizeivollzugsdienst wurden folgende Testteile für die Berechnung der prognostischen Validität verwandt (Tab. 5.2-1).

Tab. 5.2-1: Erfasste Prädiktoren zur Beantwortung der Fragestellung 2

1	Punktzahl Lückendiktat	4	Punktzahl Kulturfairer Leistungstest	7	Punktzahl Vorstellungsgespräch
2	Punktzahl Bericht Deutschleistung	5	Punktzahl Sporttest	8	Punktzahl Rundgespräch
3	Punktzahl Bericht Inhalt	6	Punktzahl Nicht kulturfairer Leistungstest	9	Gesamtpunktzahl (Gesamttestergebnis)

Eine ausführliche Beschreibung der Prädiktoren ist in Kapitel 4.3.1 umfassend dargestellt. Die zur Beantwortung der Fragestellung 1 erfassten Kriterien (Instrumente des Auswahlverfahrens) entsprechen inhaltlich den Prädiktoren zur Beantwortung der Fragestellung 2.

Analog wie in Kapitel 4.3 dargestellt, wurden die Leistungen der Teilnehmer in den unterschiedlichen Testbausteinen anhand einer neunstufigen Skala bewertet. Ein Punktwert von eins bildete die schlechteste und ein Punktwert von neun die beste Bewertung ab. Die Transformation der einzelnen Bewertungen (Rohwerte) auf die neunstufige Skala erfolgte in Abhängigkeit des jeweiligen Testverfahrens. Alle Teilnehmer erreichten in allen Testbausteinen den vorgegebenen Mindestwert und durchliefen das gesamte Testverfahren.

5.3 Beschreibung der erhobenen Kriterien

Ein Auswahlverfahren ist dann effizient und zuverlässig, wenn mittels der dadurch stattfindenden Selektion nur diejenigen Teilnehmer ausgewählt werden, die mit hinreichend großer Wahrscheinlichkeit auch beruflich erfolgreich sind. Insofern erhob sich in der vorliegenden Untersuchung die Frage, welche geeignete Kriterien abbilden kann, um den beruflichen Erfolg im Bereich der Ausbildung zum gehobenen Dienst eines Polizeivollzugsbeamten vorherzusagen. Zu den gängigen Problemen sei an dieser Stelle auf Kapitel 2.3 verwiesen. Wie aus der Literatur zu entnehmen, werden allgemein bei der Bestimmung der prognostischen Validität eines Auswahlverfahrens die Ergebnisse in der Ausbildung, d.h. der Noten als Abbild der Ausbildungsleistung und damit interpoliert als ein mögliches Kriterium für späteren Berufserfolg, herangezogen. Dabei bleibt allerdings die Frage offen und durchaus kritisch diskutiert, inwieweit die Ausbildungsleistung als

Vorhersageleistung für den beruflichen Erfolg hinreichend aussagekräftig ist. Wie in Kapitel 2 dargestellt, ist zu beachten, dass zur Gewährleistung des Berufserfolges verschiedene Faktoren zu berücksichtigen sind (z.B. Erfahrung, Soziale Kompetenz u.ä.), die zu diesem Zeitpunkt nicht erhoben werden können. Im Rahmen dieser Arbeit wurde aufgrund der in Kapitel 4.3.2 bereits dargestellten Gründe die Leistung der Studenten in der Zwischenprüfung als Kriteriumsvariable für den Studienerfolg und damit zur Analyse der prognostischen Validität des Auswahlverfahrens verwendet. Weitere Angaben zum Berufserfolg der späteren Beamten des gehobenen Polizeivollzugsdienstes waren nicht beabsichtigt. Um aber den berufspraktischen Anteil der späteren Tätigkeit zu integrieren, wurde analog wie bei der Untersuchung zur Vorhersage der Prädiktionskraft von Schulnoten, auch hier das gemeinsame Ergebnis des theoretischen sowie des praktischen Teils der Zwischenprüfung herangezogen. Das erzielte Abschlussergebnis war als Gesamtpunktzahl sowie als deren Transformation in eine Gesamtnote abbildbar. Zur Analyse der prognostischen Validität der Testteile und des Gesamtergebnisses (Prädiktorvariablen) wurde die in der Zwischenprüfung erreichte Gesamtpunktzahl (Kriterium) verwendet. Somit ergab sich eine transparente Interpretierbarkeit der Daten, da Prädiktoren und Kriterium dasselbe Niveau, nämlich Punktzahl, aufweisen. In der Tabelle 5.3-1 wird das Kriterium zusammenfassend dargestellt.

Tab. 5.3-1: Kriterium zur Bestimmung der prädiktiven Validität des Auswahlverfahrens

Kriterium	Zwischennote Setzt sich zusammen aus den erreichten Einzelergebnissen in der Fächern:
Zwischennote im Studium	Einsatzlehre, Kriminalistik, Allgemeines Verwaltungsrecht/ Polizeirecht, Strafrecht/ Ordnungswidrigkeitenrecht/ Bürgerliches Recht, Kriminologie, Psychologie, Führungslehre, Staatsrecht, Berufspraktische Studienzeit

5.4 Versuchsplan für Fragestellung 2

Zusammenfassend ergab sich für die Beantwortung der Fragestellung 2, wie in Tabelle 5.4-1 dargestellt, folgender Versuchsplan:

Tab. 5.4-1: Versuchsplan zur Beantwortung der Fragestellung 2

<table>
<tr><td colspan="3">Fragestellung 2: Analyse der prognostischen Validität des Auswahlverfahrens</td></tr>
<tr><td colspan="3">TEIL 1</td></tr>
<tr><td>TEILSTICHPROBE</td><td>PRÄDIKTOREN</td><td>KRITERIEN</td></tr>
<tr><td>Nach Ergebnis im Auswahlverfahren geeignete und eingestellte Auszubildende des gehobenen Polizeivollzugsdienstes (Einstellung Oktober 1999)</td><td>a) Punktzahl in einzelnen Bausteinen des Auswahlverfahrens :
- Lückendiktat
- Bericht Deutschleistung
- Bericht Inhalt
- Kulturfairer Leistungstest
- Sporttest
- Nicht kulturfairer Leistungstest
- Vorstellungsgespräch
- Rundgespräch
b) Gesamtpunktzahl im Auswahlverfahren</td><td>Gesamtnote der Zwischenprüfung im Studium</td></tr>
<tr><td colspan="3">TEIL 2</td></tr>
<tr><td colspan="3">Berechnung der Kosten-Nutzen-Relation des Auswahlverfahrens anhand der Payoff-Funktion</td></tr>
</table>

5.5 Durchführung

Die Durchführungen der Datenerhebung zur Beantwortung der zweiten Fragestellung fanden wieder im Zeitraum September 1998 bis April 2001 unter gleichen Bedingungen wie schon unter Kapitel 4.5 beschrieben statt. Durch die sequentielle Auswahlstrategie im Auswahlverfahren sowie der letztendlich begrenzten Anzahl von zur Ausbildung zur Verfügung stehenden Studienplätzen ergab sich eine Selektion in der zur Verfügung stehenden Population (Teilstichprobe). Zur Analyse der prognostischen Validität des Auswahlverfahrens wurden nach Beendigung der Prüfungsphase im Mai 1999 (geplante Einstellung in den gehobenen Polizeivollzugsdienst zum Oktober 1999) die Ergebnisse der erfolgreichen Teilnehmer im Auswahlverfahren (Prädiktor) erhoben. Hierzu dienten (Kapitel 5.3) zum einen die einzelnen Testbausteine des Auswahlverfahrens, zum anderen das erreichte Gesamtergebnis der Teilnehmer. Als Kriterium wurde aus in Kapitel 4 dargestellten Gründen, das zu einem späteren Zeitpunkt erhobene Ergebnis der Studenten in der Zwischenprüfung der Ausbildung zum gehobenen Polizeivollzugsdienst an der Fachhochschule herangezogen (Kriterium).

5.5.1 Beschreibung der Teilstichprobe

Um die prädiktive Validität des Auswahlverfahrens hinsichtlich des Kriteriums Ausbildungserfolg bzw. Studienerfolg erschließen zu können, wurden erneut jene Teilnehmer am Auswahlverfahren als Teilstichprobe herangezogen, die am Verfahren für die Einstellung in den gehobenen Polizeivollzugsdienst erfolgreich teilgenommen und im Oktober 1999 die Ausbildung an der Fachhochschule begonnen hatten (vgl. a. Kapitel 4.5.1.2 mit den Tabellen 4.5-7 – 4.5-12). Die von den damaligen Teilnehmern am Auswahlverfahren erreichten Ergebnisse in den einzelnen Testteilen des Auswahlverfahrens wurden codiert. Nach Beendigung der ersten Datenerhebung (Prädiktor) im Mai 1999 war es somit möglich, die Ergebnisse der in den gehobenen Polizeidienst als „Anwärter auf Probe“ eingestellten ehemaligen Bewerber, anhand ihrer Identifikationsnummern ihren bisherigen Daten zu zuordnen. Neben der im Auswahlverfahren für die Einstellung als geeignet ausgewählten Bewerber, reduzierte sich die Stichprobe aufgrund solcher Bewerber, die einen angebotenen Studienplatz an der Fachhochschule u.a. aufgrund eines Studienplatzes bei der Polizei in einem anderen Bundesland bzw. ihrer Entscheidung für eine anderweitige Ausbildung, nicht antraten. Zudem wurde die Stichprobe um diejenigen reduziert, welche zwar ihr Studium an

der FHÖV begonnen haben, jedoch aufgrund ihrer nicht ausreichenden Leistungen nicht für die nach zwei Semestern zu absolvierende Zwischenprüfung zugelassen wurden. Wie unter Kapitel 4.5.1.2 beschrieben, ergab sich abschließend eine Stichprobengröße von 127 Teilnehmern.

5.5.2 Ablauf der Untersuchung

5.5.2.1 Erhebung des Prädiktors „Auswahlverfahren“

Die Prüfungsphase für die Einstellung in den gehobenen Polizeivollzugsdienst begann im Oktober 1998 und endete im Mai 1999. Analog des Vorgehens bei der Datenaufnahme für das Kriterium im ersten Teil der Untersuchung zur Beantwortung der Fragestellung 1 (Kapitel 4.3.1) wurden hier die Ergebnisse der Teilnehmer in den einzelnen Testbausteinen des Auswahlverfahrens als Prädiktoren erfasst. Sie wurden anschließend für die spätere Berechnung anonymisiert und codiert gespeichert. Die Daten jener Teilnehmer, die für eine Einstellung aufgrund ihres Testergebnisses in Frage gekommen wären, jedoch aufgrund beschriebener unterschiedlicher Ursachen nicht an der späteren Erhebung des Kriteriums teilnehmen konnten, wurden vernichtet. Zu Fragen der Datenerhebung und Bewertung sei an dieser Stelle auf Kapitel 4.5.2.2 verwiesen. Der Ablauf der Durchführung wurde durch den vorgegebenen Auswahlrhythmus vorgegeben. Tabelle 5.5-1 zeigt eine Übersicht der erhobenen Prädiktoren und deren Kennwerte zur Bestimmung der prädiktiven Validität des Auswahlverfahrens.

Tab. 5.5-1: Kennwerte der erhobenen Testergebnisse im Auswahlverfahren als Prädiktoren

TESTBAUSTEINE n = 127	N	M_x	S_x
Lückendiktat	127	6,05	1,48
Bericht Deutschleistung	127	5,66	1,65
Bericht Inhalt	127	6,79	1,64
Kulturfairer Leistungstest	127	7,93	1,13
Sporttest	127	5,47	.84
Nicht kulturfairer Leistungstest	127	21,83	3,57
Vorstellungsgespräch	127	8,22	1,93
Rundgespräch	127	8,02	1,94
Gesamttestergebnis	127	87,69	9,28

5.5.2.2 Erhebung des Kriteriums „Erfolg im Studium“

Die Erhebung des Kriteriums wurde nach Abschluss der Zwischenprüfung und Beendigung des sich anschließenden Berufspraktikums der Studenten durchgeführt (April 2001). Hierzu wurden die Daten der an der Zwischenprüfung teilnehmenden 127 Studenten nach Abschluss dieser durch das Sekretariat der Fachhochschule aufgenommen und anonymisiert weitergegeben. Entsprechend ihrer Codierung konnten die Kriteriumsdaten weiter verarbeitet und dem späteren Prädiktorensatz zugeordnet werden. Dabei wurde der bestehende Datensatz um die Anzahl von Studenten reduziert, welche über die laufenden zwei Semester ausgeschieden waren.

In Kapitel 4.3.2 wurde beschrieben und begründet, wie der Studienerfolg anhand der in der Zwischenprüfung erreichten Gesamtpunktzahl operationalisiert wurde. Die Leistungen der Studenten wurden in den vorgegebenen Fächern in Form von Klausuren und beim Praktikum in Form eines Beurteilungsbogens (Rating) bewertet. Das erzielte Ergebnis wurde auf einer 15-Punkte-Skala abgebildet, bei der 15 Punkte die höchste und 0 die geringste Leistung darstellt. Die Gesamtpunktzahl der Zwischenprüfung (Kriterium) ergab sich wie folgt: je eine Klausur in den Fächern Einsatzlehre und Kriminalistik sowie aus einem der beiden Fächer eine zusätzliche Einzelleistung (in der Regel mündliche Prüfung oder Referat). Weiter gehen je eine Klausur in den Fächern Allgemeines Verwaltungsrecht / Polizeirecht sowie Strafrecht / Strafprozessrecht / Ordnungswidrigkeitenrecht / Bürgerliches Recht ein. Außerdem je eine Klausur in zwei der folgenden Fächer: Kriminologie, Führungslehre, Psychologie und Staatsrecht. Weiter geht die Leistung in der berufspraktischen Studienzeit als Beurteilung ein. In der Tabelle 5.5-2 werden noch einmal die erhobenen Prüfungsfächer sowie die Berechnung des Kriteriums zusammenfassend demonstriert.

In Tabelle 5.5-3 ist die Berechnung anhand einer Beispielrechnung veranschaulichend dargestellt. Tabelle 5.5-4 zeigt die berechneten Kennwerte der zur Verfügung stehenden Stichprobe für die Zwischenprüfung sowie die erreichten Gesamtpunkte (Kriterium).

Tab. 5.5-2: Grundlage zur Berechnung des Gesamtergebnisses (Kriterium) in der Zwischenprüfung

	EINZELERGEBNISSE	KRITERIUM
1	Punktzahl Klausur + mögliche Einzelleistung Einsatzlehre	Aus den erreichten Ergebnissen 1+2 oder 1+3 geht jeweils die Addition dividiert durch zwei ins Gesamtergebnis ein
2	Punktzahl Klausur + mögliche Einzelleistung Kriminalistik	
3	Punktzahl Klausur Allgemeines Verwaltungsrecht / Polizeirecht	
4	Punktzahl Klausur Strafrecht / Ordnungswidrigkeitenrecht / Bürgerliches Recht	Addition des Einzelergebnisses
5	Punktzahl Klausur aus zwei der folgenden zu wählenden Fächern Kriminologie, Psychologie, Führungslehre, Staatsrecht	Addition der Einzelergebnisse
6	Punktzahl Beurteilung der Berufspraktischen Studienzeit	Addition des Einzelergebnisses
	Gesamtergebnis	Addition aller Einzelergebnisse der Zwischenprüfung und Division durch ihre Anzahl

Tab. 5.5-3: Beispiel für ein erreichtes Gesamtergebnisses in der Zwischenprüfung (Kriterium)

	EINZELERGEBNISSE ZWISCHENPRÜFUNG *	GEWICHTUNG	ENDERGEBNIS
1	Punktzahl Klausur Einsatzlehre	Aus den erreichten Ergebnissen 1+3 geht jeweils die Addition dividiert durch zwei ins Gesamtergebnis ein	erreicht: 9 Punkte
2	Punktzahl Klausur Kriminalistik		erreicht: **14 Punkte**
3	Prüfling entscheidet sich für Referat als Einzelleistung im Fach Einsatzlehre Punktzahl Referat Einsatzlehre		erreicht in Klausur: 9 Punkte erreicht im Referat: 5 Punkte / 2 **= 7 Punkte**
4	Punktzahl Klausur Allgemeines Verwaltungsrecht / Polizeirecht	Addition des Einzelergebnisses	erreicht: **9 Punkte**
5	Punktzahl Klausur Strafrecht / Ordnungswidrigkeitenrecht / Bürgerliches Recht	Addition des Einzelergebnisses	erreicht: **8 Punkte**
6	Punktzahl Klausur aus zwei der folgenden zu wählenden Fächern: Kriminologie, Psychologie, Führungslehre, Staatsrecht	Addition der Einzelergebnisse	erreicht: **8 und 9 Punkte**
Gesamtergebnis der Zwischenprüfung		Addition der Einzelergebnisse: 14+7+9+8+8+9 Punkte Division durch Anzahl = 6 **= 9,16 Punkte**	

* 15er Skala

Wie schon in Abschnitt 4.1 und 4.2 erklärt und beschrieben, wurde für die Ermittlung der prognostischen Validität des Prädiktors „Auswahlverfahren“ die Gesamtpunktzahl statt der

erreichten Note in der Zwischenprüfung als Kriteriumsvariable verwendet. Weiter wurde im Rahmen der Bewertung des Verfahrens mittels einer schrittweisen Regression ein signifikanter R^2- Zugewinn untersucht und beleuchtet, inwieweit durch schrittweise Zunahme einzelner Testteile des Auswahlverfahrens ein signifikanter Zugewinn an Vorhersageleistung erreicht wird (inkrementelle Validität). Durch den Vergleich der bivariaten Korrelationen der einzelnen Testteile des Auswahlverfahrens mit der multiplen Korrelation (Vorhersagekraft aller Elemente des Auswahlverfahrens zusammengenommen) wurde die Hypothese untersucht, inwieweit die Prädiktionskraft aller Elemente des Auswahlverfahrens zusammengenommen höher ist, als die prognostische Validität einzelner Elemente. Die Ergebnisse wurden mittels der F-Statistik auf Signifikanz getestet. Für die Analyse der Vorhersageleistung einzelner Elemente des Auswahlverfahrens wurden einfache regressionsanalytische Analysen (Produkt-Moment-Korrelation) angewandt. Hingegen wurde für die Berechnung der prognostischen Validität des Gesamtergebnisses im Auswahlverfahren (Variablengruppen auf Prädiktorseite) multiple Regressionsanalysen gerechnet und auf Signifikanz getestet.

Tab. 5.5-4: Kennwerte der Ergebnisse nach der Zwischenprüfung (Kriterium)

GESAMTERGEBNIS ZWISCHENPRÜFUNG n = 127	N	Mx	Sx
Gesamtpunkte	127	8,82	1,45

5.5.2.3 Berechnung der Kosten-Nutzen-Relation

Analog der Darstellung in Kapitel 4.5.2.4 wurde neben der prädiktiven Validität des Auswahlverfahrens auch dessen Nutzen und Kosten bezüglich des ereichten Ergebnisses anhand der Payoff-Funktion nach dem Brodgen-Cronbach-Gleser-Modell (Höft, 2001) in Anlehnung an Kersting (2004) bestimmt. Die Berechnung der Payoff-Funktion erfolgte nach folgender Formel (Abb. 5.5-1):

$$\Delta U = NE \cdot T \cdot SDY \cdot r_{xy} \cdot \bar{z}_x - C \cdot NB$$

Abb. 5.5-1: Payoff-Funktion des Brodgen-Cronbach-Gleser-Modells (Höft, 2001)

Die inhaltliche Bedeutung der Parameter ist in der Tabelle 5.5-5 zusammengefasst.

Zur Berechnung der Payoff-Funktion wurden die verfügbaren Daten der Gesamtstichprobe und Teilstichprobe herangezogen sowie durch die erschlossenen Ergebnisse zur Bestimmung der prädiktiven Validität der Testbausteine des Auswahlverfahrens ergänzt. Da nicht alle für die Formel notwendigen Daten vorlagen, wurden verschiedene Parameter aufgrund gegebener Erfahrungswerte geschätzt bzw. interpoliert.
Für die Berechnung der Payoff-Funktion wird, wie bereits in Abschnitt 4.7 angesprochen, konservativ der unkorrigierte Validitätskoeffizient in die Analyse aufgenommen, um eine Überkorrektur zu verhindern.

Tab. 5.5-5: Parameter der Payoff-Funktion
des Brodgen-Cronbach-Gleser-Modells (Höft, 2001)

Δ U	Der Nutzenzuwachs in Geldeinheiten der durch das Testverfahren entsteht	
NE	Anzahl der Eingestellten	**n = 134**
T	Anzahl der berücksichtigten Zeiteinheiten, welche die Bewerber durchschnittlich in Jahren in einer Institution verbringen	**10 Jahre**
SDY	Standardabweichung der Berufsleistung in Geldeinheiten 40 % des jährlichen Entgeltes x 2*r*	Besoldungsstufen A 10 – A12 (Differenz zwischen den besten Mitarbeitern und durchschnittlichen Mitarbeitern) Durchschnittliches Jahresentgelt in Euro: A 10: 3052,69 * 12 = 36632,28 A 10: 3452,85 * 12 = 41434,20 A 10: 3843,33 * 12 = 46119,96 = (36632,28+41434,20+46119,96)/3 = 41395,48 41395,48 * 2 = 82,790,96 40 % **= 33116,38**
r_{xy}	Validitätskoeffizient zwischen den durch die Teilnehmer erreichten Gesamttestergebnis im Auswahlverfahren und ihrem Ergebnis in der Zwischenprüfung	zu berechnen
$\bar{z}_x$	Durchschnittlicher standardisierter Testwert (Prädiktorwert) der Ausgewählten	zu berechnen
C	Kosten für das Testverfahren pro Anwendung	Geschätzte Kosten in Euro pro Bewerber für: Mitarbeiter Annahme Bewerbungen Mitarbeiter Durchführung der Testung Mitarbeiter ärztliche Untersuchung Mitarbeiter polizeiliche Überprüfung Mitarbeiter Einstellung (siehe auch Kapitel 2.2.2)
NB	Anzahl der Bewerber	N = 735

5.6 Ergebnisdarstellung

Wie bereits in Abschnitt 4.6 beschrieben, muss zur Analyse der prognostischen Validität berücksichtigt werden, dass es sich bei der verwendeten Teilstichprobe, der nach dem Ergebnis im Auswahlverfahren geeigneten und eingestellten Bewerbern, um eine einseitig selegierte Stichprobe handelt. Durch die einseitige Auswahl der Bewerber, und somit des Wegfallens des linken Teils der Verteilung, wird in der Regel der statistische Zusammenhang zwischen Prädiktor und Kriterium in der Teilstichprobe gegenüber dem der Gesamtgruppe der Bewerber eher unterschätzt. Bei der Analyse der prognostischen Validität der Testbausteine hinsichtlich der Gesamtpunktzahl in der Zwischenprüfung wird somit ebenfalls die von Lienert (1989) angegebene Formel zur Berechnung der Selektionskorrektur durchgeführt (vgl. Lienert, 1989, S. 307). Die angegebenen Korrelationskoeffizienten sind folglich für die Varianzeinschränkung korrigiert, ergänzend werden die unkorrigierten Werte jeweils darunter in Klammern mit der Signifikanzangabe ausgewiesen. Wenn die Korrelation auf dem Niveau 0,05 % zweiseitig signifikant wird, erfolgt die Angabe des Signifikanzniveaus mit *„“, bei einer Korrelation auf dem 0,01 % -Niveau erfolgt die Angabe „** “.

5.6.1 Darstellung der prognostischen Validität des Auswahlverfahrens

Gemäß den unter Kapitel 3.3 dargestellten Hypothesen ergaben sich zu deren Beantwortung zur Analyse der prognostischen Validität des Auswahlverfahrens (Kriterium Zwischenprüfung) nachfolgende Ergebnisse.

In Tabelle 5.6-1 sind die Einzelkorrelationen (nach Pearson) der Punkte in den einzelnen Elementen des Auswahlverfahrens (Prädiktoren) mit dem Kriterium Endergebnis in der Zwischenprüfung (Gesamtpunkte) dargestellt.

In den Tabellen 5.6-2 – 5.6-4 sind die Ergebnisse der Regression (Einschluss) der Prädiktoren „Bausteine des Auswahlverfahrens“ hinsichtlich des Kriteriums „Endergebnis in der Zwischenprüfung“ (Gesamtpunkte) dargestellt.

Tab. 5.6-1: Ergebnis Korrelation Testbausteine Auswahlverfahren Endergebnis Zwischenprüfung

Anzahl der Teilnehmer	Testbausteine des Auswahlverfahrens	Endergebnis der Zwischenprüfung
127	Lückendiktat	.032 (.029)
127	Bericht Inhalt	.238 .(233**)
127	Bericht Deutschleistung	.057 (.043)
127	Kulturfairer Leistungstest	.223 (.158)
127	Sporttest	.135 (.080)
127	Nicht kulturfairer Leistungstest	.215 (.165)
127	Vorstellungsgespräch	.172 (.127)
127	Rundgespräch	.231 (.173)
127	Gesamttestergebnis	.663 (.306**)

Tab. 5.6-2: Regression: Prädiktor „Auswahlverfahren“, Kriterium „Zwischenprüfung“; Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	.378[a]	.143	.084	1,3789

a. Einflussvariablen: (Konstante) Punkte Lückendiktat, Bericht Inhalt, Bericht Deutsch, Kulturfairer Leistungstest, Sporttest, Nicht kulturfairer Leistungstest, Vorstellungsgespräch, Rundgespräch

Tab. 5.6-3: Regression: Prädiktor „Auswahlverfahren“, Kriterium „Zwischenprüfung“; ANOVA [b]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
1	Regression	36,995	8	4,624	2,432	.018[a]
	Residuen	222,476	117	1,902		
	Gesamt	259,470	125			

a. Einflussvariablen: (Konstante) Punkte Lückendiktat, Bericht Inhalt, Bericht Deutsch, Kulturfairer Leistungstest, Sporttest, Nicht kulturfairer Leistungstest, Vorstellungsgespräch, Rundgespräch
b. Abhängige Variable: Endergebnis Zwischenprüfung

Tab. 5.6-4: Regression: Prädiktor „Auswahlverfahren“, Kriterium „Zwischenprüfung“ Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz
		B	Standardfehler	Beta		
1	(Konstante)	3,558	1,633		2,179	..031
	Lückendiktat	5,540E-02	.093	.057	.596	.552
	Bericht Inhalt,	.208	.078	.237	2,651	.009
	Bericht Deutsch	-6,32E-02	.081	-.072	-.782	.436
	Kulturfairer Leistungstest	.249	.118	.195	2,107	.037
	Sporttest	6,787E-02	.151	.039	.449	.654
	Nicht kulturfairer Leistungstest	3,303E-02	.038	.082	.872	.385
	Vorstellungsgespräch	1,288E-02	.079	.017	.163	.871
	Rundgespräch	.128	.084	.173	1,529	.129

a. Abhängige Variable: Endergebnis Zwischenprüfung

5.6.2 Darstellung der inkrementellen Validität des Auswahlverfahrens

In den Tabellen 5.6-5 – 5.6-7 sind die Ergebnisse der Regression (stepwise) der Prädiktoren „Bausteine des Auswahlverfahrens“ hinsichtlich des Kriteriums „Endergebnis in der Zwischenprüfung“ (Gesamtpunkte) zur Erschließung der inkrementellen Validität dargestellt.

Tab. 5.6-5: Regression: Prädiktor „Auswahlverfahren“, Kriterium „Zwischenprüfung“; Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	.257[a]	.066	.059	1,3979
2	.311[b]	.097	.082	1,3804
3	.357[c]	.128	.106	1,3621

a Einflussvariablen: (Konstante) Punkte Bericht Inhalt
b Einflussvariablen: (Konstante) Punkte Bericht Inhalt, Kulturfairer Leistungstest
c Einflussvariablen: (Konstante) Punkte Bericht Inhalt, Kulturfairer Leistungstest, Rundgespräch

Tab. 5.6-6: Regression: Prädiktor „Auswahlverfahren“, Kriterium „Zwischenprüfung“; ANOVA [d]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
1	Regression	17,148	1	17,148	8,775	.004 [a]
	Residuen	242,322	124	1,954		
	Gesamt	259,470	124			
2	Regression	25,092	2	12,546	6,584	.002 [b]
	Residuen	234,378	123	1,906		
	Gesamt	259,470	125			
3	Regression	33,129	3	11,043	5,952	.001 [c]
	Residuen	226,341	122	1,855		
	Gesamt	259,470	125			

a Einflussvariablen: (Konstante) Punkte Bericht Inhalt
b Einflussvariablen: (Konstante) Punkte Bericht Inhalt, Kulturfairer Leistungstest
c Einflussvariablen: (Konstante) Punkte Bericht Inhalt, Kulturfairer Leistungstest, Rundgespräch
d Abhängige Variable: Endergebnis Zwischenprüfung

Tab. 5.6-7: Regression: Prädiktor „Auswahlverfahren“, Kriterium „Zwischenprüfung“ Koeffizienten [a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz
		B	Standardfehler	Beta		
1	(Konstante)	7,564	.447		16,906	.000
	Bericht Inhalt	.226	.076	.257	2,962	.004
2	(Konstante)	5,789	.975		5,936	.000
	Bericht Inhalt,	.227	.075	.258	3,014	.003
	Kulturfairer Leistungstest	.223	.109	.175	2,042	.043
3	(Konstante)	4,492	1,147		3,918	.000
	Bericht Inhalt,	.206	.075	.234	2,747	.007
	Kulturfairer Leistungstest	.266	.110	.208	2,422	.017
	Rundgespräch	.134	.064	.181	2,081	.039

a. Abhängige Variable: Endergebnis Zwischenprüfung

5.6.3 Darstellung der Kosten-Nutzen-Relation

Wie unter Kapitel 4.5.2.4 erläutert, erfolgt die Berechnung der Payoff-Funktion des Brodgen-Cronbach-Gleser-Modells für das erreichte Ergebnis im Auswahlverfahren nach der Formel: $\Delta U = NE \cdot T \cdot SD_Y \cdot r_{xy} \cdot z_x - C \cdot NB$. Die Darstellung der Berechung und das Ergebnis sind in Tabelle 5.6-8 wiedergegeben.

Die Anwendung des Prädiktors „Gesamttestergebnis“ als Auswahlinstrument ergibt im Vergleich zu einer bloßen Zufallsauswahl ($r_{xy} = 0$) für die nächsten knapp 10 Jahre einen

Nutzenzuwachs von ca. 19 Millionen Euro.

Tabelle 5.6-8: Berechnung der Payoff-Funktion des Prädiktors „Gesamttestergebnis“

ΔU	Der Nutzenzuwachs in Geldeinheiten, der durch das Testverfahren entsteht	
NE	Anzahl der Eingestellten	**134**
T	Anzahl der berücksichtigten Zeiteinheiten, welche die Bewerber durchschnittlich in Jahren in einer Institution verbringen	**10 Jahre**
SD_Y	Standardabweichung der Berufsleistung in Geldeinheiten 40 % des jährlichen Entgeltes x 2	**33116,38 Euro**
r_{XY}	Validitätskoeffizient des Gesamttestergebnisses	**r = .306**
z_X	Durchschnittliche standardisierte Testwerte (Prädiktorwert)	**1,53**
C	Kosten für das Testverfahren pro Anwendung	**2500 Euro**
NB	Anzahl der Bewerber	**735**
Nutzenzuwachs in Geldeinheiten: **= (134* 10 * 33116,38 * .306 * 1,53) – (2500 * 735)** **= 18.938.431,9 Euro**		

5.6.4 Beantwortung der Hypothesen

Die in Kapitel 3.3.2.1 formulierten Hypothesen zur Analyse der prognostischen Validität der einzelnen Testbausteine des Auswahlverfahrens und des erreichten Gesamtergebnisses der Teilnehmer hinsichtlich des Kriteriums „Gesamtpunkte Zwischenprüfung“ können wie folgt beantwortet werden:

Wie aus der Tabelle 5.6-1 ersichtlich, ergaben sich zwischen den Testbausteinen Bericht Inhalt (r = .238) und dem von den Teilnehmern erreichten Gesamttestergebnis (r = .663) auf dem 1% Niveau signifikante nachweisbare empirische Zusammenhänge.
Wie aus den Tabellen 5.6-2 und 5.6-3 zu entnehmen, ergibt sich ein korrigiertes R^2 von .084.
Die Regressionsgleichung ist mit F = 2,432 signifikant. Dabei leisten die Prädiktoren „Bericht Inhalt“ und „Kulturfairer Leistungstest“ einen signifikanten Beitrag zur Regressionsgleichung (Tab. 5.6-4).

Die in Kapitel 3.3.2.2 formulierte Hypothese zur Analyse der inkrementellen Validität des Auswahlverfahrens kann wie folgt beantwortet werden:

Wie aus den Tabellen 5.6-5 – 5.6-7 ersichtlich, wird die Regressionsgleichung mit dem Prädiktor „Bericht Inhalt“ mit einem korrigierten R^2 von .059 und dem F-Bruch von 8,775 auf

dem 1% Niveau signifikant. Bei Hinzunahme des Prädiktors „Kulturfairer Leistungstest“ ergibt sich eine Steigerung des korrigierten R^2 auf .082 und damit ein Zuwachs an inkrementeller Validität. Die Regressionsgleichung wird mit einem F-Bruch von 6,584 auf dem 1% Niveau signifikant. Bei Hinzunahme der letzten Einflussvariable „Rundgespräch“ erreichen diese drei Prädiktoren eine korrigierte multiple Korrelation von $R^2 = .102$. Die Regressionsgleichung wird ebenfalls mit einem F-Bruch von 5,952 auf dem 1 % Niveau signifikant.

5.7 Interpretation der Ergebnisse

Äquivalent zu Untersuchung 1 ist das Ziel dieser zweiten Untersuchung die Erschließung der prognostischen Validität. Als Prädiktoren dienen die einzelnen Testbausteine des Auswahlverfahrens sowie dessen Gesamtergebnis. Die erreichten Ergebnisse aller Teilnehmer am Auswahlverfahren für die Einstellung in den gehobenen Polizeivollzugsdienst standen hierfür zur Verfügung. Diejenigen Bewerber, welche aufgrund ihrer guten Ergebnisse an allen Testbausteinen im Rahmen der sequentiellen Auswahlstrategie haben teilnehmen können, und aufgrund ihres Gesamtergebnisses für die Einstellung vorgesehen werden konnten, dienen als Datenbasis für die Erhebung des Prädiktors. Kriteriumsvariable ist der Studienerfolg in der nach 1½ Jahren stattfindenden Zwischenprüfung, operationalisiert als erreichte Gesamtpunktzahl.

Bei der Zusammenstellung einzelner Testbausteine zu einem Verfahren geht es darum, anhand unterschiedlicher Informationsquellen (bspw. Leistungstest, Persönlichkeitstest, Schulnoten etc.) eine möglichst breite Datenbasis zu schaffen, mittels derer das Kriterium möglichst umfassend vorhergesagt werden kann. Ein Auswahlverfahren sollte in der Art zusammengesetzt sein, dass jeder Testbaustein durch seine Berücksichtigung im Gesamtergebnis die prognostische Validität des gesamten Auswahlverfahrens erhöht. Ein zusätzlich verwendeter Testbaustein bzw. Prädiktor, besitzt demnach dann inkrementelle Validität, wenn sich der Anteil an aufgeklärter Varianz am Kriterium bei dessen Hinzunahme in ein bestehendes Verfahren erhöht. Welche Testbausteine einen Zuwachs an inkrementeller Validität leisten, kann anhand der schrittweisen Regressionsanalyse geklärt werden.
Bei der Bewertung eines Auswahlverfahrens geht es, wie bereits in Abschnitt 4.7 beschrieben, nicht nur um die Maximierung der Vorhersageleistung der verwendeten Prädiktoren, sondern ebenso um deren Betrachtung unter Kosten-Nutzen-Aspekten. Aufgrund dessen wird für den Prädiktor aus dem Auswahlverfahren mit der höchsten prognostischen Validität die

Kosten-Nutzen-Relation anhand der in Abschnitt 2.1 beschriebenen Payoff-Funktion des Brodgen-Cronbach-Gleser-Modells berechnet.

Bei der Beantwortung der Hypothesen wurden nur zwei von neun Korrelationen auf dem 1 % Niveau signifikant. Für den Prädiktor Testbaustein „Bericht Inhalt“ ergab sich eine korrigierte prognostische Validität von r = .24 und für den Prädiktor „Gesamttestergebnis im Auswahlverfahren“ eine korrigierte prognostische Validität von r = .66 im Hinblick auf das Kriterium „Endergebnis in der Zwischenprüfung“. Durch die theoretisch mögliche Verteilung der erreichbaren Werte im Gesamtergebnis (Summe aller erzielten Testergebnisse) kommt es hier im Vergleich zu den theoretisch möglichen Verteilungen erreichbarer Werte auf der Einzeltestebene zu einer deutlichen Erhöhung der Varianz. Da sich in diesem Fall die Selektion (Bestenauslese) aufgrund der grundsätzlich erreichbaren Gesamttestergebnisse deutlicher auswirkt, zeigt sich dieser Effekt insbesondere in der Selektionskorrektur durch eine erhöhte Maßzahl für das Zusammenhangsmaß. Die weiteren Prädiktoren weisen keine signifikanten Validitätskoeffizienten mit dem Kriterium auf.

In praxi ist bereits eine Höhe des Validitätskoeffizienten von r = .50 eine sehr gute Prognose des beruflichen Erfolges, wobei man sich häufig jedoch bereits mit geringeren Validitätskoeffizienten zufrieden geben muss (Lienert (1989). Nach Rösler (1992) erreichen die beobachteten Validitätskoeffizienten vieler Studien im Durchschnitt über verschiedene Test- und Kriterienbereiche hinweg nur Werte um .2 und .3. In Anbetracht dessen kann die prädiktive Validität des Gesamttestergebnisses im Auswahlverfahren als ein gutes Ergebnis bezeichnet werden. Der Prädiktor verspricht eine hinreichend valide und somit gute Vorhersage für das erfolgreiche Abschneiden in der Zwischenprüfung.

Die prädiktive Validität des Testbausteines „Bericht Inhalt“ ist dagegen eher gering. In der eignungsdiagnostischen Praxis werden Korrelationskoeffizienten in einem Bereich von .3 als durchschnittlich bezeichnet (Schuler (1996), so dass dieser Prädiktor für eine alleinige Verwendung nahezu keinen praktischen Wert besitzt.

Die Regressionsgleichung (overall) wird mit einer multiplen Korrelation von R = .38 und einem R^2 von .14 signifikant. Durch die verwendeten Prädiktoren werden 14,3 % der Varianz des Kriteriums „Endergebnis in der Zwischenprüfung“ aufgeklärt. Dieser Prozentwert bedeutet jedoch auch, dass ein erheblicher Anteil an der Varianz des Kriteriums durch andere Einflüsse als die der verwendeten Prädiktoren determiniert wird.

Die ß-Gewichte des Prädiktors „Bericht Inhalt“ und des Prädiktors „Kulturfairer Leistungstest“ leisten einen signifikanten Beitrag zur Vorhersage des Kriteriums

„Endergebnis in der Zwischenprüfung“. Die t-Tests der anderen nicht genannten Prädiktoren werden nicht signifikant.

Zur Überprüfung der Stabilität der Koeffizienten in der Regressionsgleichung wird das gewonnene Regressionsmodell in einer anderen Stichprobe zur Vorhersage von Kriteriumswerten eingesetzt und umgekehrt. Es interessiert die Frage, ob die gewonnene Regressionsgleichung zur Vorhersage von Kriteriumswerten geeignet ist und ob man in einer anderen Stichprobe zu anderen Ergebnissen kommt. Daher wäre, wie bereits in Abschnitt 4.7 angesprochen, die Durchführung einer Kreuzvalidierung von Bedeutung gewesen. Für die Untersuchung der beiden Fragestellungen 1 und 2 war dies jedoch aus zeitlichen Gründen und aufgrund der relativ kleinen Stichprobe zur Erhebung der Kriteriumsdaten nicht umsetzbar, so dass die Frage der Übertragbarkeit der Regressionsgleichung auf andere Teilstichproben und damit eine Vorhersage des Ergebnisses in der Zwischenprüfung einer anderen Bewerberstichprobe als nicht verifiziert angesehen werden muss. Da jedoch nur zwei der acht Prädiktoren aus dem Regressionsmodell überhaupt signifikante ß-Gewichte erreichen und damit die aufgeklärte Varianz mit 14,3% vergleichsweise eher gering ist, ist die fehlende Kreuzvalidierung für das in dieser Kombination von Prädiktoren zusammengesetzte Auswahlverfahren in diesem Kontext nicht von herausragender Bedeutung. Diesbezüglich sind die Ergebnisse der Analyse der inkrementellen Validität von Interesse, da nur diejenigen Prädiktoren in das Modell aufgenommen werden, die einen signifikanten Beitrag zur Vorhersage des Kriteriums liefern.

Die Kombination der verwendeten Testbausteine zu einem Auswahlverfahren scheint zwar einen signifikanten Einfluss auf das Abschneiden in der Zwischenprüfung zu haben, bei Betrachtung der ß-Gewichte und deren Signifikanztests sind es jedoch nur zwei der acht Prädiktoren des Regressionsmodells, welche signifikante Beiträge zur Vorhersage erbringen. Wie groß ist demnach der Zugewinn an prognostischer Validität durch die einzelnen Prädiktoren?

Die Verwendung vieler einzelner Testbausteine im Auswahlverfahren unterstellt, dass diese Elemente jeweils auch einen Beitrag zur Leistung des Gesamtverfahrens erbringen. Anderenfalls würde zwar ein hoher Aufwand an Zeit, Material und Kosten für die Konstruktion und Durchführung eines Verfahrens betrieben, ohne dass dieser durch eine entsprechende Verbesserung der Entscheidungsgrundlage belohnt würde. Die Ergebnisse der Tabellen 5.6-5 bis 5.6-7 zeigen für die Testbausteine „Bericht Inhalt“, „Kulturfairer Leistungstest“ und „Rundgespräch“ inkrementelle Validität. Dieses Regressionsmodell mit

den drei genannten Prädiktoren weist eine multiple Korrelation von R = .36 auf und klärt 12,8 % der Kriteriumsvarianz auf. Dieses Ergebnis weist daraufhin, wie das der Varianzaufklärung des Regressionsmodells mit allen Prädiktoren, dass jedoch auch ein erheblicher Anteil an Varianz des Kriteriums (ca. 87%) nicht durch die Prädiktoren aufgeklärt wird. Positiv formuliert gibt es drei Testbausteine, die einen signifikanten Beitrag zur Verbesserung der Vorhersage des Endergebnisses in der Zwischenprüfung leisten. Die aufgeklärte Varianz dieses Modells liegt mit 12,8 % und einer multiplen Korrelation von R = .36 jeweils etwas unter den Werten des Regressionsmodells mit allen aufzunehmenden Prädiktoren (R^2 = .143, R = .378). Es ist die Folge der statistischen Prozedur, dass nur die Prädiktoren in das Modell aufgenommen werden, welche mit ihrer Nützlichkeit zur Vorhersage des Kriteriums über einem Minimalwert liegen. Variablen, die diesen Minimalwert überschreiten, gelten als redundant und werden nicht in das Modell aufgenommen.

Interessant ist der Aspekt, dass der Prädiktor „Rundgespräch“, der in der Regressionsgleichung aller Prädiktoren (Tabelle 5.6-4) kein signifikantes ß-Gewicht erreicht, bei der schrittweisen Regression jedoch als dritter Prädiktor mit einer über dem Minimalwert liegenden Nützlichkeit in das Modell aufgenommen wird. Für das Prozedere der schrittweisen Regression gilt, wie oben angesprochen, dass nur diejenigen Prädiktorvariablen in das Modell aufgenommen werden, die als weitere Variable das Vorhersagepotential (R^2) der bereits im Modell enthaltenen Variablen maximal erhöht. Im Gegensatz dazu werden in der Regressionsanalyse (Einschluss) alle Variablen in das Modell aufgenommen, unabhängig davon, ob sie ein Mindestmaß an Beitrag zur Vorhersage des Kriteriums leisten oder nicht. Da die beiden Prädiktoren „Kulturfairer Leistungstest“ und „Rundgespräch“ stark negativ interkorreliert sind (r = .186 **), jeweils jedoch beide positiv mit dem Kriterium korrelieren, könnte man vermuten, dass hier ein reziproker Suppressionseffekt vorliegt. In diesem Falle würden die beiden Prädiktoren als Suppressorvariablen wechselseitig irrelevante Varianzanteile unterdrücken (vgl. Bortz, 1993).

Wie nicht anders zu erwarten, trägt auch in dieser Untersuchung der durchgeführte kognitive Leistungstest einen signifikanten Anteil an der aufgeklärten Varianz bei der Bestimmung der inkrementellen Validität bei. In einer großen Zahl von Untersuchungen konnte die prädiktive Validität kognitiver Fähigkeitstests belegt werden. Schmidt und Hunter (1998) sprechen von dem wichtigsten eignungsdiagnostischen Instrument für Einstellungsentscheidungen. Schuler (1996) kommt zur Einschätzung, dass es praktisch keinen Beruf gibt, für den Maße

intellektueller Fähigkeiten nicht zur Leistungsprognose beitragen können. Hunter und Hunter (1984) kommen zu dem Ergebnis, dass die Validität der Leistungsprognose umso höher ausfällt, je höher die Komplexität des betroffenen Berufes ist. Andere Verfahren können häufig als Ergänzungen zu Intelligenztests gesehen und im Zusammenhang mit der Erschließung der inkrementellen Validität bewertet werden.
Wie schon in Kapitel 2.3 dargestellt, ergibt sich bei Kombination mit dem zweiten Prädiktor unstrukturiertes Interview gegenüber einem Intelligenztests ein Zuwachs an Validität von r = .07 (Schmidt & Hunter, 1998). Ein vergleichbarer Effekt könnte auch durch das teilweise interviewartige Rundgespräch möglich sein. Darüber hinaus wäre es denkbar, dass im Rundgespräch ähnliche Faktoren einen Beitrag zur inkrementellen Validität leisten können, wie dies im Rahmen der Durchführung des Assessment Centers von Bedeutung ist. Aus der Metaanalyse von Scholz und Schuler (1993) ergaben sich für Intelligenz, soziale Kompetenz, Dominanz und Selbstvertrauen die größten Zusammenhänge mit dem AC-Gesamturteil. Nach Klimoski und Brickner (1987) könnten allgemeine kognitive Fähigkeiten hinter den eigentlich zu erfassenden Anforderungen stehen. Da letztere schon grundlegend durch das Ergebnis im kulturfairen Leistungstest eingegangen sind, wäre in einem weiteren Schritt zu untersuchen, ob nicht Aspekte allgemeiner sozialer Kompetenz und weitere Persönlichkeitsmerkmale zusätzlich im Rundgespräch erfasst und tragend für ein hohes Ratingergebnis sein könnten. Den ersten und größten Beitrag zur inkrementellen Validität stellte die inhaltliche Bewertung der Deutschleistung dar. Neben auch hier möglicherweise zum Ausdruck kommender allgemeiner intellektueller Fähigkeiten, die über die schriftliche Darstellung erfasst werden, ist die Schriftform im Rahmen der schulischen Ausbildung an der Fachhochschule sicherlich eines der wichtigsten Instrumentarien im Rahmen der Bewertung (Klausuren während der Studienzeit und schriftliche Abschlussprüfungen).

Insgesamt wird die Hypothese, dass jedes Element im Auswahlverfahren durch seine Berücksichtigung signifikant die prognostische Validität des Gesamtverfahrens erhöht, nicht bestätigt. Positiv formuliert bedeutet dies, dass es verschiedene Elemente des Auswahlverfahrens gibt, die durch ihre Berücksichtigung im Gesamtergebnis die prognostische Validität des Auswahlverfahrens erhöhen.

Die Ergebnisse der korrelations- und der regressionsanalytischen Berechnungen zeigen, dass demnach „nur“ die drei genannten Prädiktoren aus der Kombination aller Testbausteine des Auswahlverfahrens einen Beitrag zur Vorhersage des Ergebnisses in der Zwischenprüfung

erbringen. Die restlichen Prädiktoren haben keinen bedeutsamen statistisch nachweisbaren Einfluss auf das Kriterium.

Wenn bei der Auswahl von Bewerbern eine im Sinne des Kriteriums bessere Zuordnung als bei einem zufälligen Vorgehen erreicht werden soll, ist die prädiktive Validität unerlässlich. Die Anwendung eines validen Prädiktors garantiert jedoch nicht, dass das Auswahlverfahren auch unter betriebswirtschaftlichen Aspekten rentabel ist. Diese Effizienz wird, wie in der Payoff-Funktion zu sehen ist, neben der Validität des Prädiktors noch von anderen Faktoren bestimmt. Für den Prädiktor mit der höchsten prognostischen Validität, das Gesamttestergebnis im Auswahlverfahren, ergibt sich unter Berücksichtigung der Kosten-Nutzen Analyse folgendes Bild: Im Vergleich zu einer bloßen Zufallsauswahl (r_{xy} = 0) ergibt dieser Einzelprädiktor für die nächsten knapp 10 Jahre einen Nutzenzuwachs von ca. 19 Millionen Euro. In den meisten Fällen in der eignungsdiagnostischen Praxis geht es jedoch nicht nur um den Vergleich des Nettonutzens eines Prädiktors mit dem einer zufälligen Auswahl von Bewerbern. Vielmehr werden zwei oder mehrere Verfahren bzw. Prädiktoren hinsichtlich ihrer Validitätskoeffizienten und hinsichtlich ihres Nutzens verglichen. Bei dem Vergleich des Prädiktors „Gesamttestergebnis im Auswahlverfahren“ als den Prädiktor mit der höchsten prognostischen Validität mit einem Prädiktor, welcher einen niedrigeren Validitätskoeffizienten aufweist, in diesem Falle exemplarisch der Kulturfaire Leistungstest mit einem Validitätskoeffizienten von r = .158, ergibt sich ein immer noch Nutzenzuwachs des „Gesamttestergebnisses“ von 10.048.489,94 Euro (Nutzen „Gesamttestergebnis“ 18.938.431,9 Euro - Nutzen „Kulturfairer Leistungstest“ 8.889.941,96 Euro).
In Anlehnung an das von Kersting (2004) verwendete Beispiel (Nettonutzen = 300.500 Euro) und dessen Parameter ergibt sich zur Beurteilung der Aussagekraft der vorliegenden Ergebnisse für den Prädiktor „Gesamttestergebnis“ (zwei eingestellte und 14 eingeladene Bewerber) bei einer angenommenen Verweildauer von zehn Jahren, ein zusätzlicher Nettonutzen, im Vergleich zu einer Zufallsauswahl, von 275.088,53 Euro.

Dieses Ergebnis zeigt, dass der Nutzen eines Verfahrens proportional zur Validität des Prädiktors steht. Je höher die Validität des Prädiktors, desto höher demnach auch der Nutzenzuwachs. Andererseits zeigt das Ergebnis ebenso, dass auch Prädiktoren mit niedrigeren Validitätskoeffizienten einen beachtlichen Nutzen im Vergleich zur Zufallsauswahl erbringen können. Eine Beurteilung psychologischer Eignungsdiagnostik sollte sich somit nicht nur einzig an der Höhe der Validitätskoeffizienten orientieren, sondern

auch auf die Wirkung eignungsdiagnostischer Verfahren auf Personalentscheidungen in Termini von Nutzen und Kosten dieser ausdehnen.
Anhand der Nutzenanalyse wird es somit möglich, den Nutzen jeder Entscheidungsalternative zwischen konkurrierenden Personalprogrammen als gewichtete Kombination der Teilnutzen verschiedener Kriterien und Konsequenzen der jeweiligen Entscheidungsalternative zu bestimmen (Funke & Barthel, 1995).

Erneut muss kritisch betrachtet werden, dass es für die Datenanalysen nicht möglich war, die Originalergebnisse der Testverfahren zu bekommen, sondern nur die erreichten Endergebnisse in einzelnen Testbausteinen sowie bei der Erhebung des Kriteriums zu Grunde lagen. Die erreichte Note bzw. Punktzahl der Teilnehmer wurde durch die die Prüfung abnehmende Fachhochschule übermittelt. Aus methodischen Überlegungen, wäre es jedoch wünschenswert und sinnvoll gewesen, wenn für die vorliegende Untersuchung die Berechnung sowohl der prädiktiven Validität wie auch die Analyse der Nutzen und Kosten auf Grundlage von Rohwerten hätte errechnet werden können. Hier gelten gleiche kritische Anmerkungen, wie schon unter Kapitel 4.7 umfassend dargestellt. Analog kritisch und an gleicher Stelle diskutiert, verhält es sich mit den Auswirkungen der fraglichen Reliabilität von Prädiktor und Kriterium. Insbesondere die Kritik an den Gütemaßen der Schulnoten ist auf die Notenvergabe bei der Zwischenprüfung an der Fachhochschule zu übertragen.

Wie in praxi nicht unüblich, basiert das in dieser Arbeit zugrunde liegende und untersuchte Auswahlverfahren nicht auf der Grundlage empirisch ermittelter Anforderungen und Anforderungsanalysen. Es wurde vielmehr nach inhaltlichen Überlegungen zusammengestellt. Die Testteile Lückendiktat, Bericht, Sporttest, Vorstellungsgespräch und Rundgespräch sind in der Behörde eigens entwickelt worden. Der Sporttest wurde dabei nach den bestehenden Richtlinien bundesweit festgelegt und intern normiert. Als standardisierte Prädiktoren mit entsprechenden Normen und allgemeinen Gütekriterien gelten demnach nur die beiden kognitiven Leistungstests (kulturfair und nicht kulturfair). Es ist daher nicht geklärt, inwieweit die einzelnen Testteile als Prädiktoren überhaupt eine Vorhersageleistung bezüglich der beruflichen Bewährung (Kriterium) dieser speziellen Bewerberschaft leisten können. Fraglich ist außerdem, inwieweit die aufgrund ihres Gesamtergebnisses im Einstellungsverfahren ausgewählten und eingestellten Bewerber den Anforderungen überhaupt entsprechen können, die sich im Laufe ihres Studiums und ihrer beruflichen Tätigkeit ergeben. Die hinter den Prädiktoren vermeintlich stehenden Konstrukte und deren

Operationalisierungen durch die Testbausteine im Auswahlverfahren sollen jedoch in dieser Untersuchung nicht weiter differenziert werden, so dass auf die fragliche Validität der Testbausteine als Prädiktoren nicht weiter eingegangen wird. Dies ist aber notwendigerweise ein dringliches zukünftiges Ziel einer an der DIN 33430 orientierten Eignungsfeststellung.

Eine besondere Komponente, deren Einfluss in diesem Zusammenhang zwar nicht geklärt, jedoch zumindest angesprochen werden sollte, ist die Frage nach der Validität des Kriteriums „Endergebnis im Auswahlverfahren“. Die Wahl eines repräsentativen Validitätskriteriums gehört zu den schwierigsten Aufgaben der Testentwicklung, zu deren Enderfolg sie entscheidend beiträgt (Lienert, 1989). Prädiktoren stellen eignungsdiagnostische Verfahren dar, die Persönlichkeitsmerkmale beliebiger Art sowie weitere Variablen zur Prognose des Berufserfolges operationalisieren und der Vorhersage des Kriteriums dienen. Kriterien dagegen sind die Übersetzung der Zielvorstellung eines Auswahlverfahrens in Verhaltensdimensionen. Bei den Kriterien handelt es sich nach Hossiep (1995) nicht wie bei den eignungsdiagnostischen Verfahren um Konstruktionen, die an der Wirklichkeit geprüft werden, sondern um die Wirklichkeit selbst. Überprüft werden sollte jedoch die Frage, ob die gewählten Kriterien für die Fragestellung wichtig und relevant sind. Nur mittels einer nach wissenschaftlichen Erkenntnissen folgenden Anforderungsanalyse wird es möglich sein, die Anforderungen des Polizeiberufes konkret zu erfassen, um sie dann in geeignete Prädiktoren zur Eignungsfeststellung ab zu gleichen. Diese Basis fehlt jedoch zurzeit und es kann nur angenommen werden, dass die Ausbildung (hier abgebildet als Prüfungsleistung) den Anforderungen gerecht wird.

Unter streng methodischen Gesichtspunkten muss die Beurteilung eines eignungsdiagnostischen Auswahlverfahrens alle Bewerber annehmen und testen, um unverfälschte Beziehungen zwischen Prädiktoren und Kriterien zu berechnen (Färber, 1995). Die in dieser Untersuchung verwendete Validitätsstichprobe stellt jedoch eine einseitige Auswahl von Bewerbern dar. Die Bewerber für den gehobenen Polizeivollzugsdienst müssen zum einen das Auswahlverfahren erfolgreich abgeschlossen haben, als Anwärter eingestellt werden, für die Zwischenprüfung zugelassen worden sein und diese auch abgelegt haben. Herausfallen müssen neben denjenigen, welche das Auswahlverfahren nicht bestanden haben, diejenigen, welche zwar das Eignungsfeststellungsverfahren erfolgreich absolviert haben und in den Polizeivollzugsdienst eingestellt wurden, jedoch ihre Stelle nicht angetreten haben (u.a. aufgrund eines Studienplatz in einem anderen Bundesland, einer Entscheidung für ein

„polizeifremdes“ Studium, etc.). Außerdem fallen die Daten der Bewerber heraus, welche frühzeitig, d.h. vor der Zwischenprüfung, ihr Studium beenden bzw. beenden mussten, sei es aufgrund zu schwacher Leistungen, Unzufriedenheit oder „Nichtpassung“ mit dem ausgeübten Beruf. Somit reduziert sich zwangsweise der für die statistische Analyse in Betracht kommende Datensatz auf eine einseitig selegierte Stichprobe. Unberücksichtigt bleiben somit auch die Daten jener Teilnehmer, die aufgrund nicht ausreichender Leistungen aus dem Studium ausschieden, obwohl sie letztendlich zu den „falsch positiven“ hätten zählen müssen.

Eine Randomisierung, bspw. hinsichtlich der Herkunftsbundesländer der Bewerber und der Nationalität, war nicht leistbar. Nach Czienskowski (1996) muss allerdings in vielen Untersuchungen auf eine Randomisierung verzichtet werden, da es sich um Unterschiede zwischen „natürlichen“ Gruppen handelt. Da es nicht möglich ist, die Stichprobe aus zufällig ausgewählten Probanden zusammenzusetzen, kann es zu Einschränkungen der Repräsentativität der Stichprobe kommen.
Folglich müssen die erhobenen Daten zwangsweise einer systematischen Selektion unterliegen. Durch die Varianzeinschränkung können systematische Unterschätzungen der Zusammenhangsmaße auftreten. In solchen Fällen ist es, mittels der in Kapitel 5.6 bereits beschriebenen Selektionskorrektur jedoch möglich, mit einiger Vorsicht einen Repräsentativschluss auf die Validität des Testes der Bewerberpopulation zu ziehen.

Bei der Bewertung der korrelationsstatistischen Ergebnisse auf Grundlage der klassischen Testtheorie, bei der allein die Höhe der Validitätskoeffizienten bewertet wird, sind die berechneten Validitätskoeffizienten der Prädiktoren zu niedrig und haben als einzelne Prädiktoren eher geringen praktischen Wert. Allein der Prädiktor „Bericht Inhalt“ und das „Gesamttestergebnis im Auswahlverfahren“ erreichen signifikante Ergebnisse und stehen in einem nachweisbaren statistischen Zusammenhang mit dem Kriterium. Die Ergebnisse der regressionsanalytischen Berechnungen verifizieren dieses Bild, wobei hier noch der Prädiktor „Kulturfairer Leistungstest“ in Kombination aller verwendeter Prädiktoren einen signifikanten Vorhersagebeitrag zum Kriterium leistet. Bei der Analyse der schrittweisen Regression, bei der dann die Testbausteine „Bericht Inhalt“, „Kulturfairer Leistungstest“ und das „Rundgespräch“ einen signifikanten Beitrag zur Nützlichkeit der Vorhersageleistung beitragen, liegt die Vermutung einer reziproken Suppression zwischen den Variablen „Kulturfairer Leistungstest“ und „Rundgespräch“ nahe.

Die von den drei Prädiktoren errechnete multiple Korrelation steht im Vergleich zu anderen Forschungsergebnissen jedoch im höheren Durchschnittsbereich. Bei zusätzlicher Berücksichtigung der Ergebnisse der Kosten–Nutzen Analyse und der dieses Ergebnis beeinflussenden Faktoren, wie z.B. problematische Transformation und kleine Stichprobe, erweisen sich diese Prädiktoren durchaus von praktischem Wert zur Vorhersage des Abschneidens in der Zwischenprüfung.

6. Fragestellung 3: Vergleich der Ergebnisse deutscher und nicht deutscher Teilnehmer

6.1 Einführung in die Fragestellung 3

Zur Beantwortung der Hypothesen für die Fragestellung 3 wurden die Ergebnisse der deutschen und der nicht deutschen Bewerber hinsichtlich ihrer erreichten Resultate in den beiden Anwendung findenden kognitiven Leistungstests im Auswahlverfahren für den gehobenen Polizeivollzugsdienst verglichen. Hier sollte die Frage beantwortet werden, inwieweit sich die Ergebnisse der ausländischen und die der deutschen Bewerber im kulturfairen Testverfahren sowie im nicht kulturfairen Testverfahren voneinander unterscheiden und inwieweit die Einführung eines kulturfairen Testverfahrens dazu beitragen konnte, evtl. vorhandene kulturell bedingte Benachteiligungen ausländischer Bewerber auszugleichen.

Neben der wichtigsten Aufgabe eines Auswahlverfahrens, die für die spätere Berufsausübung geeigneten Bewerber auszuwählen und jene, die nicht geeignet erscheinen, abzulehnen, war es ein weiteres Ziel im Rahmen der Beurteilung des Auswahlverfahrens zu hinterfragen, ob durch die Einführung eines kulturfairen kognitiven Leistungstest im Vergleich zu einem nicht kulturfairen Leistungstest, vermehrt für den gehobenen Polizeivollzugsdienst geeignete ausländische Mitbewerber eingestellt werden konnten. Grundvoraussetzung für diesen Bewerberkreis sind, trotz anderer kultureller Herkunft und Muttersprache, die gleichen Anforderungen in Ausbildung und Berufsausübung wie für ihre deutschen Kollegen. Dabei blieb jedoch zu beachten, dass Unterschiede bei erreichten Ergebnissen in den einzelnen Testbausteinen nicht zwangsweise auf ein geringeres Leistungs- und Intelligenzniveau

zurückgeführt werden müssen. Die Verwendung nicht sprach- und kulturfairer Testverfahren kann zur Benachteiligung dieser Bewerber führen und sie um die Chance bringen, trotz einer gegebenen generellen Eignung für den gehobenen Polizeivollzugsdienst, aufgrund geringerer Gesamtpunkte im Rahmen der Bestenauslese, nicht ausgewählt zu werden. Um diesen kulturell bedingten Effekt einer möglichen Benachteiligung der ausländischen Bewerber entgegen zu wirken, wurde ein kulturfairer kognitiver Leistungstest in die Testbatterie für die Auswahl zukünftiger Polizeibeamter aufgenommen. Dieser soll ermöglichen, die kognitive Leistungsfähigkeit der nicht deutschen Bewerber unabhängig vom Einfluss durch Muttersprache und Herkunft zu bewerten und ihnen somit eine größere Chance auf Einstellung zu geben.

Mittels eines Vergleichs der Leistungen deutscher und nicht deutscher Teilnehmer sollte die Frage beantwortet werden, welche Auswirkungen die Einführung des kulturfairen kognitiven Leistungstest im genannten Sinne hatte. Hierzu wurden die erreichten Testergebnisse deutscher und nicht deutscher Teilnehmer am Auswahlverfahren in einem kulturfairen und einem nicht kulturfairen kognitiven Leistungstest gegenüber gestellt.

6.2 Versuchsplan für Fragestellung 3

In Tabelle 6.2-1 wird der Versuchsplan für die Fragestellung 3 dargestellt.

Tab. 6.2-1: Versuchsplan für die Fragestellung 3

Fragestellung 3: Vergleich der Ergebnisse deutscher und nicht-deutscher Teilnehmer	
TEILSTICHPROBE	**UNTERSUCHUNG**
Teilnehmer am kulturfairen und nicht kulturfairen kognitiven Leistungstest im Auswahlverfahren für die Einstellung in den gehobenen Polizeivollzugsdienstes (Einstellung Oktober 1999)	Vergleich der Testergebnisse deutscher und nicht deutscher Bewerber in einem kulturfairen und einem nicht kulturfairen kognitiven Leistungstest

6.3 Durchführung

Zu den Rahmenbedingungen der Durchführung sei an dieser Stelle wieder auf Kapitel 4.5 verwiesen. Die Datenerhebung fand im gleichen Zeitraum (September 1998 bis April 2001) statt. Zeitgleich zur Erhebung der Daten für die Fragestellung 1 begann die gesonderte Aufnahme der Testergebnisse deutscher und nicht deutscher Teilnehmer.

6.3.1 Beschreibung der Stichprobe

Alle Teilnehmer am Auswahlverfahren für die Einstellung in den gehobenen Polizeivollzugsdienst (Gesamtstichprobe) bildeten auch die Stichprobe zur Beantwortung der Fragestellung 3. Entsprechend ist die Beschreibung im Kapitel 4.5.1 wiedergegeben. In den folgenden Tabellen 6.3-1 – 6.3-4 ist ein Vergleich zwischen deutschen und nicht deutschen Teilnehmern dargestellt.

Tab. 6.3-1: Zusammensetzung der Gesamtstichprobe getrennt nach Nationalität: Geschlecht

GESCHLECHT n = 735	HÄUFIGKEIT		PROZENT	
	Deutsch	Nicht deutsch	Deutsch	Nicht deutsch
Männlich	358	30	51,8	68,2
Weiblich	333	14	48,2	31,8
Gesamt	691	44	100	100

Tab. 6.3-2: Zusammensetzung der Gesamtstichprobe getrennt nach Nationalität: Schulbildung

SCHULBILDUNG n = 735	HÄUFIGKEIT		PROZENT	
	Deutsch	Nicht deutsch	Deutsch	Nicht deutsch
Fachhochschulreife	361	18	52,2	40,9
Abitur	222	13	32,1	29,5
sonstige *	108	13	15,7	29,5
Gesamt	691	44	100	100

* sonstiger, als gleichwertig anerkannter Bildungsabschluss

Tab. 6.3-3: Zusammensetzung der Gesamtstichprobe getrennt nach Nationalität: Alter der Bewerber

ALTER n = 735	HÄUFIGKEIT		PROZENT	
	Deutsch	Nicht deutsch	Deutsch	Nicht deutsch
17 – 20 Jahre	465	23	67,3	52,2
21 – 25 Jahre	123	11	17,7	25,0
26 – 30 Jahre	73	5	10,7	11,4
älter als 30 Jahre	30	5	4,3	11,4
Gesamt	691	44	100	100

Tab. 6.3-4: Zusammensetzung der Gesamtstichprobe getrennt nach Nationalität: Bewerbung für Dienstzweig

DIENSTZWEIG n = 735	HÄUFIGKEIT		PROZENT	
	Deutsch	Nicht deutsch	Deutsch	Nicht deutsch
Schutzpolizei	483	27	69,9	61,4
Kriminalpolizei	157	14	22,7	31,8
Wasserschutzpolizei	51	3	7,4	6,8
Gesamt	691	44	100	100

In der Tabelle 6.3-5 ist die Verteilung der Kennwerte der Schulnoten der Bewerber in den erhobenen Fächern getrennt nach Nationalität zusammen gefasst.

Tab. 6.3-5: Kennwerte der Schulnoten der Bewerber getrennt nach Nationalität

Schulnoten deutscher und nicht deutscher Teilnehmer im Auswahlverfahren						
n = 735	DEUTSCH			NICHT DEUTSCH		
	N	M_x	S_x	N	M_x	S_x
Deutsch	646	2,95	.75	36	3,31	.79
Mathematik	641	3,20	.98	35	3,63	.91
Englisch	635	3,15	.84	36	3,44	.94
Geschichte et al.	637	2,70	.77	40	2,88	.72
Sport	596	1,87	.75	32	2,16	.88

Wie aus der Tabelle 6.3-6 zu entnehmen, haben von den insgesamt 735 Probanden der

Stichprobe 721 (681 deutsche und 40 nicht deutsche Teilnehmer) am kulturfairen Testverfahren teilnehmen können. Aufgrund einer nicht ausreichenden Leistung – nicht mindestens vier von neun möglichen Punkten - hätte erfahrungsgemäß ein größerer Teil der Bewerber im Rahmen der sequentiellen Teststrategie aus dem Verfahren genommen werden müssen. Um jedoch auf eine möglichst große Stichprobe für die vergleichende Untersuchung zurückgreifen zu können, wurde das erste Feedback der durch die Teilnehmer erzielten Leistungen in den ersten drei Testbausteinen (Diktat, Bericht und kulturfairer Leistungstest) erst nach Durchführung des dritten Bausteins gegeben. Nur in besonderen Einzelfällen (n = 13) wurden die Probanden umgehend aus dem Verfahren genommen. Zwar wäre es für die Durchführung und Aussagekraft der Untersuchung wünschenswert gewesen, über eine vergleichbare Stichprobengröße auch für den nicht kulturfairen Leistungstest verfügen zu können, dies war jedoch aus verschiedenen Gründen nicht opportun. Zum einem sprachen sowohl ethische wie auch organisatorische Gründe dagegen, alle Bewerber am Verfahren teilnehmen zu lassen, auch wenn schon am ersten Testtag entschieden gewesen wäre, dass sie aufgrund einzelner erreichter Ergebnisse keine Möglichkeit zur Einstellung hätten bekommen können. Zum anderen wären die der Organisation und den Bewerbern durch die aussichtlose Teilnahme an drei unterschiedlichen Testtagen entstandenen Kosten nicht vertretbar gewesen. Insofern ist bei der Durchführung der Untersuchung sowie Bewertung der Ergebnisse zu berücksichtigen, dass der Anteil der nicht deutschen Teilnehmer an der Datenerhebung des nicht kulturfairen kognitiven Leistungstest aufgrund der Auswahlstrategie nur n = 13 umfasst. In Tabelle 6.3-6 findet sich eine Übersicht der Anzahl deutscher und nicht deutscher Teilnehmer an den unterschiedlichen Testbausteinen des Auswahlverfahrens für den gehobenen Polizeivollzugsdienst. Aufgrund der Durchführung der Untersuchung im Bundesland Hamburg war zu erwarten, dass sich die Bewerbungen nicht gleichmäßig und repräsentativ auf alle Bundesländer verteilen, so dass eine Einschränkung der Repräsentativität der Stichprobe gegeben ist.

Eingegangen werden sollte an dieser Stelle auf die bestehende Definition jener Teilnehmer, die als nicht deutsch dargestellt werden. Nach den gegebenen Einstellungsvorgaben, werden hierunter alle Bewerber verstanden, die mit einer zweiten Muttersprache bzw. einer nicht deutschen Muttersprache aufgewachsen sind. Dabei wird nicht unterschieden, ob ein Bewerber etwa in Deutschland aufwuchs und die Muttersprache Deutsch im Elternhaus im Vordergrund stand oder er erst zu einem späteren Zeitpunkt nach Deutschland eingewandert ist und vor Ort die deutsche Sprache hat erlernen müssen. Der überwiegende Anteil der

Bewerber ist in Deutschland geboren und hat deutsche Schulen besucht. Dabei bestand traditionell jedoch häufig eine enge Bindung an die Herkunftstradition und damit einhergehend auch der Gebrauch der Muttersprache in der Familie, da oft nicht alle Familienmitglieder deutsch sprechen.

Tab. 6.3-6: Anzahl deutscher und nicht deutscher Teilnehmer in den unterschiedlichen Testbausteinen des Auswahlverfahrens

Deutsche und nicht deutsche Teilnehmer im Auswahlverfahren						
n = 735	**DEUTSCH**		**NICHT DEUTSCH**		**GESAMT**	
	Häufigkeit	**Prozent***	**Häufigkeit**	**Prozent**	**Häufigkeit**	**Prozent**
Diktat	691	94,0	44	6,0	735	100,0
Bericht	691	94,0	43	5,9	734	99,9
Kulturfairer Leistungstest	681	92,7	40	5,4	721	98,0
Sporttest	637	86,7	26	3,5	663	90,2
Nicht kulturfairer Leistungstest	422	57,4	13	1,8	435	59,2
Vorstellungs- und Rundgespräch	201	27,3	6	0,8	207	28,2
Im Gesamtverfahren erfolgreich	162	22,0	5	0,7	167	22,7

* die Prozentangaben beziehen sich auf die Gesamtstichprobe

Die sich für die Stichprobe ergebenden weiteren Kennwerte hinsichtlich der erreichten Ergebnisse im kulturfairen sowie nicht kulturfairen Leistungstest sind in Tabelle 6.3-7 dargestellt.

Tab. 6.3-7: Kennwerte deutscher und nicht deutscher Teilnehmer im kognitiven Leistungstest

Kognitiver Leistungstest n = 735	**N**		**M_x**		**S_x**	
	DEUTSCH	**NICHT DEUTSCH**	**DEUTSCH**	**NICHT DEUTSCH**	**DEUTSCH**	**NICHT DEUTSCH**
kulturfairer kognitiver Leistungstest	681	40	7,17	6,20	1,60	1,68
nicht kulturfairer kognitiver Leistungstest	422	13	19,0	16,85	4,74	3,44

6.3.2 Ablauf der Untersuchung

Mit Beginn der Prüfungsphase für die Einstellung in den gehobenen Polizeivollzugsdienst zum Oktober 1999 wurden regelmäßig nach Durchführung des Auswahlverfahrens ab Oktober 1998 die in den einzelnen Testbausteinen erreichten Ergebnisse der Bewerber aufgenommen. Der sequentielle Aufbau des Auswahlverfahrens führte dazu, dass nicht von

allen Teilnehmern die Ergebnisse in beiden kognitiven Leistungstests erhoben werden konnten und sich die Stichprobe reduzierte. Auch wenn grundsätzlich größere Stichproben zu genaueren und differenzierteren Ergebnissen führen als kleinere (Bortz, 1993), konnte aus o.g. Gründen die Erhebung aller Ergebnisse nicht erreicht werden. Nach Beendigung der Datenerhebung Ende Mai 1998 wurden die Daten statistisch ausgewertet. Die Analysen sollten darüber Aufschluss geben, ob der Einsatz eines kulturfairen Testverfahrens im Kontext der Reduzierung einer möglichen Leistungsminderung durch eine andere Herkunft und Muttersprache der ausländischen Mitbewerber, bei der Auswahl für den gehobenen Polizeivollzugsdienst angemessen und effizient ist. Mittels eines t-Testes für unabhängige Stichproben (s.a. Bortz, 1993) wurden die Ergebnisse der deutschen und nicht deutschen Teilnehmer in den beiden kognitiven Leistungstests miteinander verglichen und auf Signifikanz getestet.

6.4 Ergebnisdarstellung

6.4.1 Darstellung der Ergebnisse im Vergleich deutscher und nicht deutscher Teilnehmer

In Tabelle 6.4-1 sind die Gruppenstatistiken deutscher und nicht deutscher Teilnehmer im kulturfairen kognitiven Leistungstest wiedergegeben.

Tab. 6.4-1: Gruppenstatistiken deutscher und nicht deutscher Teilnehmer im kulturfairen kognitiven Leistungstest

Gruppenstatistiken n = 721	**NATIONALITÄT**	**N**	**MX**	**S_x**	**STANDARDFEHLER MITTELWERT**
KULTURFAIRER KOGNITIVER LEISTUNGSTEST	Deutsch	681	7,17	1,60	6,13E-02
	Nicht deutsch	40	6,20	1,68	.27

In Tabelle 6.4-2 ist das Ergebnis des Mittelwertvergleiches deutscher und nicht deutscher Teilnehmer im kulturfairen kognitiven Leistungstest abgebildet.

Tab. 6.4-2: Test für unabhängige Stichproben im kulturfairen kognitiven Leistungstest

TEST BEI UNANHÄNGIGEN STICHPROBEN										
		LEVENE-TEST DER VARIANZENGLEICHHEIT		T-TEST FÜR DIE MITTELWERTVERGL EICHE						
		F	SIG.	T	DF	SIG. 2-SEITIG	MITTLERE DIFFERENZ	STANDARDF. DIFFERENZ	95% KONFIDENZINTERVALL DIFFERENZ	
									Untere	Obere
KULTURFAIRER KOGNITIVER LEISTUNGSTEST	Varianzen sind gleich	.001	.980	3,732	719	.000	.97	.26	.46	1,49
	Varianzen sind nicht gleich			3,565	43,33	.001	.97	.27	.42	1,52

In der Tabelle 6.4-3 sind die Gruppenstatistiken deutscher und nicht deutscher Teilnehmer im nicht kulturfairen kognitiven Leistungstest wiedergegeben.

Tab. 6.4-3: Gruppenstatistiken deutscher und nicht deutscher Teilnehmer im nicht kulturfairen Leistungstest

Gruppenstatistiken n = 435	NATIONALITÄT	N	MX	S_x	STANDARDFEHLER MITTELWERT
NICHT KULTURFAIRER KOGNITIVER LEISTUNGSTEST	Deutsch	422	19,00	4,74	.23
	Nicht deutsch	13	16,85	3,44	.95

In Tabelle 6.4-4 ist das Ergebnis des Mittelwertvergleiches deutscher und nicht deutscher Teilnehmer im nicht kulturfairen kognitiven Leistungstest abgebildet.

Tab. 6.4-4: Test für unabhängige Stichproben im nicht kulturfairen kognitiven Leistungstest

TEST BEI UNABHÄNGIGEN STICHPROBEN										
		LEVENE-TEST DER VARIANZENGLEICHHEIT		T-TEST FÜR DIE MITTELWERTVERGL EICHE						
		F	SIG.	T	DF	SIG. 2-SEITIG	MITTLERE DIFFERENZ	STANDARDF. DIFFERENZ	95% KONFIDENZINTERVALL DIFFERENZ	
									Untere	Obere
NICHT KULTURFAIRER KOGNITIVER LEISTUNGSTEST	Varianzen sind gleich	2,564	.110	1,629	433	.104	2,16	1,33	-.45	4,76
	Varianzen sind nicht gleich			2,201	13,446	.046	2,16	.98	.05	4,27

6.4.2 Beantwortung der Hypothesen

Die in Kapitel 3.3.3.1 formulierten Hypothesen zur Analyse der Leistungsunterschiede zwischen deutschen und nicht deutschen Teilnehmer in einem kulturfairen sowie einem nicht kulturfairen kognitiven Leistungstest werden wie folgt beantwortet:
Wie aus der Tabelle 6.4-2 zu entnehmen, unterscheiden sich die Mittelwerte der beiden Gruppen (deutsche und nicht deutsche Teilnehmer) im kulturfairen Leistungstest, bei Annahme der Varianzhomogenität, auf dem 1% Niveau signifikant voneinander. Die deutschen Bewerber erreichen eine höhere Punktzahl als die nicht deutschen ($M_{deutsch}$ = 7,17; $M_{nicht\ deutsch}$= 6,20).
Bei der Interpretation der Ergebnisse in Tabelle 6.4-4 ist zu berücksichtigen, dass von der Heterogenität der Varianzen auszugehen ist, da per Konvention als Anpassungsstrategie zur Erhaltung der H_0 α = 20% getestet wird. Bei Anpassung der Freiheitsgrade für heterogene Varianzen ergibt sich somit auch ein signifikanter Unterschied zwischen den Mittelwerten der deutschen ($M_{deutsch}$ = 19,00) und nicht deutschen ($M_{nicht\ deutsch}$ = 16,85) Bewerber im nicht kulturfairen kognitiven Leistungstest (p= .046). Entsprechend der Hypothese 1 unter 3.3.3.2 ist festzustellen, dass sich, parametrisch getestet, signifikante Leistungsunterschiede zwischen deutschen und nicht deutschen Teilenehmern im nicht kulturfairen Leistungstest zeigen.
Angesichts der geringen Stichprobengröße (n = 13) bei den nicht deutschen Bewerbern und der gegebenen Varianzheterogenität konnte jedoch ein aus diesem Grund zusätzlich gerechneter non parametrischer Test (Mann-Whitney U- Test) das Ergebnis nicht weiter absichern (Mann-Whitney-U = 2080; Z = -1,48; Signifikanz (2-seitig) = 0,13).

In der Tabelle 6.4-4 ist dargestellt, dass sich die Ergebnisse der deutschen und der nicht deutschen Bewerber bei Annahme der Varianzhomogenität im nicht kulturfairen kognitiven Leistungstest nicht signifikant voneinander unterscheiden. Es zeigen sich somit keine Leistungsunterschiede zwischen den deutschen und den nicht deutschen Bewerbern ($M_{deutsch}$ = 19,00; $M_{nicht\ deutsch}$= 16,85).

6.5 Interpretation der Ergebnisse

Ziel der Untersuchung 3 war es, zu hinterfragen, inwieweit die Einführung eines kulturfairen Leistungstests dazu beitragen kann, vermehrt auch ausländische Bewerber in den gehobenen Polizeivollzugsdienst einzustellen. Generell gelten für diesen Bewerberkreis die gleichen Anforderungen wie für ihre deutschen Mitbewerber. Allerdings wurde die Möglichkeit

berücksichtigt, dass durch eine andere Muttersprache und kulturelle Herkunft auf Seiten der nicht deutschen Bewerber geringere Leistungen in sprachgebundenen Testverfahren auftreten können. Diese sind möglicherweise Ergebnis unterschiedlicher Herkunft und fehlender Sozialisierung und nicht zwangsweise Ursache eines geringeren kognitiven Leistungsniveaus. Um diesen möglichen Effekt aufzufangen, wurde ein kulturfairer kognitiver Leistungstest in die Testbatterie für die Auswahl der zukünftigen Polizeivollzugsbeamten aufgenommen. Auf dessen Ergebnisgrundlage soll das kognitive Leistungsniveau, unabhängig von Herkunft und Muttersprache, beurteilt werden.

In Deutschland besteht in vielen ansässigen nicht deutschen Familien eine enge kulturelle Bindung an die Herkunftstradition und damit einhergehend der Gebrauch der Muttersprache innerhalb der Familie. Durch diese enge Tradition und in vielen Fällen auch eine mangelnde gesellschaftliche Integration, kann sich aufgrund dessen eigenständige Gesellschaften bzw. Subkulturen in einer Großgesellschaft mit eigener Sprache und kulturellem Hintergrund bilden. Diese Subkulturenbildung kann besonders im polizeilichen Kontext zu Problemen führen. Durch den Gebrauch unterschiedlicher Sprachen und Unkenntnis über den kulturellen Hintergrund, den damit verbundenen Ritualen und Denkansätzen der nicht deutschen Mitbürger, kann es zu Missverständnissen zwischen der Polizei und des nicht deutschen Mitbürgers kommen. Derartige Irrtümer könnten dadurch aufgefangen werden, dass vermehrt Polizeibeamte nicht deutscher Herkunft eingestellt werden, die in derartigen Situationen eine womöglich adäquatere Einschätzung der Situation leisten und sprachliche Hindernisse überbrücken könnten. Ein weiterer positiver Effekt wäre eine höhere Akzeptanz der Beamten seitens der nicht deutschen Mitbürger.

Neben dem zusätzlich eingeführten sprachfreien und kulturfairen Leistungstest wurde allerdings weiterhin auch ein sprachgebundener nicht kulturfairer Leistungstest zur differenzierteren Bewertung einzelner kognitiver Fähigkeiten zu einem späteren Zeitpunkt im Rahmen der sequentiellen Auswahlstrategie angewandt. Aufgrund dieser Rahmenbedingungen sollten durch die Untersuchung 3, die Ergebnisse der deutschen und der nicht deutschen Bewerber in den beiden im Auswahlverfahren verwendeten kognitiven Leistungstests für die Einstellung in den gehobenen Polizeivollzugsdienst, verglichen werden. Es sollte die Frage beantwortet werden, ob sich die Ergebnisse der deutschen Bewerber und die der ausländischen Bewerber im kulturfairen Testverfahren sowie im nicht kulturfairen Testverfahren voneinander unterscheiden und inwieweit die Einführung eines kulturfairen

Testverfahrens unter diesen Umständen dazu beitragen konnte, evtl. vorhandene kulturell bedingte Benachteiligungen ausländischen Bewerber auszugleichen.

Die Ergebnisse des t-Testes für unabhängige Stichproben zur Beantwortung der Fragestellung, ob und inwieweit sich die Leistungen der deutschen Bewerber von den Leistungen der nicht deutschen Bewerber im kulturfairen Leistungstest und im nicht kulturfairen Leistungstest voneinander unterscheiden, erbringt ein, auf den ersten Blick, überraschendes Ergebnis. Schon im kulturfairen Testverfahren, mit dem Ziel verwandt, mögliche sprachliche und kulturelle Benachteiligungen nicht deutscher Teilnehmer aufzufangen, erreichen die nicht deutschen Teilnehmer eine niedrigere Punktzahl als die deutschen Teilnehmer. Der von den deutschen Teilnehmern erreichte Mittelwert $\bar{x}_{deutsch}$ = 7,17 Punkten stellt im Vergleich zu dem Mittelwert der nicht deutschen Teilnehmer von $\bar{x}_{nicht\ deutsch}$ = 6,20 Punkten ein signifikant besseres Ergebnis dar. Es zeigen sich somit statistisch nachweisbare Leistungsunterschiede zwischen den beiden Teilstichproben zu Gunsten der deutschen Teilnehmer.
Entsprechend der unter Abschnitt 3.3.3.2 formulierten Hypothese, steht das Ergebnis der Analyse von Mittelwertsunterschieden bezüglich des Ergebnisses im nicht kulturfairen Leistungstest. Auch hier ergaben sich zumindest parametrisch nachweisbare statistische Unterschiede in den Ergebnissen der deutschen und der nicht deutschen Teilnehmer. Die deutschen Teilnehmer erreichen eine durchschnittliche Punktzahl von $\bar{x}_{deutsch}$ = 19,0 Punkten, während die nicht deutschen Teilnehmer eine mittlere Punktzahl von $\bar{x}_{nicht\ deutsch}$ = 16,85 Punkten erreichen und somit nachweisbare Leistungsunterschiede zwischen den beiden Teilnehmergruppen bestehen. Ein angesichts der geringen Stichprobengröße (n = 13) bei den nicht deutschen Bewerbern und der gegebenen Varianzheterogenität gerechneter non parametrischer Test konnte jedoch das Ergebnis nicht weiter absichern.

Diese Ergebnisse sind insofern überraschend, da der Einsatz des kulturfairen Leistungstests als nicht sprachgebundenes Intelligenzverfahren die vermuteten sprachlichen und kulturellen Defizite auffangen und sie hinsichtlich der Anforderungen mit den deutschen Bewerbern gleich stellen sollte. Grundgedanke dieses Vorgehens war, dass die Verwendung nicht sprach- und kulturfairer Testverfahren zu einer möglichen Benachteiligung der nicht deutschen Bewerber führen könnte und sie trotz einer gegebenen generellen Eignung für den gehobenen Polizeivollzugsdienst, aufgrund zu geringerer Gesamtpunkte aus dem weiteren Verfahren ausscheiden.

Die Erfahrung der durchführenden Dienstelle bei Annahme der Bewerbungen zeigte, dass ein überwiegender Anteil der Bewerber in Deutschland gebürtig ist und aufgrund dessen auch deutschsprachige Schulen besucht hat. Die möglicherweise bestehenden kulturell bedingten Subkulturen sollte das Auswahlverfahren auffangen. Das Ziel des Einsatzes des kulturfairen Leistungstests für diese Bevölkerungsgruppe konnte mit dem Ergebnis nicht erreicht werden. Als Erklärung könnte angenommen werden, dass sich die Teilstichprobe der nicht deutschen Bewerber, die am Auswahlverfahren für den gehobenen Polizeivollzugsdienst de facto teilgenommen haben, von der a priori definierten Teilstichprobe unterscheidet.
Viele der nicht deutschen Bewerber, die am Auswahlverfahren teilgenommen haben, sind im Sinne der vorher abgegebenen Definition keine „wirklichen“ Ausländer. Die meisten nicht deutschen Teilnehmer sind in Deutschland geboren, haben ihre Ausbildung an deutschen Schulen erworben und sind aufgrund dessen sozialisiert und integriert.

Das Ergebnis der statistisch signifikanten Unterscheidung in beiden Leistungstests zugunsten der deutschen Bewerber spricht für die Annahme, dass die Stichprobe der nicht deutschen Bewerber im Ganzen schlechtere Ergebnisse erzielt hat. In diesem Kontext sollten die Ergebnisse aus Tabelle 6.3-6 Berücksichtigung finden. In dieser wird dargestellt, wie groß der Anteil an jeweils deutschen und nicht deutschen Teilnehmern ist, die an den einzelnen Stufen des sequentiellen Auswahlverfahrens haben teilnehmen können. Hier zeigt sich dass 23,44 % der deutschen im Vergleich zu nur 11,36 % der nicht deutschen Teilnehmer am Auswahlverfahren überhaupt erfolgreich waren. Die letzte Stufe des Auswahlverfahrens (Rundgespräch) haben 29 % der deutschen und nur 13,36 % der nicht deutschen Teilnehmer erreicht. Es sind somit insgesamt 76,56 % der deutschen Teilnehmer am Auswahlverfahren ausgeschieden und im Gegensatz dazu insgesamt 88,64 % der nicht deutschen Teilnehmer.
Die Ergebnisse der Zusammensetzung der Stichprobe unterstützen weiterhin die Annahme, dass die nicht deutschen Teilnehmer insgesamt schlechtere Ergebnisse erzielen als die deutschen Teilnehmer. Weiter zeigt die Tabelle 6.3-2, dass ein größerer Anteil der deutschen Teilnehmer am Auswahlverfahren einen höheren Schulabschluss aufweisen kann als die nicht deutschen Teilnehmer. Von den deutschen Teilnehmer haben 32,1 % das Abitur im Vergleich zu 29,5% der nicht deutschen Teilnehmer; 52,2 % der deutschen Teilnehmer besitzen die Fachhochschulreife, hingegen nur 40,9 % der nicht deutschen Teilnehmer. Die Betrachtung der Kennwerte der Schulnoten hinsichtlich der Zusammensetzung der Stichprobe stützt ebenfalls die Annahme, dass die nicht deutschen Teilnehmer insgesamt geringere Leistungen erzielen als die deutschen Teilnehmer. In allen aufgeführten Schulnoten weisen

die deutschen Bewerber bessere Noten auf als die nicht deutschen Teilnehmer.

Weiter muss bei der Interpretation der Ergebnisse die kleine Stichprobe der nicht deutschen Teilnehmer Berücksichtigung finden. Insgesamt nehmen nur N = 44 nicht deutsche, im Vergleich zu N = 691 deutschen Bewerbern, am gesamten Auswahlverfahren teil. Durch die sequentielle Auswahlstrategie wurde die kleinere Stichprobe der nicht deutschen Teilnehmer relativ weiter reduziert. Im Testbaustein „Kulturfairer Leistungstest“ liegen noch N = 40 Ergebnisse der nicht deutschen Teilnehmer vor, während die Stichprobe im Testbaustein „Nicht Kulturfairer Leistungstest“ bereits auf N = 13 geschrumpft ist. Da der Anteil an geeigneten Teilnehmern in der Stufe „Nicht Kulturfairer Leistungstest“ aufgrund der sequentiellen Auswahlstrategie eher höher ist als in der Stufe des „Kulturfairen Leistungstest“ (selegierte Stichprobe, s.a. auch Kapitel 5.7), trägt dieser Aspekt ebenso dazu bei, dass sich die beiden Teilstichproben in ihren Leistungen nicht signifikant unterscheiden.

Ein weiterer Punkt, welcher die Auswertung der Ergebnisse und deren Interpretation erschwert, ist die vorhandene Heterogenität der beiden Stichproben der deutschen und der der nicht deutschen Teilnehmer. Diese ist nicht unproblematisch, da aufgrund dessen nicht eindeutig nachzuvollziehen ist, wie die Ergebnisse entstanden sind, entweder hauptsächlich durch die uneinheitliche Zusammensetzung der Stichprobe (Stichprobeneffekt) oder aufgrund des unterschiedlichen Leistungsniveaus der Teilnehmer, das sich in den Ergebnissen widerspiegelt. Somit können die vorliegenden Ergebnisse als eine mögliche Tendenz bewertet werden, nicht aber als ein eindeutig empirisch belegbarer Effekt.

Zusammengefasst lässt sich sagen, dass das oben genannte Ziel nicht erreicht werden konnte. Die Ergebnisse sprechen nicht dafür, dass es gelungen ist, vermehrt geeignete Ausländer für den Polizeiberuf zu gewinnen. Es wird deutlich, dass es sich bei den erfolgreichen Bewerbern nicht generell Menschen mit den gewünschten kulturellen Bindungen handelt, sondern eher um in Deutschland geborene und sozialisierte Bewerber und zudem Bewerber, mit schlechteren Zugangsvoraussetzungen (Schulnoten, Schulbildung etc.) als ihre zum Vergleich herangezogenen deutschen Teilnehmer. Da es keine empirischen Belege für die Annahme gibt, das ausländische Bewerber generell kognitiv weniger leistungsfähig sind, ist das Ergebnis im kulturfairen Testverfahren eher ein Hinweis darauf, dass hier eine Population angesprochen wurde, die ein deutlich geringeres kognitives und allgemeines Leistungsniveau besitzt. Bewerber ausländischer Herkunft, scheinen sich weniger für den Beruf des

Polizeibeamten zu interessieren, wenn sie eine entsprechende Bildung besitzen. So ist anzunehmen, dass das vorliegende Ergebnis dafür spricht, dass es bisher nicht ausreichend gelungen ist, geeignete Bewerber, die im politisch gewollten Sinne den kulturellen Hintergrund mitbringen, anzusprechen.

7. Integration der Ergebnisse und Ausblick

Wie schon in Kapitel 2.1 dargestellt, ist die Arbeits- und Anforderungsanalyse die Grundlage der Personalauswahl und Leistungsbeurteilung. Für den Arbeitsplatz und das Aufgabenfeld werden diejenigen Tätigkeiten analysiert, welche für das erfolgreiche Ausführen der Arbeitstätigkeit erforderlich sind. Die tätigkeitsrelevanten und tätigkeitsübergreifenden Anforderungen werden dann in Verfahren umgesetzt, welche diese erfassen und abbilden sollen und somit eine effiziente Personalauswahl erst möglich machen. Die Konstruktion der Prädiktoren und Kriterien sollte demnach, wenn möglich, auf den Ergebnissen einer Anforderungsanalyse basieren. Das in dieser Arbeit zu beurteilende Auswahlverfahren für den gehobenen Polizeivollzugsdienst stützte sich, wie es in praxi durchaus nicht unüblich ist, nicht auf die Ergebnisse einer Arbeitsplatzanalyse, sondern wurde durch erfahrene Mitarbeiter aufgrund inhaltlicher Überlegungen entwickelt. Insofern stellt sich zentral die Frage, ob die im Auswahlverfahren eingesetzten Prädiktoren, die beruflichen Fähigkeiten für den gehobenen Polizeivollzugsdienst valide abbilden und somit eine Vorhersage des beruflichen Erfolges leisten können.

In der vorliegenden Untersuchung 1 war es das Ziel, die prognostische Validität in den Zeugnissen der Bewerber um einen Studienplatz im Rahmen der Ausbildung zum gehobenen Polizeivollzugsdienst zu überprüfen. In einem ersten Teil der Untersuchung wurden hierzu die Schulnoten Deutsch, Englisch, Mathematik, Sport sowie die Note in einem der Fächer Geschichte, Politik oder Wirtschaft als Prädiktoren erhoben. Das Kriterium war das Gesamtergebnis der Teilnehmer im Auswahlverfahren. Im zweiten Teil der Untersuchung 1 wurden erneut die gleichen Schulnoten als Prädiktoren erhoben. Es sollte untersucht werden, welche Aussagekraft Schulnoten hinsichtlich des Kriteriums „Zwischenprüfung an der Fachhochschule“ besitzen. Neben Aspekten der Güte von Schulnoten als potentielle Prädiktoren im Auswahlprozess, sollte auch ihre Rolle unter Abwägung von Kosten-Nutzen Relationen beurteilt werden. Aus diesem Grund wurde die Kosten-Nutzen Funktion mittels

des Brodgen-Cronbach-Gleser-Modells berechnet.
Bei Beantwortung der Hypothesen zeigte sich, dass die errechneten bivariaten Korrelationskoeffizienten zwischen Schulnoten und dem Erfolg im Auswahlverfahren keine zufrieden stellende Höhe erreichen konnten (Deutsch: r = -.12; Mathematik: r = -.16; Englisch: r = -.19; Sport: r = -.04; Durchschnittsnote: r = -.21). Orientiert an den Ergebnissen anderer Studien (siehe auch Kapitel 2.3 und 4) leistete nur die errechnete Durchschnittsnote einen vergleichbaren Beitrag zur Vorhersage des Endergebnisses im Auswahlverfahren und könnte als Vorauswahlkriterium einen valideren Beitrag für die Aussagekraft der Bewerbungsunterlagen liefern.
Das Ergebnis konnte auch durch die regressionsanalytischen Berechnungen bestätigt werden (Determinationskoeffizient = .063), so dass insgesamt nur ca. 6%, der Varianz des Kriteriums durch die Schulnoten erklärt wurden.

Für die Vorhersage des Kriteriums „Gesamtnote in der Zwischenprüfung“ ergaben sich zufrieden stellend höhere signifikante bivariate korrigierte Korrelationen (Deutsch: r = .22; Mathematik: r = .25; Geschichte et al.: r = .32). Vergleichbar war auch hier das Bild in der Analyse der multiplen linearen Regression. Hier wies neben der Note in Geschichte auch die Durchschnittsnote einen vergleichsweise hohen Validitätskoeffizient von r = .3 auf.
Die Kosten-Nutzen Analyse zeigt im Vergleich zu einer bloßen Zufallsauswahl ($r_{xy} = 0$) einen deutlichen monetären Nutzenzuwachs.

Auch wenn der Beitrag der Prädiktoren „Schulnoten“ zur Vorhersage beider Kriterien (Endergebnis im Auswahlverfahren, Note in der Zwischenprüfung) überwiegend gering bzw. nur unzureichend ist – bei der Bewertung ist die Diskussion unterschiedlicher Aspekte wie Multikollinearität durch hohe Interkorrelation, Suppressionseffekte, Auswirkungen bei Skalentransformationen, nicht empirisch belegter Gütekriterien von Prädiktor und Kriterium sowie Effekte der Stichprobenselektion und fehlende Kreuzvalidierung zu beachten - können Schulnoten dazu beitragen, den monetären Nutzen des Auswahlverfahrens zu verbessern. In Anbetracht der fehlenden Güte, sind jedoch auch Kosten und Nutzen Überlegungen in Relation zur Zufallsauswahl kein überzeugendes Argument für die Verwendung von Schulnoten als alleinige Prädiktoren zur Vorhersage des Ergebnisses im Auswahlverfahren bzw. des Studienerfolges. Jedoch können einzelne Schulnoten im Rahmen der Vorauswahl, als Baustein einer Testbatterie und damit einhergehender Vorselektion, die Güte des Auswahlverfahrens erhöhend, eingesetzt werden.

In der Untersuchung 2 wurde die prädiktive Validität des aktuellen Auswahlverfahrens für die Einstellung in den gehobenen Polizeivollzugsdienst erschlossen. Als Prädiktoren dienten die von den später eingestellten Bewerbern erreichten Punkte in den einzelnen Testbausteinen (selektive Teilsstichprobe) des Auswahlverfahrens sowie dessen Gesamtergebnis. Kriteriumsvariable war die erreichte Gesamtpunktzahl der Studenten nach Einstellung in den gehobenen Polizeivollzugsdienst in der nach 1 ½ Jahren stattfindenden Zwischenprüfung an der Fachhochschule. Neben der Erschließung der prädiktiven Validität wurde weiter die inkrementelle Validität der zu einem Gesamtverfahren zusammengestellten Testbausteine untersucht und zudem die Kosten-Nutzen-Relation berechnet.

Analog zu Untersuchung 1 ist auch bei der Bewertung der Ergebnisse wieder zu berücksichtigen gewesen, dass leider keine Rohwerte vorlagen, sondern nur Punktwerte in den einzelnen Testbausteine, die nach unterschiedlichen Maßstäben transformiert wurden. Verzerrungen der Ergebnisse sind, wie unter Kapitel 4 und 5 beschrieben, daher nicht auszuschließen gewesen.

Bei der Beantwortung der Hypothesen zur prädiktiven Validität wurden zwei von acht Korrelationen auf dem 1 % Niveau signifikant („Bericht Inhalt“: r = .238; „Gesamt-testergebnis“: r = .663). Wie schon erwähnt, bedeutet in praxi ein Validitätskoeffizienten von r = .50 eine sehr gute Prognose des beruflichen Erfolges. Insbesondere der Prädiktor „Gesamttestergebnis“ verspricht eine gute Vorhersage für das erfolgreiche Abschneiden in der Zwischenprüfung.
Die ß-Gewichte des Prädiktors „Bericht Inhalt“ leisten auch in der Regressionsgleichung (R = .378; R^2 = .143) neben dem Prädiktors „Kulturfairer Leistungstest“ einen signifikanten Beitrag zur Vorhersage des Kriteriums „Endergebnis in der Zwischenprüfung“. Allerdings werden durch die verwendeten Prädiktoren nur 14,3 % der Varianz des Kriteriums aufgeklärt.

Angesichts der Tatsache, dass nur zwei der acht Prädiktoren aus dem Regressionsmodell signifikante ß-Gewichte erreichen, sind die Ergebnisse der Analyse der inkrementellen Validität von Interesse, da nur diejenigen Prädiktoren in das Modell aufgenommen werden, die einen signifikanten Beitrag zur Vorhersage des Kriteriums liefern. Die Testbausteine „Bericht Inhalt“, „Kulturfairer Leistungstest“ und „Rundgespräch“ zeigten inkrementelle Validität im Regressionsmodell (R = .357) und klärten 12,8 % der Kriteriumsvarianz auf. Auch hier wurde deutlich, dass ein erheblicher Anteil an der Varianz des Kriteriums (ca.

87%) nicht durch die Prädiktoren aufgeklärt wird.
Auffallend war der Prädiktor „Rundgespräch“, der in der Regressionsgleichung aller Prädiktoren (Tabelle 5.6-4) kein signifikantes ß-Gewicht erreicht, bei der schrittweisen Regression jedoch als dritter Prädiktor in das Modell aufgenommen wurde. Möglicherweise bietet die negative Interkorrelation der Prädiktoren „Kulturfairer Leistungstest“ und „Rundgespräch“ (r = .186), die beide positiv mit dem Kriterium korrelieren, einen Hinweis auf einen reziproken Suppressionseffekt.
Die Hypothese, dass jedes Element im Auswahlverfahren durch seine Berücksichtigung signifikant die prognostische Validität des Gesamtverfahrens erhöht, konnte somit nicht bestätigt werden.

Im Rahmen der Kosten-Nutzen Analyse für die Untersuchung 2 ergab sich für den Prädiktor „Gesamttestergebnis“ im Vergleich zu einer bloßen Zufallsauswahl (r_{xy}= 0) für die nächsten knapp 10 Jahre ein deutlicher Nutzenzuwachs. Im Vergleich dieses Prädiktors mit der höchsten prognostischen Validität mit einem Prädiktor, welcher einen niedrigeren Validitätskoeffizienten aufweist (kulturfairer Leistungstest) ergab sich über den gleichen Zeitraum ebenfalls ein zufrieden stellender Nutzenzuwachs.

Bei der Gesamtbewertung der Untersuchung 2 ist zu bedenken, dass auch hier oben genannte Aspekte (Multikollinearität, Suppressionseffekte, Auswirkungen bei Skalentransformationen, Gütekriterien von Prädiktor und Kriterium, Stichprobenselektion und fehlende Kreuzvalidierung etc.) die Aussagekraft des Ergebnisses schmälern. Es wurde jedoch deutlich, dass die Prädiktoren „Bericht Inhalt“, „Kulturfairer Leistungstest“ und „Rundgespräch“ diejenigen Testbausteine sind, welche die höchste Nützlichkeit für die Vorhersage des Kriteriums aufweisen. Unter Berücksichtigung der prädiktiven und inkrementellen Validität sowie der Kosten–Nutzen Analyse erweisen sich diese Prädiktoren durchaus von praktischem Wert zur Vorhersage des Abschneidens in der Zwischenprüfung.

Mit Durchführung der Untersuchung 3 sollte das Ziel, Einführung eines kulturfairen Leistungstests zur vermehrten Einstellung ausländischer Mitbürger in den gehobenen Polizeivollzugsdienst, hinterfragt werden. Im t-Test für unabhängige Stichproben erreichten im kulturfairen Leistungstest die nicht deutschen Teilnehmer eine niedrigere Punktzahl als die deutschen Teilnehmer ($M_{x\ \text{nicht deutsch}}$= 6,20; $M_{x\ \text{deutsch}}$= 7,17 Punkte). Die Hypothese, dass sich keine Unterschiede zwischen beiden Gruppen im kulturfairen Testverfahren ergeben, musste

zurückgewiesen werden. Die Analyse der Mittelwertsunterschiede im nicht kulturfairen Leistungstest ergab dagegen zwar wie erwartet parametrisch einen nachweisbaren statistischen Unterschied ($M_{x\ nicht\ deutsch}$ = 16,85; $M_{x\ deutsch}$ = 19,0 Punkte), diese konnte allerdings nicht non parametrisch abgesichert werden.
Insgesamt konnte das Ziel des Einsatzes des kulturfairen Leistungstests nicht erreicht werden. Es ist eher anzunehmen, dass die Teilstichprobe der nicht deutschen Bewerber sich von der a priori definierten Teilstichprobe unterscheidet. Trotz der Berücksichtigung der kleinen Stichprobe und unter Mitbewertung der in den anderen Testbausteinen erreichten Ergebnisse der Teilnehmer, ist eher anzunehmen, dass die Bewerber nicht deutscher Herkunft a priori schlechtere Leistungen erzielten. Um der politischen Vorgabe Rechnung tragen zu können, muss in erster Linie eine Bewerbergruppe mit weitergehenden kognitiven und anderen Fähigkeiten angesprochen werden. Dieses Ziel scheint bislang nicht erreicht.

Im Rahmen der vorliegenden drei Untersuchungen wurde ein grundlegendes Problem bei der Durchführung deutlich. Insbesondere für die Analyse der Vorhersageleistung von Prädiktordaten können nur „vollständige“ Datenpaare verwendet werden (Prädiktor- und Kriteriumsdaten). Wie generell bei Bewährungskontrollen, besaßen somit nicht alle zum Auswahlverfahren eingeladenen Bewerber zwei Datenpaare. In der Folge kommt es zu einer abgeschnittenen Verteilung, d.h. leistungsschwach beurteilte Bewerber fallen aufgrund der sequentiellen Auswahlstrategie und der Tatsache, dass Organisationen nur Erfolg versprechende Kandidaten einstellen, heraus. Möglicherweise wären aber gerade die leistungsschwächeren Kandidaten bei der Überprüfung der Ausbildungsergebnisse gescheitert, bzw. hätten deutlich schlechtere Ergebnisse erzielen können. Aus diesem Grund werden die Zusammenhänge zwischen Eignungsbeurteilungen und den berufsrelevanten Kriterien meist deutlich unterschätzt (Althoff, 1986). Als Folge der mangelnden Streuung zwischen den Bewerbern können die Voraussetzungen für korrelationsstatistische Analysen (z.B. Zufallsverteilung und Normalverteilung der Daten) nicht erfüllt werden (vgl. Bortz, 1993). Neben der oben beschriebenen „natürlichen“ Selektion der Bewerber durch das Auswahlverfahren und u.a. der Zulassung zur Zwischenprüfung, ist eine Randomisierung der Bewerber nicht möglich gewesen. Folglich müssen die erhobenen Daten zwangsweise einer systematischen Selektion unterliegen. Nach Czienskowski (1996) muss allerdings in vielen Untersuchungen auf eine Randomisierung verzichtet werden, da es sich um Unterschiede zwischen „natürlichen“ Gruppen handelt, wie es auch in der vorliegenden Arbeit der Fall ist.

Trotzdem Bewerbungen für den Polizeivollzugsdienst aus allen Bundesländern angenommen wurden, kann es zu Einschränkungen der Repräsentativität der Stichprobe kommen. Insbesondere die unterschiedlichen Schulsysteme und Wahlmöglichleiten in den Oberstufen der einzelnen Bundesländern (z.B. gibt es ein Einheitsabitur nur in den drei Ländern Baden-Württemberg, Bayern und Schleswig-Holstein) könnten sich bei der Untersuchung zur prädiktiven Validität der Schulnoten den empirischen Zusammenhang reduzierend ausgewirkt haben.

Aufgrund der geringen Stichprobenanzahl (n = 4) konnte die Frage nach möglichen Beziehungen zwischen den Prüfungsergebnissen (Auswahlverfahren und Ergebnissen der Zwischenprüfung an der Fachhochschule) deutscher und nicht deutscher Teilnehmer sowie deren Ergebnisse im kulturfairen und nicht kulturfairen kognitiven Testverfahren nicht empirisch untersucht werden. Hier wäre zukünftig eine langfristig angelegte Studie mit einem größeren N aussagekräftiger und wünschenswert.

Generell ist kritisch zur Untersuchung 3 anzumerken, dass der Vergleich zwischen einem kulturfairen und einen nicht kulturfairen Testverfahren im Grunde der Intention des ersten widerspricht. Durch den Einsatz des kulturfairen Testverfahrens sollten mögliche sprachliche und kulturelle Defizite nivelliert werden, um mehr Bewerber nicht deutscher Herkunft für den Polizeivollzugsdienst gewinnen zu können. Wenn es zukünftig gelingen sollte, generell geeignete Bewerber ausländischer Herkunft zu gewinnen, die zwar einerseits über ausreichende Fähigkeiten für den Polizeivollzugsdienst verfügen, andererseits jedoch aufgrund kulturell bedingter Einflüsse benachteiligt wären, so wäre die Hinzunahme eines sprachgebundenen weiteren Leistungstest eine Maßnahme, die das Ziel konterkariert und ad absurdum führt.

Weiter ist kritisch anzumerken, dass das bisherige Auswahlverfahren zwar auf den bisherigen Erfahrungen und Überlegungen der durchführenden Organisation fußt, jedoch bislang keiner regelmäßigen empirisch wissenschaftlichen Überprüfung sich hat stellen müssen. Eine regelmäßige Beurteilung und empirischen Überprüfung der durchgeführten Eignungsbeurteilung ist nicht zuletzt nach Einführung der DIN 33430 zur Qualitätssicherung und Anpassung an die sich regelmäßig verändernden Rahmenbedingungen sowohl fachlich wie auch ökonomisch notwendig.

Auch die Anpassung des Auswahlverfahrens an die zwischenzeitlich durchgeführten Anforderungsanalysen für den Polizeivollzugsdienst sollte zeitnah erfolgen. Es ist anzunehmen, dass die vorliegenden Ergebnisse Auswirkung auf die Zusammensetzung des Auswahlverfahrens haben, zudem die Ergebnisse dieser Arbeit daraufhin deuten, dass einige Testbausteine des Auswahlverfahrens unzulänglich sind. Nach Abstimmung des Auswahlverfahrens mit den Anforderungen und Integration bisheriger Ergebnisse, sollte eine wissenschaftlich fundierte regelmäßig durchgeführte Bewertung sich nicht nur an den Ergebnissen der Zwischenprüfung orientieren, wie es in dieser Arbeit aus zeitlichen Gründen geschehen ist, sondern auch an der Abschlussprüfung. Weiter wäre es vorausblickend wünschenswert, wenn es darüber hinaus gelänge, sowohl das Auswahlverfahren, wie auch das erreichte Ergebnis der Ausbildung, an Kriterien des Berufserfolges empirisch zu überprüfen. Trotz der bekannten Schwierigkeiten bei der Definition geeigneter Kriterien für den Berufserfolg bleibt es das übergeordnete Ziel, nicht nur den Ausbildungserfolg, sondern in erster Linie den Berufserfolg vorher zu sagen.

In dieser Arbeit blieb der Aspekt der sozialen Validität überwiegend unberücksichtigt. Doch gerade das Ziel, vermehrt geeignete nicht deutsche Bewerber für den Beruf anzusprechen, macht unter Umständen deutlich, wie wichtig auch dieser Faktor für die hier durchgeführte Eignungsuntersuchung ist. Das vorliegende Ergebnis belegt, dass die Gruppe der bisher angesprochenen nicht deutschen Bewerber eher schlechtere Ergebnisse erzielen konnte als im Vergleich mit ihren deutschen Mitbewerbern. Damit bliebe anzunehmen, dass diejenigen, welche gleich gute Ergebnisse erzielen könnten, sich nicht für den Beruf der Polizeibeamten interessieren oder aber auch durch das Image der Organisation abgeschreckt werden. Die Auswahlsituation stellt häufig den Erstkontakt zwischen Bewerber und Organisation dar. Aufgrund fehlender anderer Erfahrungen können Bewerber in diesem Stadium den Eindruck gewinnen, dass die Auswahlsituation und das von der Organisation gezeigte Verhalten repräsentativ für die Organisation seien. D.h., die Auswahlsituation vermittelt ein Bild der Organisation, welches, insbesondere durch abgelehnte Bewerber, die keine Gelegenheit finden, diesen selektiven Eindruck zu kompensieren, an andere potentielle Bewerber weiter getragen wird. Die Frage, ob Personen, die das Auswahlverfahren bereits absolviert haben, anderen eine Bewerbung bei der Organisation empfehlen, hängt damit durchaus auch davon ab, welches im Auswahlverfahren wahrgenommene Image der Organisation zugeschrieben und weiter getragen wird. Nicht nur zur Beantwortung der Frage, wie weitere geeignete Bewerber nicht deutscher Herkunft interessiert werden können, sondern generell werden

Organisation in Zeiten schlechter werdender Schulabschlüsse zunehmend in Konkurrenz um geeignete Bewerber stehen. Insofern sollte es ein Nahziel künftiger Arbeiten sei, diesen Aspekt mit zu erfassen.

8. Literatur

Amelang, M & Zielinski, W. (1994). Psychologische Diagnostik und Intervention. Berlin: Springer – Verlag.

Amthauer, R. (1970). Intelligenz-Struktur-Test 70 (IST 70). Göttingen: Hogrefe.

Amthauer, R., Brocke, B., Liepmann, D. & Beauducel, A. (1997). Intelligenz-Struktur-Test 2000 (IST 2000). Göttingen: Hogrefe.

Anastasi, A. (1985). The use of personality assessment in industry: Methodological and interpretive problems. In H.J. Bernardin & D.A. Bownas (Eds.), Personality assessment in organizations, 1-20. New York: Praeger.

Althoff, K. (1977). Zusammenhänge zwischen Ergebnissen von Eignungstests und beruflicher Bewährung. Schriftenreihe der Polizeiführungsakademie, Heft 1, 6-27.

Althoff, K. (1984). Zur prognostischen Validität von Intelligenz- und Leistungstests im Rahmen der Eignungsdiagnostik. Psychologie und Praxis. Zeitschrift für Arbeits- und Organisationspsychologie, Band 28, 144-148.

Althoff, K. (1986). Zur Aussagekraft von Schulzeugnissen im Rahmen der Eignungsdiagnostik. Psychologie und Praxis. Zeitschrift für Arbeits- und Organisationspsychologie, Band 30 (N.F.4), 77-85.

Arvey, R.D. & Campion, J.E. (1982). The employment interview: A summary and review of recent research. Personnel Psychology, 35, 281-322.

Backhaus, K., Erichson, B., Plinke, W. und Weiber, R. (1994). Multivariate Analysemethoden. (7. Auflage). Berlin: Springer-Verlag.

Backhaus, J. & Wagner, R. (1994). Ausbilder – Taschenbuch 1995. Stuttgart: Deutscher Sparkassenverlag.

Barrick, M. R. & Mount, M.K. (1991). The Big Five Personality Dimensions and Job Performance: A Meta-Analysis. Personnel Psychology, 44, 1 – 26.

Barthel, E. & Schuler, H. (1989). Nutzenkalkulation eignungsdiagnostischer Verfahren am Beispiel eines biographischen Fragebogens. Zeitschrift für Arbeits- und Organisationspsychologie, 33, 73 – 83.

Baron-Boldt, J. (1989). Die Validität von Schulabschlußnoten für die Prognose von Ausbildungs- und Studienerfolg: Eine Metaanalyse nach dem Prinzip der Validitätsgeneralisierung. Frankfurt am Main: Peter Lang.

Baron-Boldt, J., Funke, U. & Schuler, H.(1989). Prognostische Validität von Schulnoten. Eine Metaanalyse der Prognose des Studien- und Ausbildungserfolges. In R.S. Jäger, R. Horn & K. Ingenkamp (Hrsg.), Tests und Trends 7, 11-39. Weinheim: Beltz.

Barthel, E. & Schuler, H. (1989). Nutzenkalkulation eignungsdiagnostischer Verfahren am Beispiel eines biographischen Fragebogens. Zeitschrift für Arbeits- und Organisationspsychologie, 33, 73 – 83.

Behrens, R. & Merkel, R. (1989). Erfolgreiche Personalauswahl. Leitfaden für Unternehmer und Führungskräfte. Köln: TÜV-Rheinland.

Berthel, J. (1995). Personal-Management. Stuttgart: Schäffer-Poeschel.

Bliesener, T. (1991). Die Validität Biographischer Daten bei der Prognose des Berufserfolgs: Erste Ergebnisse einer metaanalytischen Studie zur Validitätsgeneralisierung. In H. Schuler & U. Funke (Hrsg.), Eignungsdiagnostik in Forschung und Praxis. Stuttgart: Verlag für Angewandte Psychologie.

Booth, J.F. (1991). Die Anwenderschnittstelle - Schlüssel zum anwenderfreundlichen und validen computergestützten Testen. In H. Schuler & U. Funke (Hrsg.), Eignungsdiagnostik in Forschung und Praxis, Bd.10, 70-76. Stuttgart: Verlag für Angewandte Psychologie.

Borg, I. & Staufenbiel, T. (1997). Theorien und Methoden der Skalierung. In K. Pawlik (Hrsg.), Methoden der Psychologie, Band 11. Bern: Verlag Hans Huber.

Borkenau, P. & Ostendorf, F. (1993). NEO-Fünf-Faktoren-Inventar (NEO-FFI). Göttingen: Hogrefe.

Boudreau, J. (1989). Economic considerations in estimating the utility of human resource productivity improvement programs. Personnel Psychology, 36, 551 – 576.

Bortz, J. (1984). Lehrbuch der empirischen Forschung für Sozialwissenschaftler. Berlin: Springer-Verlag.

Bortz, J. (1989). Statistik für Sozialwissenschaftler. 3.Auflage. New York, Berlin, Heidelberg: Springer Verlag.

Bortz, J. & Döring, N. (1995). Forschungsmethoden und Evaluation für Sozialwissenschaftler (2.Auflage). Berlin: Springer-Verlag.

Brandstätter, H. (1979). Die Ermittlung personaler Eigenschaften kognitiver Art. In G. Reber (Hrsg.) Personalinformationssysteme, 74 – 95. Stuttgart: Poeschel.

Brickenkamp, R. (1962). Aufmerksamkeits-Belastungstest (d2). Göttingen: Hogrefe.

Brosius, G. und Brosius F. (1995). SPSS. Base System und Professional Statistics. Bonn: International Thomson Publishing.

Bühner, M. & Schmidt-Atzert, L. (2003). Arbeitsgedächtnis & Aufmerksamkeit: Ein Beitrag zur Struktur und Validität beider Konzepte. 7. Arbeitstagung der Fachgruppe Differentielle Psychologie und Persönlichkeitspsychologie und Psychologische Diagnostik in Halle, 29. bis 30. September 2003.

Cascio, W.F. (1991). Costing human resources: The financial impact of behavior in organizations, 3rd. Kent: Publishing Company.

Conrad, Baumann & Mohr (1980). Mannheimer Test zur Erfassung des physikalischen-technischen Problemlösens (MTP). Göttingen: Hogrefe.

Comer, R.J. (1995). Klinische Psychologie. Heidelberg: Spektrum Akademischer Verlag.

Cronbach, L.J. & Gleser, G.C. (1965) . Psychological tests and personnel decisions. 2nd ed. Urbana: University of Illinois Press.

Czienskowski, U. (1996). Wissenschaftliche Experimente: Planung, Auswertung, Interpretation. Weinheim: Psychologie Verlags Union.

Dachler, H.P. (1995). Managementdiagnostik als sozialer Prozess. In W. Sarges (Hrsg.), Managementdiagnostik. Göttingen: Hogrefe.

Dicker, H. (1989). Die Reliabilität der Beurteilung von Mathematikarbeiten. In: K. Ingenkamp (Hrsg.): Die Fragwürdigkeit der Zensurengebung. Texte und Untersuchungsberichte. Weinheim: Beltz Verlag.

DIN 33430 (2002). Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen. Normausschuss Gebrauchstauglichkeit und Dienstleistungen (NAGD) im DIN Deutsches Institut für Normung e.V. Berlin: Deutsches Institut für Normung e.V.

Dörner, D. (1996). Die Logik des Misslingens. Strategisches Denken in komplexen Situationen. Reinbeck: Rowohlt.

Dörner, D. , Kreuzig, H.W., Reither, F. & Stäudel, T. (Hrsg.) (1983). Lohhausen: Vom Umgang mit Unbestimmtheit und Komplexität. Bern: Huber.

Dorsch, F., Häcker, H, Stapf, K.-H. (1987). Psychologisches Wörterbuch. Bern: Verlag Hans Huber.

DuBois, P.H. (1970). A history of psychological testing. Boston: Allyn & Bacon.

Dücker, H. (1959). Konzentrations-Leistungs-Test (KLT). Göttingen: Hogrefe.

Dumke, D. (1973). Schülerleistung und Zensur. Ergebnisse aus der Arbeit der Niedersächsischen Lehrerfortbildung, Heft 23. Hannover: Hermann Schroedel Verlag.

Eckhardt, H.H. & Schuler, H. (1992). Berufseignungsdiagnostik. In R.S. Jäger & F. Petermann (Hrsg.), Psychologische Diagnostik, 533-551. Weinheim: Psychologische Verlags Union.

England, G.W. (1971). Development and Use of Weighted Application Blanks. Industrial Relations Center. University of Minnesota.

Färber, B. (1995). Probleme der Evaluation. In W. Sarges (Hrsg.), Managementdiagnostik. Göttingen: Hogrefe.

Fiedler, K. (1997). Die Verarbeitung sozialer Informationen für Urteilsbildung und Entscheidungen. In W. Stroebe, M. Hewstone & G. M. Stephenson (Hrsg.), Sozialpsychologie, 143 – 175. Berlin: Springer.

Fisseni, H.J. (1990). Lehrbuch der psychologischen Diagnostik. Göttingen: Hogrefe.

Feger, H. (1983). Planung und Bewertung von wissenschaftlichen Beobachtungen. In H.Feger & J. Bredenkamp (Hrsg.), Datenerhebung. Enzyklopädie der Psychologie. Bd.1, Forschungsmethoden, 1-75. Göttingen: Hogrefe.

Flanagan, J. C. (1949). Critical requirements for research personnel: A study of observed behaviours of personnel in research laboratories. Pittsburgh: American Institute for Research.

Frieling, E. & Hoyos, C. Graf (1978). Fragebogen zur Arbeitsanalyse. (FAA) Bern: Huber.

Frieling, E. & Sonntag, K. (1999). Lehrbuch Arbeitspsychologie. Bern: Huber.

Funke, U. (1986). Die Validität verschiedener eignungsdiagnostischer Verfahren bei Lehrstellenbewerbern. Zeitschrift für Arbeits- und Organisationspsychologie, 30, 92-97.

Funke, U. & Barthel, E. (1995). Nutzenanalyse von Personalauswahlprogrammen. In W. Sarges (Hrsg.), Managementdiagnostik. Göttingen: Hogrefe.

Funke,U., Krauß,J., Schuler,H. & Stapf,K.H. (1987). Zur Prognostizierbarkeit wissenschaftlich technischer Leistungen mittels Personenvariablen: Eine Metaanalyse der Validität diagnostischer Verfahren im Bereich Forschung und Entwicklung. Gruppendynamik, 18, 407 – 428.

Fruhner, R., Schuler, H., Funke, U. & Moser, K. (1991). Einige Determinanten der Bewertung von Personalauswahlverfahren. Zeitschrift für Arbeits- und Organisationspsychologie. 35, 4, 170 – 178.

Gittler, G. (1990). Dreidimensionaler Würfeltest. Weinheim: Beltz.

Graudenz, H. (1987). Eignungsuntersuchung aus der Sicht der Bewerber. DGP Information, 38, 1 – 28.

Groffmann, K.J. (1983). Die Entwicklung der Intelligenzmessung. In: Enzyklopädie der Psychologie, Band 2, Intelligenz- und Leistungsdiagnostik. Göttingen: Hogrefe.

Guthke, J., Böttcher, H.R. & Sprung, L. (1990). (Hrsg.), Psychodiagnostik Band 1. Berlin: VEB Deutscher Verlag der Wissenschaften.

Guttman, L. (1972). Measurement as structural theory. Psychometrika, 36, 329-347.

Hanisch, G. (1990). Problematik der Leistungsfeststellung durch schriftliche Arbeiten am Beispiel der Mathematik. Habilitationsschrift an der Grund- und Integrativwissenschaftlichen Fakultät der Universität Wien.

Heyse, H. & Kersting, M. (2004). Anforderungen an den Prozess der Eignungsbeurteilung. In L. Hornke /U. Winterfeld (Hrsg.), Eignungsbeurteilungen auf dem Prüfstand: DIN 33430 zur Qualitätssicherung. Heidelberg: Spektrum Akademischer Verlag.

Hollmann, H. (1991). Validität in der Eignungsdiagnostik. Göttingen: Hogrefe.

Hollmann, H. & Reitzig, G. (1995). Referenzen und Dokumentenanalyse. In W. Sarges (Hrsg.), Managementdiagnostik, 463-470.Göttingen: Hogrefe.

Höft, S. (2001). Erfolgsüberprüfung personalpsychologischer Arbeit. In Schuler, H. (Hrsg.), Lehrbuch der Personalpsychologie. Göttingen: Hogrefe.

Höft, S. & Funke, U. (2001). Simulationsorientierte Verfahren der Personalauswahl. In H. Schuler (Hrsg.), Lehrbuch der Personalpsychologie. Göttingen: Hogrefe.

Hörmann, H. (1964). Aussagemöglichkeiten psychologischer Diagnostik. Göttingen: Hogrefe.

Horn, W. (1962). Leistungsprüfsystem (LPS). Göttingen: Hogrefe.

Hornke, L.F. (1991). Neue Itemformen für computergestütztes Testen. In H. Schuler & U. Funke (Hrsg.), Eignungsdiagnostik in Forschung und Praxis, Bd.10, 67-77. Stuttgart: Verlag für Angewandte Psychologie.

Hossiep, R. (1995). Berufseignungsdiagnostische Entscheidungen. Zur Bewährung eignungsdiagnostischer Ansätze. Göttingen: Hogrefe.

Hossiep, R., Paschen, M. & Mühlhaus, O. (2000). Persönlichkeitstests im Personalmanagement. Grundlagen, Instrumente und Anwendungen. Göttingen: Verlag für Angewandte Psychologie.

Hoyos, C.G. (1986). Die Rolle der Anforderungsanalyse im eignungsdiagnostischen Prozess. Psychologie und Praxis. Zeitschrift für Arbeits- und Organisationspsychologie, Band 30, 59-67.

Huber, M. (1996). Betreuung von Opfern und Angehörigen. In: M. Hermanutz, C. Ludwig & H.P. Schmalzl. Moderne Polizeipsychologie in Schlüsselbegriffen. 33 – 39.Stuttgart: Boorberg Verlag.

Hüttemann, J. (2002). Berliner Intelligenzstruktur-Test (BIS-4). In U. P. Kanning und H. Holling. Handbuch personaldiagnostischer Instrumente. Göttingen: Hogrefe.

Hunter, J.E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. Journal of Vocational Behavior, 29, 340-362.

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

Imming, G. (1920). Die Arbeitsprobe. Praktische Psychologie, 2, 338 – 344.

Ingenkamp, K. (1969) Möglichkeiten und Grenzen des Lehrerurteils und der Schultests. In: H. Roth (Hrsg.): Begabung und Lernen, 407-431. Stuttgart: Klett.

Ingenkamp, K. (1989). Die Fragwürdigkeit der Zensurengebung. Texte und Untersuchungsberichte. Weinheim: Beltz Verlag.

Ingenkamp, K. (1977). Zur Problematik der Zensurengebung. In C. Schwarzer & R. Schwarzer (Hrsg.) Diagnostik im Schulwesen. Studientexte zur pädagogischen Diagnose, Beratung und Entscheidung,15 – 36. Braunschweig: Westermann Verlag.

Ingenkamp, K. (1995). Lehrbuch der pädagogischen Diagnostik. Weinheim: Beltz Verlag.

Jäger, A. O. (1960). Zum prognostischen Wert psychologischer Eignungsuntersuchungen. Psychologische Rundschau, Band 11, 160-178.

Jäger, A. O. (1966). Prognose und Bewährung in der Eignungsdiagnostik. Psychologische Rundschau, Band 17, 186-208.

Jäger, A.O. (1984). Personalauslese. In A. Mayer & B. Herwig (Hrsg.), Betriebspsychologie. Handbuch der Psychologie, Bd.9, 613-667. Göttingen: Hogrefe.

Jäger, A.O. (1986). Validität von Intelligenztests. Diagnostica, 32 (4), 272-289.

Jäger, A.O. & Althoff, K. (1994). Wilde-Intelligenz-Test (WIT). Göttingen: Hogrefe.

Jäger, A.O. Süß, H.M. & Beauducel, A. (1997). Berliner Intelligenzstruktur-Test.(BIS). Göttingen: Hogrefe.

Jetter, W. (1996). Effiziente Personalauswahl. Durch strukturierte Einstellungsgespräche die richtigen Mitarbeiter finden. Stuttgart: Schäffer-Poeschel.

Jeserich, W. (1991). Mitarbeiter auswählen und fördern. Assessment Center Verfahren. Handbuch der Weiterbildung für die Praxis in Wirtschaft und Verwaltung. 6. Auflage. München: Hanser

Jödden, C. (2002). Wilde-Intelligenz-Test (WIT). In U. P. Kanning und H. Holling. Handbuch personaldiagnostischer Instrumente. Göttingen: Hogrefe.

Kallus, K. W. & Jahnke, W. (1992). Klassenzuordnungen. In S. Jäger und F. Petermann (Hrsg.), Psychologische Diagnostik, 2.Aufl., 170 – 186, Weinheim: Psychologische Verlags Union.

Kannheiser, W. & Frieling, E. (1992). Arbeitsstrukturierung und Arbeitsanalyse. In D. Frey, Graf C. Hoyos & D. Stahlberg (Hrsg.), Angewandte Psychologie, 129-146. Weinheim: PVU.

Kanning, U.P. (1999). Die Psychologie der Personenbeurteilung. Göttingen: Hogrefe.

Kanning, U.P. (2002a). Nicht -standardisierte Methoden. In U.P. Kanning & H. Holling (Hrsg.), Handbuch personaldiagnostischer Instrumente, 118-124. Göttingen: Hogrefe.

Kanning, U.P. (2002b). Tipps für die Anwendung nicht-standardisierter Methoden. In U.P. Kanning & H. Holling (Hrsg.), Handbuch personaldiagnostischer Instrumente, 493-543. Göttingen: Hogrefe.

Kanning, U.P. & Holling, H. (2002). Handbuch personaldiagnostischer Instrumente. Göttingen: Hogrefe.

Kaufmann, W. (1984). Wie man die "beste" Führungskraft auswählt. Die Assessment-Center-Methode zur Auswahl des Führungskräftenachwuchses. Management Zeitschrift, 53, 474 – 476.

Kastner, M. (1995). Klinische Urteilsbildung. In W. Sarges (Hrsg.), Management-Diagnostik. Göttingen: Hogrefe.

Kersting, M. (1998). Differentielle Aspekte der sozialen Akzeptanz von Intelligenztests und Problemlöseszenarien als Personalauswahlverfahren. Zeitschrift für Arbeits- und Organisationspsychologie, 42, 61 – 75.

Kersting, M. (2004a). Kosten und Nutzen beruflicher Eignungsbeurteilungen. In L. Hornke & U. Winterfeld (Hrsg.), Eignungsbeurteilungen auf dem Prüfstand: DIN 33430 zur Qualitätssicherung. Heidelberg: Spektrum Akademischer Verlag.

Kersting, M. (2004b). Qualitätssicherung und -verbesserung: Zur Überprüfung der Gültigkeit berufsbezogener Eignungsbeurteilungen. In L. Hornke & U. Winterfeld (Hrsg.), Eignungsbeurteilungen auf dem Prüfstand: DIN 33430 zur Qualitätssicherung. Heidelberg: Spektrum Akademischer Verlag.

Kisser, R. (1992). Adaptive Strategien. In R.S. Jäger & F. Petermann (Hrsg.), Psychologische Diagnostik, 161-170. Weinheim: Psychologie Verlags Union.

Kleine, D. & Jäger, A.O. (1989). Kriteriumsvalidität eines neuartigen Tests zum Berliner Intelligenzstrukturmodell. Eine Untersuchung an brasilianischen Schülern und Studenten. Diagnostica, 35, 17 – 37.

Kleinevoss, R. & Sonnenberg, H. G. (1987). Die prognostische Validität einer eignungs-diagnostischen Prozedur – dargestellt mit einem konfiguralen Diskriminanzmodell. Psychologie und Praxis. Zeitschrift für Arbeits- und Organisationspsychologie, 31, 15 – 21.

Kleinmann, M. (1997). Assessment Center, Stand der Forschung – Konsequenzen für die Praxis. Göttingen: Verlag für Angewandte Psychologie.

Kleinmann, M. & Strauß, B. (1998). Potentialfeststellung und Personalentwicklung. Göttingen: Hogrefe.

Klimoski, R. & Brickner, M. (1987). Why do assessment centers work ? The puzzle of assessment center validity. Personal Psychology, 40, 243 – 260.

Krapp, A. (1979). Prognose und Entscheidung. Weinheim: Beltz.

Kubinger, K.D. (1993). Testtheoretische Probleme der Computerdiagnostik. Zeitschrift für Arbeits- und Organisationspsychologie, 37, 130 – 137.

Lammers, F. (1998). Personalentwicklung „off the job". In: M. Kleinmann & B. Strauß. Potentialfeststellung und Personalentwicklung. 199 – 218. Göttingen: Hogrefe.

Lang von Wins, T. & Rosenstiel, L. (1998). Potentialfeststellungsverfahren. In: M. Kleinmann & B. Strauß. Potentialfeststellung und Personalentwicklung. 71 – 96. Göttingen: Hogrefe.

Lienert, G.A. (1967). Die Draht-Biege-Probe (DBP). Göttingen: Hogrefe.

Lienert, G.A. (1987). Schulnoten-Evaluation. Frankfurt: Athenäum.

Lienert, G.A. (1989). Testaufbau und Testanalyse. 4.Auflage. Weinheim: PVU.

Lienert, G.A. & Orlik, P.(1965). Eine Maßzahl zur Bestimmung der Präzision psychologischer Planversuche. Zeitschrift Psychologie, 172, 203-216.

Lienert, G.A. & Raatz, U. (1994). Testaufbau und Testanalyse. Weinheim: Psychologie Verlags Union.

Ludwig, C. (1996). Stress. In: M. Hermanutz, C. Ludwig & H.P. Schmalzl. Moderne Polizeipsychologie in Schlüsselbegriffen. 216 – 228. Stuttgart: Boorberg Verlag.

Malamed, T. (1992). Use of biodata for predicting academic success over thirty years. Psychological Reports, 71, 31 – 38.

Mattenklott, A. (1992). Diagnostische Urteilsbildung. In R.S. Jäger & F. Petermann (Hrsg.), Psychologische Diagnostik. Weinheim: PVU.

Maukisch, H. (1978). Einführung in die Eignungsdiagnostik. In A. Mayer (Hrsg.), Organisationspsychologie, 105-136. Stuttgart: Poeschel.

McDaniel, M.A. (1994). The Validity of Employment Interviews: A Comprehensive Review and Meta-Analysis. Journal of Applied Psychology, 79, 599-616.

Mead, A. D. & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. Psychological Bulletin, 114, 449 – 458.

Meehl, P.E. (1954). Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. Minneapolis: University of Minnesota Press.

Meier, B.D. (2003). Ist der Erfolg im Jurastudium vorhersagbar? Empirische Befunde zum Zusammenhang zwischen Schulnoten und Abschneiden im Ersten Juristischen Staatsexamen. Beiträge zur Hochschulforschung. Heft 4, 25, 18 – 35. Bayrisches Staatsinstitut für Hochschulforschung und Hochschulplanung.

Mertens, D. (1974). Schlüsselqualifikationen. Thesen zur Schulung für eine moderne Gesellschaft. Mitteilungen aus der Arbeitsmarkt- und Berufsforschung, 36 – 43.

Michel, L. & Conrad, W. (1982). Theoretische Grundlagen psychometrischer Tests. In K.J. Groffmann & L. Michel (Hrsg.). Enzyklopädie der Psychologie, II: Psychologische Diagnostik, 1, 1 – 129.

Michel, L. & Mai, N. (1968). Entscheidungstheorie und Probleme der Diagnostik bei Cronbach & Gleser. Diagnostica, 14, 99-121. Göttingen: Hogrefe.

Moser, K. & Rhyssen, D. (2001). Referenzen als eignungsdiagnostische Methode. Zeitschrift für Arbeits- und Organisationspsychologie, 45, 40 – 46.

Moser, K. & Zempel, J. (2001). Personalmarketing. In H. Schuler (Hrsg.), Lehrbuch der Personalpsychologie. Göttingen: Hogrefe.

Moosbrugger, H. & Heyden, M. (1997). FAKT. Frankfurter Konzentrationsleistungs-Test. Bern: Huber.

Mummendey, H.D. (1987). Die Fragebogen-Methode. Grundlagen und Anwendung in Persönlichkeits-, Einstellungs- und Selbstkonzeptforschung. Göttingen: Hogrefe.

Neubauer, A. & Urban, E. (1991). Der Vergleich von computergestützter Testdarbietung und Standardvorgabe am Beispiel von Ravens Advanced Progressive Matrices. In H. Schuler & U. Funke (Hrsg.), Eignungsdiagnostik in Forschung und Praxis, 10, Stuttgart: Verlag für Angewandte Psychologie.

Nevo, B. (1986). Face validity and others related variables. In B. Nevo und R.S. Jäger (Eds.), Psychological Testing: The examinee perspective, 49 – 68. Göttingen: Hogrefe.

Otte, Rolf (1992). Die Methodik zur Prüfung von Offiziersbewerbern der Bundeswehr. In Schuler, H. und Stehle, W. (Hrsg.), Assessment Center als Methode der Personalentwicklung. Stuttgart: Verlag für Angewandte Psychologie.

Owens, W.A. (1976). Background data. In M.D. Dunette (Eds.), Handbook of industrial and organisational psychology. Chicago: Rand McNally.

Penning, S. (1996). Soziale Kompetenz und Persönliche Kompetenz. In: M. Hermanutz, C. Ludwig & H.P. Schmalzl. Moderne Polizeipsychologie in Schlüsselbegriffen. 204 – 215. Stuttgart: Boorberg Verlag.

Petersen, R. (2002). Biographie orientierte Personalauswahl im Kontext angewandter Eignungsdiagnostik. Unveröffentlichte Dissertation Universität zu Kiel.

Poortinga, Y.H., Coetsier, P., Meuris, G., Miller, K.M., Samsonowitz, V., Seisdedos, N. & Schlegel, E. (1982). A survey of attitudes towards tests among psychologists in six Western European countries. International Review of Applied Psychology, 31, 7 – 34.

Raststetter, D. (1999). Bewerbungsunterlagenscreening. Analyse eines hochselektiven, kaum erforschten Auswahlinstrumentes. Zeitschrift für Arbeitswissenschaft, 53, 37-44.

Raven, J.C. (1958). Advanced Progressive Matrices. London: Lewis.

Reilly, R.R. & Chao, G.T. (1982). Validity and Fairness of some alternative employee selection procedures. Personnel Psychology, 35 (1), 1-62.

Richter, P., Rudolf, M. & Schmidt, C.F. (1996). Fragebogen zur Analyse belastungsrelevanter Anforderungsbewältigung (FABA). Frankfurt: Sweet Test Services.

Rochel, H. (1983). Planung und Auswertung von Untersuchungen im Rahmen des allgemeinen linearen Modells. Berlin: Springer.

Rohmert, W. & Landau, K. (1979). Das arbeitswissenschaftliche Erhebungsverfahren zur Tätigkeitsanalyse (AET). Bern: Huber.

Rosenthal, R. & Jacobson, L. (1971). Pygmalion im Unterricht. Weinheim: Beltz.

Rosenstiel, L. v. (2001). Die Bedeutung von Arbeit. In H. Schuler (Hrsg.), Lehrbuch der Personalpsychologie, Göttingen: Hogrefe.

Rösler, F. (1992). Personalauslese, Training und Personalentwicklung in Organisationen. In Frey, D., Hoyos, C.G. & Stahlberg, D. (Hrsg.), Angewandte Psychologie. Ein Lehrbuch. Weinheim: Psychologie Verlags Union.

Rüppell, H. (1991). Computergestützte Simulationen und komplexe Probleme. In H. Schuler & U. Funke (Hrsg.), Eignungsdiagnostik in Forschung und Praxis, Bd.10, 91-121. Stuttgart: Verlag für Angewandte Psychologie.

Salgado, J. F. (1997). The Five Factor Model of Personality and Job Performance in the European Community. Journal of Applied Psychology, 82, 1, 30 – 43.

Sarbin, T.R., Taft, R. & Bailey, D.E. (1960). Clinical inference and cognitive theory. New York: Holt, Rinehart & Winston.

Sarges, **W. (1994).** Interviews. In W. Sarges (Hrsg.), Management-Diagnostik, 475-489. Göttingen: Hogrefe.

Sarges, **W. (1994).** Management-Diagnostik (2.Auflage). Göttingen: Hogrefe.

Sarges, W. (1996). Weiterentwicklung der Assessment Center-Methode. Göttingen: Verlag für Angewandte Psychologie.

Sarges, W. (2000). Einleitende Überlegungen zu Persönlichkeitstests im Personalmanagement. In: R. Hossiep, M. Paschen und O. Mühlhaus. Persönlichkeitstests im Personalmanagement. Grundlagen, Instrumente und Anwendungen, Göttingen: Verlag für Angewandte Psychologie.

Sarges, W. & Weinert, A. B. (1991). Früherkennung von Management-Potentialen. In W.E. Feix (Hrsg.), Personal 2000 – Visionen und Strategien erfolgreicher Personalarbeit, 267 – 301, Frankfurt, Wiesbaden: Gabler.

Sarges, W. & Wottawa, H. (2001). Handbuch wirtschaftspsychologischer Testverfahren. Lengerich: Pabst Science Publishers.

Sawyer, J. (1966). Measurement and prediction, clinical and statistical. Psychological Bulletin, 66, 178-200.

Scheinecker, M. & Wallner, S. (2003). Methoden der Potenzialanalyse in österreichischen Unternehmen. Empirische Untersuchung der Einsatzhäufigkeit verschiedener Verfahren der Potenzialanalyse. Wien: Trigon.

Schmale, H. & Schmidtke, H. (1966). Berufseignungstest (BET). Bern: Huber.

Schmale, H. & Schmidtke, H. (1969). Eignungsprognose und Ausbildungserfolg. Köln: Westdeutscher Verlag.

Schmidt, F.L. & Hunter, J.E. (1977). Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 62, 529-540.

Schmidt, F.L. & Hunter, J.E. (1981). Employment Testing: Old theories and new research findings. American Psychologist, 36, 1128-1137.

Schmidt, F.L. & Hunter, J.E. (1998). Meßbare Personenmerkmale: Stabilität, Variabilität und Validität zur Vorhersage zukünftiger Berufsleistung und berufsbezogenen Lernens. In M. Kleinmann & B. Strauß (Hrsg.), Potentialfeststellung und Personalentwicklung, 15-43. Göttingen: Verlag für Angewandte Psychologie.

Schmidt, F.L., Hunter, J.E., & Caplan, J.R. (1981). Validity generalization results for two job groups in the petroleum industriy. Journal of Applied Psychology, 66, 261 – 273.

Schmidt, F.L. & Hunter, J.E. & Pearlman, K.(1982). Assessing the economic impact of personnel programs of workforce productivity. Personnel Psychology, 35, 333-347.

Schmidt, F.L., Hunter, J.E., Pearlmann, K. & Shane, G.S. (1979). Further tests of the Schmidt Hunter Bayesian Validity Generalization Model. Personel Psychology, 32, 257-281.

Schmotzer, C., Kubinger, K.D. & Maryschka, C. (1994). Rechnen in Symbolen. Frankfurt: Sweet Test Service.

Schneewind, K.A., Schröder, G. & Cattell, R.B. (1994). Der 16-Persönlichkeitsfaktoren-Test (16 PF). 3. Aufl.. Bern: Huber.

Scholz, G. & Schuler, H. (1993). Das nomologische Netzwerk des Assessment Centers: Eine Metaanalyse. Zeitschrift für Arbeits- und Organisationspsychologie, 37, 73 – 85.

Sonntag, K. (1998). Personalentwicklung „on the job“. In: M. Kleinmann & B. Strauß. Potentialfeststellung und Personalentwicklung. 175 – 198. Göttingen: Hogrefe.

Schorr, A. (1991). Diagnostische Praxis in der Arbeits- und Organisationspsychologie. Teilergebnisse aus einer repräsentativen Umfrage zur diagnostischen Praxis. In H. Schuler & U. Funke (Hrsg.), Eignungsdiagnostik in Forschung und Praxis, 10, Stuttgart: Verlag für Angewandte Psychologie.

Schuler, H. (1988). Berufseignungsdiagnostik. Zeitschrift für Differentielle und Diagnostische Psychologie, Band 9, Heft 3, 201-213.

Schuler, H. (1989). Die Validität des Assessment Centers. In C. Lattmann (Hrsg.), Das Assessment-Center Verfahren der Eignungsbeurteilung. Sein Aufbau, seine Anwendung und sein Aussagegehalt, 223 – 250, Heidelberg: Physica-Verlag.

Schuler, H. (1990). Personenauswahl aus der Sicht der Bewerber: Zum Erleben eignungsdiagnostischer Situationen. Zeitschrift für Arbeits- und Organisationspsychologie, 34, 184-191.

Schuler, H. (1992a). Das Multimodale Einstellungsinterview. Diagnostica, 38, 281-300.

Schuler, H. (1992b). Assessment-Center als Auswahl und Entwicklungsinstrument: Einleitung und Überblick. In H. Schuler & W. Stehle, Assessment-Center als Methode der Personalentwicklung, Beiträge zur Organisationspsychologie. 2. Auflage. Göttingen, Stuttgart: Verlag für Angewandte Psychologie.

Schuler, H. (1996). Psychologische Personalauswahl. Einführung in die Berufseignungsdiagnostik. Göttingen: Verlag für Angewandte Psychologie.

Schuler, H. (1998). Noten und Studien- und Berufserfolg. In D.H. Rost (Hrsg.), Handwörterbuch Pädagogische Psychologie, 370-374. Weinheim: Psychologische Verlags Union.

Schuler, H. (2001). Arbeits- und Anforderungsanalyse. In H. Schuler (Hrsg.), Lehrbuch der Personalpsychologie. Göttingen: Hogrefe.

Schuler, H. & Funke, U. (1989). Berufseignungsdiagnostik. In E. Roth (Hrsg), Organisationspsychologie, Enzyklopädie der Psychologie D/III/3, 281-320. Göttingen: Hogrefe.

Schuler, H. & Funke, U. (1991). Eignungsdiagnostik in Forschung und Praxis. Stuttgart: Verlag für Angewandte Psychologie.

Schuler, H. & Funke, U. (1993). Diagnose beruflicher Eignung und Leistung. In H. Schuler (Hrsg.), Lehrbuch der Organisationspsychologie. Bern: Huber.

Schuler, H. & Höft, S. (2001). Konstruktorientierte Verfahren der Personalauswahl. In H. Schuler (Hrsg.), Lehrbuch der Personalpsychologie. Göttingen: Hogrefe.

Schuler, H. & Marcus, B. (2001). Biographieorientierte Verfahren der Personalauswahl. In H. Schuler (Hrsg.), Lehrbuch der Personalpsychologie. Göttingen: Hogrefe.

Schuler, H. & Moser, K. (1995). Geschichte der Management-Diagnostik. In W. Sarges (Hrsg.), Management-Diagnostik, 32 - 42. Göttingen: Hogrefe.

Schuler, H. & Stehle, W. (1983). Soziale Validität eignungsdiagnostischer Verfahren: Anforderungen für die Zukunft. In H. Schuler & W. Stehle (Hrsg.), Organisationspsychologie und Unternehmenspraxis: Perspektiven der Kooperation. Stuttgart: Hogrefe.

Schuler, H. & Stehle, W. (1992). Assessment Center als Methode der Personalentwicklung. Stuttgart: Verlag für Angewandte Psychologie.

Schuler, H. Funke, U., Moser, K. & Donat, M. (1995). Personalauswahl in Forschung und Entwicklung. Eignung und Leistung von Wissenschaftlern und Ingenieuren. Göttingen: Hogrefe.

Schuler, H. (Hrsg.) (2001). Lehrbuch der Personalspsychologie. Göttingen: Hogrefe.

Schweizer, K. (Hrsg.) (1999). Methoden für die Analyse von Fragebogendaten. Göttingen: Hogrefe.

Seibt, H. & Kleinmann, M. (1991). Personalvorauswahl von Bewerbern: Derzeitiger Stand und Alternativen. In H. Schuler & U. Funke (Hrsg.), Eignungsdiagnostik in Forschung und Praxis, 174-177.Göttingen: Hogrefe.

Semmer, N. & Udris, I. (1995). Bedeutung von Arbeit. In H. Schuler, (Hrsg.), Lehrbuch der Organisationspsychologie. Bern: Huber.

Starch, D. & Elliot, E. (1913). Reliability of grading work in mathematics. School review, Vol. 21, 254 – 259.

Steck, P. (1997). Aus der Arbeit des Testkuratoriums: Psychologische Testverfahren in der Praxis. Ergebnisse einer Umfrage unter Testanwendern, Diagnostica, 43, 3, 267 – 284.

Stehle, W. (1995). Biographische Fragebogen. In W. Sarges (Hrsg.), Managementdiagnostik. Göttingen: Hogrefe.

Stehle, W. & Barthel, E. (1984). Lohnen sich psychologische Auswahlverfahren? Ein Nutzen-Kosten Analyse. Personalwirtschaft, 11/84, 381-386. Göttingen: Hogrefe.

Tent, L. (1998). Zensuren. In D.H. Rost (Hrsg.). Handwörterbuch Pädagogische Psychologie. Weinheim: PVU.

Thornton, G.C., Gaugler, B.B., Rosenthal, D.B. & Bentson, C. (1992). Die prädiktive Validität des Assessment Centers - eine Metaanalyse. In Schuler, H. & Stehle, W. (Hrsg.), Assessment Center als Methode der Personalentwicklung. Stuttgart: Verlag für Angewandte Psychologie.

Ulich, E. (1991). Arbeitspsychologie. Stuttgart: Poeschel.

Weinert, A.B. (1998). Organisationspsychologie. Ein Lehrbuch (4.Auflage). Weinheim: Psychologie Verlagsunion.

Weiss, R. (1989a). Leistungsbeurteilung in den Schulen. Notwendigkeit oder Übel? Problemanalysen und Verbesserungsvorschläge. Wien: Jugend und Volk Verlag.

Weiss, R. (1989b). Die Zuverlässigkeit der Ziffernbenotung bei Aufsätzen und Rechenarbeiten. In K. Ingenkamp (Hrsg.), Die Fragwürdigkeit der Zensurengebung. Texte und Untersuchungsberichte. Weinheim: Beltz Verlag.

Westhoff, K. (1992) (Band-Hrsg.). Entscheidungsorientierte Diagnostik. Bonn: Deutscher Psychologen Verlag GmbH.

Westhoff, K. (2000). Das psychologisch-diagnostische Interview. Report Psychologie, 25, 18-24.

Wiesner, W.H. & Cronshaw, S.F. (1988). A meta-analytic investigation of the impact of interview format and degree of structure on the validity of the employment interview. Journal of Occupational Psychology, 72, 484-487.

Wittmann, W.W. & Matt, G.E. (1986). Aggregation und Symmetrie. Grundlagen einer multivariaten Reliabilitäts- und Validitätstheorie, dargestellt am Beispiel der differentiellen Validität des Berliner Intelligenzstrukturmodells. Diagnostica, 32, 309 – 329.

Wolff, P. & Voullaire, C. (1968). Eignungsbegutachtung von Körperbehinderten für einen Verwaltungsberuf. Eine Bewährungskontrolle. Diagnostica, 14, 70 – 87.

Wottawa, H. & Thierau, H. (**1998)** Lehrbuch der Evaluation (2.Auflage). Bern: Verlag Hans Huber.

9. Anhang

1. Nicht dargestellte deskriptive Daten der Stichprobe

Alter am Prüfungstag

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	17	77	10,5	10,5	10,5
	18	180	24,5	24,5	35,0
	19	145	19,7	19,7	54,7
	20	86	11,7	11,7	66,4
	21	49	6,7	6,7	73,1
	22	32	4,4	4,4	77,4
	23	23	3,1	3,1	80,5
	24	18	2,4	2,4	83,0
	25	12	1,6	1,6	84,6
	26	20	2,7	2,7	87,3
	27	8	1,1	1,1	88,4
	28	24	3,3	3,3	91,7
	29	15	2,0	2,0	93,7
	30	11	1,5	1,5	95,2
	31	13	1,8	1,8	97,0
	32	9	1,2	1,2	98,2
	33	5	,7	,7	98,9
	34	3	,4	,4	99,3
	35	2	,3	,3	99,6
	37	2	,3	,3	99,9
	41	1	,1	,1	100,0
	Gesamt	735	100,0	100,0	

Statistiken

		Schulnote Deutsch	Schulnote Mathematik	Englisch	Schulnote Geschichte, Wirtschaft, Politik	Schulnote Sport	Notendurchschnitt
N	Gültig	682	676	671	677	628	689
	Fehlend	53	59	64	58	107	46

Schulnote Deutsch

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1	9	1,2	1,3	1,3
	2	174	23,7	25,5	26,8
	3	332	45,2	48,7	75,5
	4	163	22,2	23,9	99,4
	5	4	,5	,6	100,0
	Gesamt	682	92,8	100,0	
Fehlend	99	53	7,2		
Gesamt		735	100,0		

Schulnote Mathematik

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1	32	4,4	4,7	4,7
	2	121	16,5	17,9	22,6
	3	238	32,4	35,2	57,8
	4	237	32,2	35,1	92,9
	5	48	6,5	7,1	100,0
	Gesamt	676	92,0	100,0	
Fehlend	99	59	8,0		
Gesamt		735	100,0		

Schulnote Englisch

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1	18	2,4	2,7	2,7
	2	120	16,3	17,9	20,6
	3	287	39,0	42,8	63,3
	4	227	30,9	33,8	97,2
	5	19	2,6	2,8	100,0
	Gesamt	671	91,3	100,0	
Fehlend	99	64	8,7		
Gesamt		735	100,0		

Schulnote Geschichte, Wirtschaft, Politik

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1	33	4,5	4,9	4,9
	2	225	30,6	33,2	38,1
	3	323	43,9	47,7	85,8
	4	95	12,9	14,0	99,9
	5	1	,1	,1	100,0
	Gesamt	677	92,1	100,0	
Fehlend	99	58	7,9		
Gesamt		735	100,0		

Schulnote Sport

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1	209	28,4	33,3	33,3
	2	290	39,5	46,2	79,5
	3	120	16,3	19,1	98,6
	4	9	1,2	1,4	100,0
	Gesamt	628	85,4	100,0	
Fehlend	99	107	14,6		
Gesamt		735	100,0		

Deskriptive Statistik

	N	Minimum	Maximum	Mittelwert	Standardabweichung
Punkte Diktat	735	1	9	5,40	1,64
Punkte Bericht Inhalt	734	1	9	5,24	1,69
Punkte Bericht Deutschleistung	734	1	9	5,90	2,19
Punkte kulturfair L.	721	1	9	7,12	1,62
Punkte Sport	663	1,0	14,2	4,703	1,428
Punkte nicht kulturf. L. Gesamtbewertung	435	8	33	18,94	4,72
Punkte Vorstellungsgespräch	207	,00	13,75	7,2874	2,6475
Punkte Rundgespräch	207	,00	12,25	6,9783	2,6317
Endergebnis Punkte	735	1,0	115,2	48,474	25,615
Gültige Werte (Listenweise)	207				

2. Verteilung der Schulnoten in der Stichprobe

Notendurchschnitt

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1,2	2	,3	,3	,3
	1,3	1	,1	,1	,4
	1,4	6	,8	,9	1,3
	1,6	15	2,0	2,2	3,5
	1,7	1	,1	,1	3,6
	1,8	1	,1	,1	3,8
	1,8	15	2,0	2,2	6,0
	2,0	31	4,2	4,5	10,4
	2,2	38	5,2	5,5	16,0
	2,3	6	,8	,9	16,8
	2,3	1	,1	,1	17,0
	2,4	67	9,1	9,7	26,7
	2,5	11	1,5	1,6	28,3
	2,6	79	10,7	11,5	39,8
	2,8	15	2,0	2,2	41,9
	2,8	90	12,2	13,1	55,0
	3,0	92	12,5	13,4	68,4
	3,2	70	9,5	10,2	78,5
	3,3	17	2,3	2,5	81,0
	3,3	2	,3	,3	81,3
	3,4	60	8,2	8,7	90,0
	3,5	14	1,9	2,0	92,0
	3,6	32	4,4	4,6	96,7
	3,7	2	,3	,3	97,0
	3,8	3	,4	,4	97,4
	3,8	11	1,5	1,6	99,0
	4,0	6	,8	,9	99,9
	4,2	1	,1	,1	100,0
	Gesamt	689	93,7	100,0	
Fehlend	99,0	46	6,3		
Gesamt		735	100,0		

Deskriptive Statistik

	N	Minimum	Maximum	Mittelwert	Standardabweichung
Schulnote Deutsch	682	1	5	2,97	,76
Schulnote Mathematik	676	1	5	3,22	,98
Schulnote Englisch	671	1	5	3,16	,84
Schulnote Geschichte, Wirtschaft, Politik	677	1	5	2,71	,77
Schulnote Sport	628	1	4	1,89	,75
Notendurchschnitt	689	1,2	4,2	2,806	,542
Gültige Werte (Listenweise)	570				

3. Korrelationen der Schulnoten

Korrelationen

		Notendurchschnitt	Schulnote Deutsch	Schulnote Mathematik	Schulnote Englisch	Schulnote Geschichte Politik Wirtschaft	Schulnote Sport	Gesamtnote aus allen Fächern	Gesamtpunkte aus allen Fächern
Notendurchschnitt	Korrelation nach Pe	1,000	,651**	,642**	,650**	,655**	,494**	,288**	-,283**
	Signifikanz (2-seitig	,	,000	,000	,000	,000	,000	,001	,002
	N	121	119	119	119	121	114	121	121
Schulnote Deutsch	Korrelation nach Pe	,651**	1,000	,169	,421**	,361**	,258*	,200*	-,212*
	Signifikanz (2-seitig	,000	,	,069	,000	,000	,006	,029	,021
	N	119	119	117	118	119	113	119	119
Schulnote Mathem	Korrelation nach Pe	,642**	,169	1,000	,219*	,303**	,117	,261**	-,265**
	Signifikanz (2-seitig	,000	,069	,	,018	,001	,219	,004	,004
	N	119	117	119	117	119	112	119	119
Schulnote Englisch	Korrelation nach Pe	,650**	,421**	,219*	1,000	,243**	,093	,103	-,065
	Signifikanz (2-seitig	,000	,000	,018	,	,008	,327	,265	,484
	N	119	118	117	119	119	113	119	119
Schulnote Geschic Politik Wirtschaft	Korrelation nach Pe	,655**	,361**	,303**	,243**	1,000	,188*	,323**	-,303**
	Signifikanz (2-seitig	,000	,000	,001	,008	,	,045	,000	,001
	N	121	119	119	119	121	114	121	121
Schulnote Sport	Korrelation nach Pe	,494**	,258*	,117	,093	,188*	1,000	-,052	,009
	Signifikanz (2-seitig	,000	,006	,219	,327	,045	,	,582	,922
	N	114	113	112	113	114	114	114	114
Gesamnote aus all Fächern	Korrelation nach Pe	,288**	,200*	,261**	,103	,323**	-,052	1,000	-,970**
	Signifikanz (2-seitig	,001	,029	,004	,265	,000	,582	,	,000
	N	121	119	119	119	121	114	127	127
Gesamtpunkte aus Fächern	Korrelation nach Pe	-,283**	-,212*	-,265**	-,065	-,303**	,009	-,970**	1,000
	Signifikanz (2-seitig	,002	,021	,004	,484	,001	,922	,000	,
	N	121	119	119	119	121	114	127	127

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

*. Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.

4. Gewichtung der Durchschnittnote

Aufgenommene/Entfernte Variablen[b]

Modell	Aufgenommene Variablen	Entfernte Variablen	Methode
1	Schulnote Sport, Schulnote Geschichte, Wirtschaft, Politik, Schulnote Englisch Schulnote Mathematik, Schulnote Deutsch[a]	,	Eingeben

a. Alle gewünschten Variablen wurden aufgenommen.

b. Abhängige Variable: Notendurchschnitt Deutsch,Mathe,Englisch,GPW,Sport

Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,683[a]	,467	,462	,399

a. Einflußvariablen : (Konstante), Schulnote Sport, Schulnote Geschichte, Wirtschaft, Politik, Schulnote Englisch, Schulnote Mathematik, Schulnote Deutsch

ANOVA[b]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
1	Regression	78,580	5	15,716	98,821	,000[a]
	Residuen	89,696	564	,159		
	Gesamt	168,276	569			

a. Einflußvariablen : (Konstante), Schulnote Sport, Schulnote Geschichte, Wirtschaft, Politik, Schulnote Englisch, Schulnote Mathematik, Schulnote Deutsch

b. Abhängige Variable: Notendurchschnitt Deutsch,Mathe,Englisch,GPW,Sport

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz
		B	Standardfehler	Beta		
1	(Konstante)	1,418	,105		13,448	,000
	Schulnote Deutsch	2,212E-02	,025	,031	,881	,379
	Schulnote Mathematik	-1,27E-02	,018	-,023	-,696	,487
	Schulnote Englisch	1,781E-02	,022	,028	,793	,428
	Schulnote Geschichte, Wirtschaft, Politik	,483	,022	,682	22,077	,000
	Schulnote Sport	-5,43E-03	,023	-,008	-,239	,811

a. Abhängige Variable: Notendurchschnitt Deutsch,Mathe,Englisch,GPW,Sport

5. Korrelation der Testbausteine

Korrelationen

		Punkte Dik	Punkte Bericht Deutschleistung	Punkte Bericht Inh	Punkte Kulturfaer Leistungtest	Punkte Sp	Punkte Nicht Kulturfaer Leistungtest	Punkte Vorstellungsgespräc	Punkte Rundgespr h	Endergebr
Punkte Diktat	Korrelation nach Pe	1,000	,466	,137	,218	,018	,276	,032	-,045	,394
	Signifikanz (2-seitig	,	,000	,000	,000	,648	,000	,649	,523	,000
	N	735	734	734	721	663	435	207	207	735
Punkte Bericht Deutschleistung	Korrelation nach Pe	,466	1,000	,336	,190	-,029	,160	,103	,103	,387
	Signifikanz (2-seitig	,000	,	,000	,000	,452	,001	,141	,140	,000
	N	734	734	734	721	663	435	207	207	734
Punkte Bericht Inh	Korrelation nach Pe	,137	,336	1,000	,076	,026	,048	,077	,064	,221
	Signifikanz (2-seitig	,000	,000	,	,042	,507	,315	,272	,359	,000
	N	734	734	734	721	663	435	207	207	734
Punkte Kulturfairer Leistungstest	Korrelation nach Pe	,218	,190	,076	1,000	,102	,449	,013	-,010	,481
	Signifikanz (2-seitig	,000	,000	,042	,	,008	,000	,849	,881	,000
	N	721	721	721	721	663	435	207	207	721
Punkte Sport	Korrelation nach Pe	,018	-,029	,026	,102	1,000	-,007	-,017	,038	,503
	Signifikanz (2-seitig	,648	,452	,507	,008	,	,893	,808	,586	,000
	N	663	663	663	663	663	432	207	207	663
Punkte Nicht Kultu Leistungstest	Korrelation nach Pe	,276	,160	,048	,449	-,007	1,000	,131	,095	,747
	Signifikanz (2-seitig	,000	,001	,315	,000	,893	,	,061	,175	,000
	N	435	435	435	435	432	435	207	207	435
Punkte Vorstellungsgespr	Korrelation nach Pe	,032	,103	,077	,013	-,017	,131	1,000	,748	,824
	Signifikanz (2-seitig	,649	,141	,272	,849	,808	,061	,	,000	,000
	N	207	207	207	207	207	207	207	207	207
Punkte Rundgespr	Korrelation nach Pe	-,045	,103	,064	-,010	,038	,095	,748	1,000	,779
	Signifikanz (2-seitig	,523	,140	,359	,881	,586	,175	,000	,	,000
	N	207	207	207	207	207	207	207	207	207
Endergebnis	Korrelation nach Pe	,394	,387	,221	,481	,503	,747	,824	,779	1,000
	Signifikanz (2-seitig	,000	,000	,000	,000	,000	,000	,000	,000	,
	N	735	734	734	721	663	435	207	207	735

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

*. Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.

Lebenslauf

Name	**Isabel Talea Petersen**, geb. Engels	
Geburtsdatum, -ort	16.04.1973 in Emden, Kreis Ostfriesland	
Familienstand	Verheiratet, zwei Kinder	
Schulausbildung	1979 – 1983	Grundschule Dornholzhausen, Bad Homburg
	1983 – 1984	Orientierungsstufe Gymnasium Bad Homburg
	1984 - 1985	Orientierungsstufe Gymnasium Cuxhaven
	1985 - 1992	Lichtenberggymnasium Cuxhaven, Abschluss: Allgemeine Hochschulreife
Studium	1992 – 1998	Studium der Psychologie an der CAU zu Kiel Abschluss der Diplomprüfungen am 25.05. 1998 Schwerpunktfächer: Klinische Psychologie Pädagogische Psychologie
Berufserfahrung	01.09.1998 - 11.12.1999	Sozialpädagogische Mitarbeiterin in der Kindervilla im August-Lütgens-Park, Hamburg
	13.12.1999 - zurzeit	Erziehungsurlaub