RUPRECHT-KARLS-UNIVERSITÄT-HEIDELBERG

Konstruktion und Evaluation eines Studierendenauswahlverfahrens für Psychologie an der Universität Heidelberg

Moritz Heene

Dissertation

zur Erlangung des akademischen Grades eines

Dr. phil.

der Fakultät für Verhaltens- und Empirische

Kulturwissenschaften

der Ruprecht-Karls-Universität Heidelberg

Tag der Disputation: 20.07.2007

Erstgutachter: Prof. Dr. Manfred Amelang

Zweitgutachter: Prof Dr. Joachim Werner

Hiermit erkläre ich, Moritz Heene, dass ich die vorliegende Dissertation selbstständig angefertigt, nur die angegebenen Hilfsmittel benutzt und die Zitate gekennzeichnet habe. Die vorliegende Dissertation wurde in dieser oder anderer Form noch nicht als Prüfungsarbeit verwendet oder einer anderen Fakultät als Dissertation vorgelegt.

Heidelberg, den 11. Dezember 2006 Moritz Heene

DANKSAGUNG

Vielen habe ich zu danken. An erster Stelle möchte ich mich besonders bei meiner Freundin Kirstin Mayser bedanken, die mich immer unterstützt hat, auch wenn ich nicht gerade selten wegen der Arbeit an der Dissertation Symptome von „dissertalem Autismus“ zeigte. Ein sehr großer Dank gilt meinem Betreuer Prof. Dr. Manfred Amelang zum einen für die ausgezeichnete fachliche Unterstützung zum Thema Studierendenauswahl, zum anderen für die für mich nötige „Narrenfreiheit“ bei der Bearbeitung des Themas und nicht zuletzt für meine Zeit am Heidelberger Psychologischen Institut, die er überhaupt erst ermöglichte. Hr. Prof. Dr. Joachim Werner möchte ich ebenfalls für seine statistisch-methodische Unterstützung danken, die mir so manche Schuppen von den Augen fallen ließ. Untrennbar mit dieser Danksagung verbunden ist ebenso meine geschätzte Kollegin Dr. Ricarda Steinmayr. Mein Dank betrifft sowohl ihre fachliche Hilfe, insbesondere bei der Anforderungsanalyse, dem Korrekturlesen und weiterer inhaltlicher Ratschläge, als auch die menschliche Unterstützung während der gesamten Zeit. Hr. Dipl.-Psych Samuel Greiff möchte ich ebenfalls herzlich für den fachlichen Austausch und das Korrekturlesen danken. Ein besonderer Dank gilt auch Hr. Prof. Dr. Peter Schönemann, der mir den etablierten Boden unter den Füßen wegzog und mich wieder laufen lehrte. Ihm widme ich diese Arbeit.

Herzlich bedanken möchte ich mich auch bei Fr. Dipl.-Psych. Birgit Koopmann und Hr. cand. psychol. Philipp Wollscheid für die Hilfe bei der Dateneingabe und der besonders wertvollen und mühevollen Beurteilung freier Antworten einiger Testaufgaben.

Dank gilt auch schließlich allen Studienteilnehmern, die an dieser Studie teilgenommen haben, wie auch dem Forschungsdezernat der Universität Heidelberg, welches mich großzügig mit Versuchspersonengeldern ausgestattet hatte.

0. INHALTSVERZEICHNIS

1. Einleitung

„Bekanntlich gibt es Menschen, denen durch, sei
es angeboren, sei es durch äußere Erziehung
erweckte Anlage, ihr kunftiger [sic!] Beruf von früher
Jugend an vor Augen steht. Das sind die Glücklichen,
die unbekümmert ihre Straße wandern können,
auch wenn sie sonst mancherlei Hindernissen
begegnen .... Ich gehöre nicht zu diesen Glücklichen“
(Wundt, 1920, S. 57, zit. nach Oldfield, 1939).

Die Frage nach der passenden Wahl eines Studiums scheint nicht erst eine heutiger Zeit zu sein, sondern beschäftigte offensichtlich auch die Urväter der modernen Psychologie wie Wilhelm Wundt. Spiegelt sich in obigem Zitat die Frage des *Individuums* nach der richtigen Wahl eines Berufes oder Studiengangs wider, so stellt sich in heutiger Zeit seitens der *Hochschulen* zunehmend auch die Frage nach der Auswahl geeigneter Studierender. Dies nicht zuletzt vor dem Hintergrund von drei Jahrzehnten Bildungsentwicklung in Deutschland, mit ihrer seit dem Ende der 1970er Jahre fast verdoppelten Studierendenzahl (Statistisches Bundesamt, 2003) und den daraus resultierenden stark begrenzten Hochschulkapazitäten. Es liegt daher in der Natur der Sache, dass eine Hochschule bestrebt ist, die limitierte Anzahl an Studienplätzen an möglichst geeignete Personen zu vergeben. Seit der Reform des Hochschulrahmengesetzes (Reich, 2002) sind zudem die Hochschulen gefordert, selbst aktiv bei der Hochschulzulassung mitzuwirken, um so eine bessere Übereinstimmung der Qualifikationsprofile von Studienbewerbern mit den Anforderungen einzelner Studienfächer zu erzielen. In diesem Zusammenhang wird daher in letzter Zeit der Ruf nach dem Einsatz von Verfahren zur Identifizierung von geeigneten Studienbewerbern im tertiären Bildungssektor immer dringlicher.

Auch die vorliegende Arbeit ist in diesem Zusammenhang zu sehen. Sie beschäftigt sich mit der Konstruktion und ersten Evaluationsschritten eines Studieneingangsverfahren für Studienplatzbewerber im Fach Psychologie an der Universität Heidelberg. Der Rahmen dieser Arbeit umschließt dabei die einzelnen inhaltlichen wie psychometrischen Konstruktionsschritte des

Verfahrens sowie erste kriteriumsbezogene Validitätsuntersuchungen der Testverfahren und der Abiturdurchschnittsnote als bislang vorrangiges Hochschulzulassungkriterium.
Kapitel 2 erläutert grundlegende Aspekte des deutschen Hochschulzugangs. Im Anschluss folgt in Kapitel 4 bis 4.5 ein Überblick über den derzeitigen Forschungsstand der Studieneignungsdiagnostik im Hinblick auf nationale und internationale Forschung. Hieraus werden in Kapitel 5 die Fragestellungen und Hypothesen abgeleitet. Mit Kapitel 8 beginnt die Darstellung des empirischen Teils dieser Arbeit nebst theoretischer und psychometrischer Hintergründe zu den Skalenanalysen (Kapitel 9 und 10). Die Kapitel 11 bis 18 stellen die empirischen Ergebnisse dieser Arbeit dar, um in Kapitel 19 die Resultate im Hinblick auf den bisherigen Forschungsstand zu diskutieren und einen Ausblick zu geben. Kapitel 20 schließt die Arbeit mit einer Zusammenfassung der Ergebnisse ab.

2. Hochschulzugang in Deutschland: rechtliche und politische Rahmenbedingungen

Die vergangenen drei Jahrzehnte deutscher Hochschulentwicklung sind geprägt von einer „bis dahin historisch nicht gekannte[n] soziale[n] Bildungsexpansion...“ (Bultmann, 2001, S. 10). So begrüßenswert die Ausweitung der Bildungschancen gesamtgesellschaftlich betrachtet auch sein mag, so schlecht vorbereitet sahen und sehen sich die Hochschulen angesichts der daraus resultierenden Probleme. Betrachtet man auf der einen Seite die seit dem Bestehen der Bundesrepublik sich mehr als verzwanzigfachte Zahl an Studierenden (Statistisches Bundesamt, 2005) bzw. die seit dem Ende der 1970er Jahre fast verdoppelte Studierenden-anzahl (Statistisches Bundesamt, 2003), so sieht man demgegenüber auf der Seite der Hochschulausgaben eine Stagnation. So stieg die Anzahl wissenschaftlicher Stellen seit Ende der 1970er Jahre lediglich um 10 Prozent, der Anteil der Hochschulausgaben am Brutto-sozialprodukt reduzierte sich sogar von 1.3 auf 0.9 Prozent (Bultmann, 2001, S. 11). Im internationalen Vergleich liegt Deutschland damit unterhalb des Durchschnitts, wobei in anderen Ländern die Investitionen sogar noch ansteigen (Egeln & Heine, 2005). Die augen-scheinlichsten Merkmale dieses Missverhältnisses von Anforderungen an die Hochschulen und deren Möglichkeiten manifestieren sich bspw. in überfüllten Hörsälen, unzureichenden Betreuungsrelationen, Streichung von Studiengängen, hohen Studienabbruchquoten und Langzeitstudierenden (Deidesheimer Kreis, 1997, S. 9; Dierkes & Merkens, 2002; Hörner, 1999). Weniger augenscheinlich, aber in den Konsequenzen noch weitreichender, muss man die Probleme im Kontext globalisierter Arbeits- und Bildungsmärkte sehen. Internationale Studien belegen den Zusammenhang zwischen Bildungskapital und Wirtschaftswachstum (OECD, 2003, S. 201ff.). Nach Dierkes und Merkens (2002, S. 8) haben „nur die Nationen und Regionen, die in die Wissensbasis ihrer Bevölkerung investieren ... eine Chance, in diesem Rennen unter den Gewinnern zu sein .... Investitionen in das Humankapital sind damit ein Schlüsselfaktor im immer intensiveren und globaleren Wettbewerb“. Vor diesem Hintergrund wurde der Ruf nach Reformen des Bildungssystems immer lauter. Deutlich tritt in den damit verbundenen Debatten die Forderung nach einer Erweiterung der Autonomie und des Auswahlrechts der Hochschulen in den Vordergrund. (Deidesheimer Kreis, 1997; Dierkes & Merkens, 2002; Fedrowitz & Zempel, 1996; Wissenschaftsrat, 2004). Meist wird dies mit dem Argument verknüpft, eine bessere Passung zwischen der jeweiligen Hochschule und den Studierenden im Sinne des Person-Job-Fit-Ansatzes (Amelang, 1997) zu erreichen. Die hierbei wesentlichen Aspekte der Passung betreffen die Perspektive des Individuums (Studienbewerber) als auch der Insitution (Hochschule). Seitens des Studienbewerbers führe

dies zu einem Mehr an Informationen über die angestrebte Ausbildung, die zu bewältigenden spezifischen Anforderungen des Studienfaches und somit sehr wahrscheinlich zu einer fundierten Entscheidung in Bezug auf das Studium und/oder die Universität (Amelang & Funke, 2005; Deidesheimer Kreis, 1997, S. 164f.). Seitens der Hochschule resultiere eine bessere Kapazitätsauslastung wegen weniger „falsch positiv“ zugeordneter Studienbewerber und höherer Erfolgsquoten. Durch Studieneignungstests könne somit die Anzahl an Studienabbrechern, Langzeitsstudierenden und Prüfungswiederholern reduziert werden (Deidesheimer Kreis, 1997, S. 164f.). Die Hochschulrektorenkonferenz (2004) fasst die Ziele der autonomen Studierendenauswahl daher wie folgt zusammen:

a) Profilbildung der Hochschulen und nationaler bzw. internationaler Wettbewerb zwischen den Hochschulen
b) Qualitätssteigerung in Studium und Lehre, um Studienabbruchquoten zu verringern
c) Verbesserung der Passung zwischen Studienplatz und -bewerber
d) Beifolgende Überprüfung der spezifischen Studierfähigkeit
e) Förderung der Entscheidungssicherheit des Studienplatzbewerbers durch beratende Funktion

Mittlerweile erlaubt auch die Gesetzeslage durch eine Novellierung des Hochschulrahmengesetzes (HRG), dass bei Studiengängen, in „denen nach Feststellung der Zentralstelle zu erwarten ist, dass im allgemeinen Auswahlverfahren [d.h. über Schulabschlussnote bzw. Wartezeit] die Auswahl ... zu unvertretbar hohen Anforderungen an den Grad der Qualifikation ... führen würde, ... an die Stelle des allgemeinen Auswahlverfahrens ... ein besonderes Auswahlverfahren treten [soll]“ (Reich, 2002, S. 304). In bundesweit zulassungsbeschränkten Studiengängen sind zudem im Rahmen einer sog. 20-20-60-Regelung nun Studierendenauswahlverfahren möglich. Dies bedeutet eine Vergabe der Studienplätze anhand entsprechender Quoten: 20% nach Abiturnote, weitere 20% nach Wartezeit und 60% nach Maßgabe eines von den Hochschulen selbst entwickelten Auswahlverfahrens. Für die Feststellungsverfahren wesentlich ist, dass in diesen keine Kenntnisse der Hochschulzugangsberechtigung abgefragt werden dürfen, wie Reich (2002) ausführt:

> Im Feststellungsverfahren sollen grundsätzlich nicht die Kenntnisse festgestellt werden, die bereits Gegenstand der Bewertung in der Hochschulzugangsberechtigung sind; es soll den Bewerberinnen und Bewerbern insbesondere Gelegenheit geben, in

> den bisherigen Abschlüssen nicht ausgewiesene Fähigkeiten und Kenntnisse nachzuweisen, die für den Studienerfolg von Bedeutung sein können .... Zu diesem Zweck können insbesondere entsprechende Testverfahren durchgeführt werden. Das Feststellungs-Verfahren ist ... einheitlich zu gestalten. (S. 305)

Damit ist allerdings die Art des Feststellungsverfahrens lediglich grob umrissen, was einigen Interpretationsspielraum erlaubt. Allerdings können die angestrebten positiven Aspekte der größeren Autonomie in der Studierendenauswahl nur dann erreicht werden, wenn die psychometrische Qualität der Verfahren als notwendige Grundvoraussetzung gegeben ist (Deutsches Institut für Normung, 2002; Westmeyer, 2005; Wissenschaftsrat, 2004). Damit ist zugleich die Rolle der Psychologie in der Entwicklung und Evaluation von Auswahlverfahren zentral benannt. In der vorliegenden Arbeit wird daher ein eigens für den Studiengang Psychologie konstruiertes Verfahren evaluiert. Im Fokus stehen hierbei die Umsetzung spezifischer Anforderungen des Psychologiestudiums in geeignete Testverfahren, deren psychometrische Beurteilung nach Maßgabe der Rasch-Modelle sowie hinsichtlich retrospektiver und prädiktiver Validität gegenüber Maßen des Studienerfolges im Vergleich zur Abiturdurchschnittsnote.

3. Vorhersage des Studienerfolgs

Jede Theorie zur Vorhersage des Studienerfolgs muss zunächst diejenigen Annahmen benennen, auf deren Basis sie eignungsdiagnostische Urteile fällt. Die bisherige Eignungsdiagnostik im Rahmen der Studierendenauswahl bezieht sich hierbei im Wesentlichen auf die Annahmen der klassischen Trait-Diagnostik, wie sie der Deidesheimer Kreis (1997, S. 78f.) in seinen Grundannahmen formuliert hat:

a) Es gibt eine Reihe von Merkmalen, die in unterschiedlicher Weise für den Erfolg in verschiedenen Studiengängen wichtig sind.
b) Die Eignungsmerkmale sind bei den Studienbewerbern in unterschiedlichem Maße stark ausgeprägt.
c) Der jeweilige Ausprägungsgrad lässt sich mithilfe entsprechender diagnostischer Instrumente abschätzen.

d) Die Eignungsmerkmale sind relativ überdauernd und erlauben mithin eine längerfristige Prognose.
e) Fähigkeiten sind Kernbestandteil der Eignungsmerkmale. Sie können sich auf verschiedenste Weise entwickelt haben bzw. erworben worden sein – auch außerhalb der Schule. Rückschlüsse auf Vorliegen und Ausprägung solcher Fähigkeiten können auch aus bereits erbrachten schulischen und außerschulischen Leistungen gezogen werden.

Im Folgenden wird der Untersuchungsgegenstand zwar auf Basis dieser Annahmen behandelt, allerdings sollen auch nicht-intellektuelle Personenmerkmale wie Persönlichkeitsmerkmale in die Darstellung einfließen, sofern sich in der wissenschaftlichen Literatur Zusammenhänge mit Studienleistungen zeigen.
Vorab werden jedoch zuerst wesentliche Begrifflichkeiten geklärt, bevor in Kapitel 3 und 4 auf die Facetten von Kriterien und Prädiktoren des Studienerfolgs eingegangen wird.

3.1 Begriffsbestimmung: Studieneignung, Studierfähigkeit und Studienerfolg

Eine allgemeine Definition des Begriffs der Eignung geben Häcker und Stapf (1998, S. 208f.), indem sie diesen beschreiben als das „Insgesamt der im Individuum liegenden Bedingungen für das Eintreten positiv bewerteter Ereignisse“. Der Deidesheimer Kreis (1997, S. 89) fasst demgemäß diesen Begriff in Bezug auf *Studieneignung* zusammen als „eine breite Palette individueller Merkmale, die sich allgemein dem kognitiven und dem motivational-affektiven Bereich zuordnen lassen“ welche für die Studienanforderungen von Belang sind. Im weitesten Sinn kann man daher Studieneignung als die Übereinstimmung der in einer Person vorhandenen kognitiven und motivational-affektiven Merkmale mit den Anforderungen des Studiums ansehen.
Demgegenüber fällt es schwerer, eine allgemeingültige Definition der *Studierfähigkeit* zu geben (Konegen-Grenier, 2001, S. 23). Die Definitionen unterscheiden sich hierbei insbesondere in der Anzahl der Begriffskomponenten. So postuliert Heldmann (1984) aus den Ergebnissen einer Umfrage an Hochschulen die folgenden fünf Dimensionen allgemeiner Studierfähigkeit:

a) Ausbildungsbereitschaft: Lern- und Leistungsbereitschaft, Interesse, Freude am Studium etc.
b) Vorhandensein elementarer Voraussetzungen für wissenschaftliches Arbeiten: schriftliches und mündliches Ausdrucksvermögen, effektive Lerntechniken, Problemlösestrategien etc.
c) Formen geistigen Tätigseins: allgemeines Denkvermögen, Abstraktionsfähigkeit, Auffassungsgabe etc.
d) Ausprägung der Persönlichkeit: Belastbarkeit, Selbstständigkeit, Motivation etc.
e) Interesse und Engagement: engagierte Haltung, Interessen außerhalb des Studiums, Kontaktfähigkeit etc.

Bereits früher identifizierte Trost (1975) demgegenüber durch eine fächerübergreifende Literaturübersicht zwei Dimensionen, allerdings mit sehr ähnlicher inhaltlicher Akzentuierung wie Heldeman (1984):

a) Intellektuelle Fähigkeiten: Schulleistungen, allgemeine Intelligenz, intellektuelle Studierfähigkeit, fachspezifische Fähigkeiten etc.
b) Nicht-intellektuelle Prädiktoren: effektive Lern- und Arbeitstechniken, Persistenz, Unabhängigkeit, emotionale Stabilität etc.

Von diesen allgemeinen Bestimmungsstücken der Studierfähigkeit muss man diejenigen für spezifische Studienfächer unterscheiden. Hier wird allerdings eine genaue Definition für die jeweiligen Studienfächer durch die unterschiedlichen Anforderungen zwischen den einzelnen Hochschulen erschwert wie überhaupt auch dadurch, dass deutsche Hochschulen bislang noch kaum fachspezifische Anforderungen expliziert haben (Wissenschaftsrat, 2004, S. 86), was allerdings auch in Bezug auf US-amerikanische Hochschulen kritisiert wird (Sternberg & Expertise, 2004).
Zwar gibt es im Rahmen von Anforderungsanalysen im Berufskontext reichhaltige Erfahrung (Schuler, 2001, S. 43ff.), doch ist eine direkte Übertragbarkeit berufsspezifischer Anforderungen auf solche des Studiums kritisch. Z.B. finden schriftliche und mündliche Leistungsüberprüfungen, wie sie typisch für den universitären Kontext sind, sehr wenige Äquivalente im Berufskontext, wo hierfür insbesondere Vorgesetztenurteile eingesetzt werden. Die Forschung zu Anforderungsanalysen für fachspezifische Studienanforderungen steht also erst am Anfang,

wozu diese Arbeit ebenso einen Beitrag leisten soll (zu weiteren Ansätzen s. u.a. Pixner, Zapf & Schüpbach, 2005; Zimmerhofer, 2003).

Ähnlich schwierig wie eine Begriffbestimmung der Studieneignung gestaltet sich eine für *Studienerfolg*, was nicht zuletzt an der sehr unterschiedlichen Operationalisierbarkeit des Begriffes liegt. Im allgemeinen Sprachgebrauch versteht man unter dem „Erreichen eines Studienziels“ den Abschluss eines einmal angefangenen Studiums mit dem Hauptexamen. Hierfür läge das Kriterium der Abbruchquote als Operationalisierungsmaß nahe. Ergänzend müssen jedoch weitere Indikatoren neben diesem rein dichotomen Indikator herangezogen werden, um die *Güte* der Studienleistung zu erfassen. Hierzu bieten sich Noten in Zwischen- oder Abschlussprüfungen an, solche in einzelnen Lehrveranstaltungen, die Studiendauer, Anzahl von Prüfungswiederholungen, Dozenten- und Kommilitonenbeurteilungen, aber ebenso Selbsteinschätzungen der Studienzufriedenheit. Allerdings stellt der Wissenschaftsrat (2004, S. 87) fest, dass „...die meisten wissenschaftlichen Untersuchungen zum Thema Studienerfolg ... darunter einen Studienabschluss mit guter Note“ verstehen. Die überwiegend leichte Verfügbarkeit dieses Erfolgskriteriums ist sicherlich der Hauptgrund für seine Verwendung. Angesicht der Fülle möglicher Erfolgskriterien gehen Rindermann und Oubaid (1999) auch unter Einbeziehung individueller, gesellschaftlicher sowie hochschulbezogener Gesichtspunkte von einem multikausalen Bedingungsgefüge für die erfolgreiche Absolvierung eines Studiums aus, wie es in Abbildung 1 dargestellt ist:

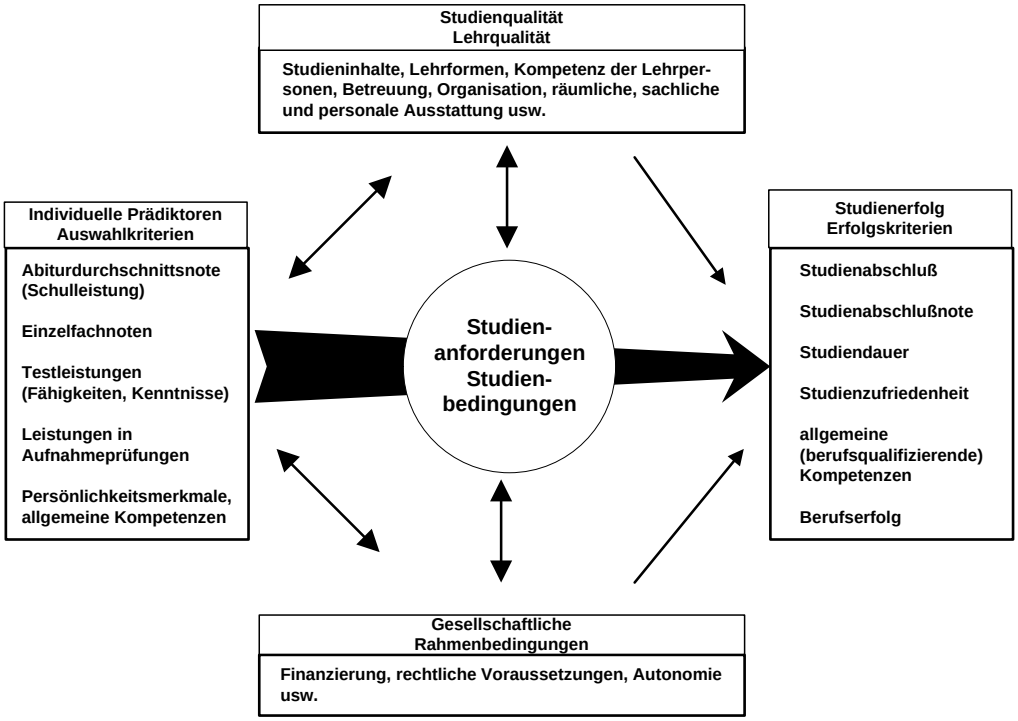


Abbildung 1: Bedingungsmodell des Studienerfolgs (nach Rindermann und Oubaid, 1999)

Die folgenden Kapitel sind der näheren Darstellung der möglichen Kriterien und Prädiktoren des Studienerfolgs gewidmet, wie sie im Modell von Rindermann und Oubaid (1999) enthalten sind.

3.2 Kriterien des Studienerfolgs

Studienerfolg ist wegen seiner multidimensionalen, facettenreichen und komplexen Natur ein schwierig hinreichend zu erfassendes Konstrukt. Trost und Bickel (1979, S. 19) verweisen in diesem Zusammenhang daher darauf, dass „Studienerfolg ... ein mehrschichtiges und facettenreiches Phänomen [ist], dem man mit Verwendung eines einzigen Index auf keinen Fall gerecht werden kann“. Einzelindikatoren als sogenannte. „aktuelle“ Kriterien sind daher stets mit den Problemen ihrer Beziehungen zu dem „eigentlichem“ Kriterium des Studienerfolgs behaftet (s. hierzu z.B. Amelang & Zielinski, 1997, S. 187f.). Im Optimalfall liegt eine Kriteriums*relevanz* als Überschneidungsbereich zwischen Indikator („aktuelles“ Kriterium) und Konstrukt („eigentliches“ Kriterium) vor. Demgegenüber kann sich jedoch schlechtestenfalls eine Kriteriums*kontamination* als Eigenständigkeit des Indikators gegenüber dem eigentlichen Kriterium ergeben, oder eine Kriteriums*defizienz* als nicht durch das aktuelle Kriterium abgebildete Aspekte des Konstruktes Studienerfolg. Erschwerend tritt hinzu, dass

die bislang untersuchten Studienerfolgsmaße wie Studienabschlussnote, Studiendauer und Studienabschluss, Studienzufriedenheit, allgemeine Kompetenzen sowie Berufserfolg lediglich mäßig miteinander korrelieren (Trost & Bickel, 1979), obgleich alle Maße bedeutsame Aspekte des Konstruktes abbilden.
Im Folgenden soll ein Überblick über objektive und subjektive Kriterien des Studienerfolgs gegeben werden, wobei jeweils deren Vor- und Nachteile dargestellt werden.

3.2.1 Studienabschluss

Das grundlegenste Studienerfolgskriterium überhaupt stellt der Abschluss eines Studiums mit einem bestandenen Examen dar (Rindermann & Oubaid, 1999, S. 175; Trost & Bickel, 1979, S. 7). Ihm gegenüber steht der Studiumsabbruch als Misserfolgskriterium. Letzterer ist allerdings von einem Studienfach- oder Hochschulwechsel sowie von einer Studienunterbrechung zu unterscheiden (Rindermann & Oubaid, 1999, S. 175). Gründe für einen Studienabbruch liegen in „wirtschaftlichen Problemen ... über organische und psychische Erkrankungen, schwere Krisen in der Partnerbeziehung, Einsicht in mangelnde Befähigung oder Eignung für das Studium bis zu einer plötzlichen Änderung der beruflichen Perspektiven“ (Trost & Bickel, 1979, S. 8). Aktuellere Untersuchungen des Hochschulinformationssystems (Lewin, 1999) korrespondieren weitgehend mit den von Trost und Bickel (1979) genannten; hier tun sich vornehmlich die folgenden Problemfelder des Studienabbruchs auf: fachliche Überforderung, Distanz zum Studium, günstige Berufschance auch ohne Examen, Wunsch nach mehr Praxisbezug in der Ausbildung, zu lange Studienzeiten, schlechte Arbeitsmarktchancen nach dem Abschluss, Kritik an der Didaktik und den Begleitumständen des Studiums, wie auch finanzielle und familiäre Gründe (Lewin, 1999, S. 20). Da die Abbruchquote jedoch im Fach Psychologie an der Universität Heidelberg mit ca. 2.5% bei jährlich 85 Studienanfängern sehr gering liegt (J. Funke, persönliche Mitteilung vom 24.4.2005), spielt dieses Kriterium in der vorliegenden Arbeit praktisch keine Rolle.

3.2.2 Hochschulnoten

Studienabschlussnoten oder Zwischenprüfungsnoten stellen das am häufigsten herangezogene Studienerfolgskriterium dar, da sie „über fachspezifische Prädiktionskraft für die Arbeitsmarktchancen [verfügen] ... und als inhaltlich valides Maß des Studienerfolgs [gelten]“

(Rindermann & Oubaid, 1999, S. 175). Gerade den Zwischenprüfungs- oder Vordiplomnoten kommt eine für den einzelnen Studierenden bedeutsame Rückmeldungsfunktion über die bislang erbrachte Leistung zu. Jedoch sind Abschlussnoten demgegenüber als höherrangiges Kriterium anzusehen, gerade weil sich in ihr der längerfristige Studienerfolg manifestiert und ihrer Bedeutung für den Arbeitsmarkt als einem Indikator der beruflichen Qualifikation.

Studiennoten sind allerdings von erheblichen psychometrischen Mängeln beschlagen. Trost (1975, S. 95ff.) verweist im Zusammenhang mit mündlichen Examensnoten auf objektivitäts-mindernde Einflüsse besonders durch unterschiedliche Bewertungsmaßstäbe, wie auch durch Vorurteile, persönliche Zu- und Abneigungen und Mängel bei der Protokollierung der Leistungen des Prüflings. Darüber hinaus hängen Prüfungsleistungen generell auch von Persönlichkeitsmerkmalen des Prüfers, des Studienfaches und der Hochschule ab (Amelang, 1974; 1978, S. 1014; Preiser, 1975; Trost, 1975, S. 95ff.). Trost und Bickel (1979, S. 12) kommen in einem Überblick über mehrere Studien daher auch zu dem Schluss, dass die Beurteilerstrenge stark variiert, unterschiedliche Maßstäbe an unterschiedlichen Hochschulen eines Faches bestehen, die Retestreliabilitäten gering ausfallen und die Bedeutung der Studienabschlussnote in unterschiedlichen Studienfächern variiert. Wie einige Autoren betonen, muss vor diesem Hintergrund mit deutlichen Validitätsbeschränkungen gerechnet werden (s. u.a. Amelang, 1978; Amelang & Hoppensack, 1977; Trost, 1975, S. 98; Trost & Bickel, 1979, S. 12). Jedoch betonen Trost und Bickel (1979, S. 12): „Dennoch wird man bei einer differenzierten Untersuchung zur Vorhersagbarkeit des Studienverhaltens auf die Erhebung dieses Kriteriums – unter anderen – nicht verzichten können“.

Trotz der psychometrischen Einwände gegenüber Studiennoten sollen sie auch in der vorliegenden Arbeit als notwendiges, wenn auch nicht hinreichendes Studienleistungs-kriterium herangezogen werden. Zum einen bieten sie nicht zuletzt wegen ihrer leichten Verfügbarkeit als Kriterium die Möglichkeit zum Ergebnisvergleich mit zahlreichen anderen Studien. Zum anderen kommt ihnen gerade wegen ihrer inhaltlichen Validität auch eine tragende Rolle in ihrer Bedeutung für die Studierenden als Rückmeldung über die bislang erbrachte Studienleistung zu.

3.2.3 Studiendauer

Die Studiendauer wurde gerade in den letzten Jahren vor dem Hintergrund deutlich von der Regelstudienzeit abweichende Studienzeiten als weiteres Studienerfolgsmerkmal gefordert und berücksichtigt (Giesen & Gold, 1996). In diesem Sinne argumentieren auch Rindermann und Oubaid (1999, S. 175), wenn sie betonen: „Ein Studium gilt dann als erfolgreich, wenn in kurzer Zeit ein qualifizierter Abschluss erreicht wurde“. Jedoch erweist sich dieses Kriterium als problematisch, da es nicht alleine die individuellen Studienleistungen abbildet, sondern zum einen ebenso die universitären Rahmenbedingungen (Studien- und Prüfungsorganisation, Besetzung von Lehrstühlen, Betreuungsverhältnis etc.), zum anderen auch private Umstände, welche Einfluss auf die Dauer der Studiendauer nehmen können wie etwa eine parallele Berufstätigkeit oder die Erziehung von Kindern. In diesem Sinne ist zu erwarten, dass sich zwischen Leistungsmaßen und Studiendauer lediglich geringe Zusammenhänge zeigen. Demgemäß finden sich auch nur geringe Korrelationen zwischen Studiendauer und Schulleistungsmaßen von durchschnittlich $r = .16$ (Rindermann & Oubaid, 1999, S. 179). Daniel (1996) weist bezüglich der mittleren Studiendauer auf die große Varianz zwischen den Universitäten hin und Giesen und Gold (1996) verweisen auf diejenige zwischen Studienfächern. Quereinsteiger, Studienfach- und Universitätswechsler verzerren zudem die Statistiken erheblich (Rindermann & Oubaid, 1999, S. 175). Hitpass, Ohlsson und Thomas (1984) konnten zudem bei individuell ungleichen formalen und intellektuellen Eingangsvoraussetzungen keine statistisch signifikanten Unterschiede der durchschnittlichen Studiendauer nachweisen. Auch dieser Befund lässt das Kriterium Studiendauer insgesamt als problematisch erscheinen, da es deutlich stärker von institutionellen Rahmenbedingungen der jeweiligen Hochschulen beeinflusst zu sein scheint. Die Brauchbarkeit der Studiendauer als Studienerfolgskriterium ist dadurch sehr begrenzt.

Aus diesen Gründen wird auch in der vorliegenden Arbeit auf die Studiendauer als Erfolgskriterium verzichtet. Zu groß erweisen sich die Einflüsse institutionaler und sozialer Rahmenbedingungen auf die Studiendauer, als dass sie als notwendiges Studienerfolgsmaß herangezogen werden könnte.

3.2.4 Subjektive Kriterien: Studienzufriedenheit

Unter Studienzufriedenheit wird im Allgemeinen die positive oder negative Bewertung des eigenen Studiums verstanden (Westermann, Heise, Spies & Trautwein, 1996). Dieses Merkmal wird insgesamt seltener als Kriterium herangezogen, da die Wahrnehmung der Studienzufriedenheit stark von den subjektiv als wichtig erachteten Studienzielen abhängt. Ohnehin ist eine eindeutige Definition von Studienzufriedenheit bislang noch nicht zu erkennen. So ermittelten Westermann, Heise, Spies und Trautwein (1996) anhand von Befragungen von Studierenden die folgenden sehr heterogenen Hauptmerkmale der Studienzufriedenheit:

a) Studienbedingungen
b) Studieninhalte
c) Lehrverhalten der Lehrenden
d) Benotung
e) Berufliche Relevanz
f) Anerkennung
g) Kontakte zu Lehrenden
h) Unterstützende Kontakte zu Kommilitonen
i) Hochschulverhalten und –politik
j) Randbedingungen des Studiums

Die Gewichtung dieser Hauptbestimmungsstücke der Studienzufriedenheit kann jedoch von Studierendem zu Studierendem stark schwanken, sodass verallgemeinerbare Aussagen hierbei schwierig sind. Gleichwohl stellen sie neben den intellektuellen Merkmalen wichtige Informationsquellen für die Art des Studien*verlaufs* dar. Indessen kommen sie in der Prognose des Studienerfolgs selten vor. Zum einen wegen ihres hohen Erhebungsaufwandes, zum anderen wegen ihres „weichen“ Charakters, der nicht zuletzt auch aus der Unbestimmtheit des Konstruktes resultiert.

Aufgrund des hohen Erhebungsaufwandes von Studienzufriedenheit, ihrer noch ungeklärten Konstruktvalidität und der starken Abhängigkeit ihrer Beurteilung von subjektiven Studienzielsetzungen und Gewichtungen ihrer Facetten, wird in der vorliegenden Arbeit auf dieses Kriterium verzichtet.

3.2.5 Zusammenfassung

„The criterion problem is like the weather; everyone talks about it, but few try to do anything about it" (Fishman & Pasanella, 1960, zit. nach Trost, 1975, S. 94). Dieses Zitat scheint auch fast 50 Jahre nach seiner Entstehung im Zusammenhang mit Studienerfolgsprognose nichts von seiner Relevanz eingebüßt zu haben. Studiennoten bzw. Zwischennoten als vorläufiges Erfolgskriterium stellen zwar das am häufigsten verwendete Kriterium dar, welches aber nur einen begrenzten Aspekt des eigentlichen Kriteriums abdeckt. Weitere Kriterien, wie der Abschluss eines Studiums überhaupt, seine Dauer, die Studienzufriedenheit und schließlich der Berufserfolg als distalstes Kriterium, stellen weiterhin wesentliche Bestimmungsstücke des Studienerfolges dar, wobei auch sie durch zahlreiche Einflussfaktoren mitbestimmt werden und dadurch keinen befriedigenden Status als Kriteriumsindikatoren erlangen. Die konzeptuelle Mehrdimensionalität von Studienerfolg, sein Facettenreichtum, aber eben auch die Kriteriumsdefizienz seiner hier dargestellten Indikatoren müssen sich zwangsläufig mindernd auf die Vorhersagekraft jedes Prädiktors auswirken. Auf eine Verbesserung der Kriteriumsreliabilität, insbesondere derjenigen von Hochschulnoten, zielen demgemäß auch die meisten Autoren ab, indem sie etwa eine größere Standardisierung von Hochschulprüfungen fordern (s. in diesem Sinne etwa Amelang, 1978). Jedoch scheint damit lediglich nur ein kleiner Problembereich der Kriteriumsproblematik angesprochen. Trost (1975, S. 94) sowie Amelang (1978) weisen nämlich ebenso auf das meist übersehene theoretische Defizit hinsichtlich der Definition von Studienerfolg hin. Selbst die hier dargestellten Indikatoren des Studienerfolges dürften in ihrer Gesamtheit lediglich einen Bruchteil dessen erfassen, was Studienerfolg ausmacht. Diesbezüglich besteht immer noch Forschungsbedarf, will man nicht weiterhin in quasi naiv-empiristischer Weise *Fakten* bzw. Korrelationen zwischen verschiedenen Variablen „*sammeln*", sondern *Theorien* des Studienerfolges *prüfen*. Auch eine multiple Kombination der hier dargestellten Kriterien zu einem facettenreichen Gesamtmaß des Studienerfolgs könnte nicht befriedigen. Sollten sich nämlich mit einem solchen Globalmaß höhere Kriteriumsvaliditäten mit den verwendeten Prädiktoren erzielen lassen, so würde dies erst die Frage aufwerfen, *warum* ein Prädiktor oder eine Prädiktorkombination mit dem Globalmaß korreliert, einmal davon abgesehen, dass ein solches Vorgehen vielmehr die Anleitung einer technischen Operation denn eine theoretisch fundierte Definition darstellt. Bereits vor nunmehr gut dreißig Jahren verwies Trost (1975) ganz in diesem Sinne auf das Ungleichgewicht zwischen statistisch-diagnostischer Verfeinerung und mangelnder Klärung dessen, was überhaupt vorhergesagt werden soll:

> Die Vernachlässigung gerade dessen, was durch den vielfältigen Einsatz diagnostischer Instrumente und mithilfe immer komplizierterer statistischer Werkzeuge vorhergesagt werden soll, kennzeichnet bis heute den Stand der pädagogischen Forschung in den angelsächsischen Ländern wie in der Bundesrepublik. (S. 94)

Es soll dabei nicht unterschlagen werden, dass das theoretisch-definitorische Defizit des Studienerfolges ebenso die vorliegende Arbeit betrifft, da auch sie lediglich auf Studiennoten als Erfolgskriterium fußt. Eine theoriegeleitete Erweiterung des Kriteriumsraumes beispielsweise um die Dimensionen „continous learning, intellectual interest, and curiosity“ oder auch breiter gefasste wie „growth, achievement, and accomplishments in cross-disciplinary skills“ wie sie etwa Camara (2005, S. 55 und S. 75) fordert, wäre sicherlich sehr wünschenswert. Leider fehlen allerdings bislang Verfahren, derartige Dimensionen zu erfassen. Eine theoriegeleitete Ableitung weitere Kriteriumfacetten und deren Umsetzung in geeignete Verfahren neben einer solchen von Prädiktoren in Testverfahren in dieser Arbeit hätte allerdings ihren Rahmen gesprengt. Daher beschränkt sich also auch diese Arbeit auf Studiennoten als das bislang relativ, jedoch nicht absolut gesehen, beste Studienerfolgskriterium.

4. Prädiktoren des Studienerfolgs

Nachdem im vorigen Kapitel wesentliche Bestimmungstücke des Studienerfolgs erläutert wurden, werden in diesem Kapitel werden die wichtigsten Prädiktoren des Studienerfolges dargestellt und vor dem Hintergrund der Hauptgütekriterien Objektivität, Reliabilität und prognostischer Validität beurteilt.
Die Literatur bietet mittlerweile ein Fülle infrage kommender und empirisch untersuchter Prädiktorklassen. Die Darstellung der vorliegenden Arbeit orientiert sich dabei an der folgenden Klassifikation von Rindermann und Oubaid (1999):

- Schulnoten
  - Abiturdurchschnittsnote
  - Einzelfachnoten
- Testleistungen
  - Intelligenztests
  - Kenntnistests
  - Studierfähigkeitstests
- Leistungen in Aufnahmeprüfungen und Arbeitsproben
- Interviews bzw. Auswahlgespräche
- Situative Auswahlverfahren und Assessment-Center
- Essays
- Nicht-intellektuelle Prädiktoren: Persönlichkeitsmerkmale und allgemeine Kompetenzen

Die folgenden Darstellungen der Ergebnisse zu Studierendenauswahlverfahren beschränken sich allerdings aus zwei Gründen auf solche zu Schulnoten und Testleistungen sowie auf diejenigen zu Persönlichkeitsmerkmalen. Erstens kommt diesen Personenmerkmalen eine besondere Bedeutung zu, da sie diejenige Prädiktoren darstellen, welche dem empirischen Teil der vorliegenden Arbeit zugrunde liegen bzw. aus den Ergebnissen einer Anforderungsanalyse zu hypothetisch studienerfolgsrelevanten Personenmerkmalen abgeleitet worden waren (s. hierzu Kap. 7.1 bis 7.3). Zweitens weisen insbesondere Schulnoten und Studierfähigkeitstests in zahlreichen Studien und Metaanalysen von allen bislang untersuchten Prädiktoren die besten Validitäten in Bezug auf Studienerfolg auf (Hell, Trapmann, Weigand, Hirn & Schuler, 2005; Höppel & Moser, 1993; Köller & Baumert, 2002; Rindermann & Oubaid, 1999; Robbins et al., 2004; Schuler, Funke & Baron-Boldt, 1990; Trapmann, Hell & Schuler, 2005a, 2005b; Trost, 1975, 2003; Trost & Freitag, 1991). Zu ausführlicheren Darstellungen der übrigen

Prädiktorklassen sei daher lediglich auf weiterführende Literatur verwiesen (s. hierzu insbesondere Deidesheimer Kreis, 1997; Rindermann & Oubaid, 1999; Trost, 1975, 2003, 2005; Trost & Bickel, 1979).

4.1 Schulnoten

Abiturnoten stellen das bislang am häufigsten eingesetzte leistungsorientierte Auswahlinstrument für die Vergabe von Studienplätzen an deutschen Universitäten dar (Rindermann & Oubaid, 1999, S. 176). Gründe hierfür liegen in deren leichter Verfügbarkeit und in ihrem Status als Indikatoren des erreichten Leistungsstandes im jeweiligen Schulfach. Als Prädiktoren kommen hierbei sowohl Abitureinzelnoten als auch die Abiturdurchschnittsnote in Betracht. Die Gesamtabschlussnote gilt zudem auch als Maß einer globalen intellektuellen Leistungsfähigkeit (Köller & Baumert, 2002, S. 15) und zeigt nicht zuletzt deshalb auch die höchste konzeptionelle Nähe zum Kriterium der Hochschulreife. Neben diesen intellektuellen Anteilen enthält die Abiturdurchschnittsnote jedoch ebenso Persönlichkeitsanteile, welche durchaus prädiktiven Wert für Studienerfolg aufweisen wie z.B. Fleiss, Lerninteressen, Arbeitshaltungen und Merkmale des Selbstkonzeptes (Baron-Boldt, Schuler & Funke, 1988, S. 21; Gold & Souvignier, 2005, S. 217; Trost, 1975, S. 11). Die Abiturdurchschnittsnote stellt somit eine „globale Schätzung über viele Fächer, abgegeben von verschiedenen Lehrern und aufgrund einer jahrelangen Beobachtung im Unterricht dar“ (Konradt, 1997, S. 156) und sollte daher den Studienerfolg besonders gut vorhersagen.

4.1.1 Objektivität und Reliabilität

Ungeachtet ihrer hohen diagnostischen und pädagogischen Funktion als Belohnung unterrichtsfördernden Verhaltens wird an Schulnoten ihre geringe Objektivität, unbefriedigender Beurteilerübereinstimmung und niedrige Retestreliabilität kritisiert (s. insbesondere Ingenkamp, 1971). Die Durchführungsobjektivität von Schulnoten leidet bereits unter den unterschiedlichen Modalitäten mündlicher und schriftlicher Leistungserhebungen. Die Forschung zur Auswertungsobjektivität von Schulnoten konnte ein ganzes Bündel an beeinflussenden Faktoren identifizieren. Hierbei zeigen sich deutliche Effekte durch die Persönlichkeit des Lehrers, Klausurreihenfolge und dem Geschlecht des Schülers (Baron-Boldt, Schuler & Funke, 1988; Möller & Köller, 1997; Schrader & Helmke, 1990).

Trost (1975, S. 12f.) nennt auf geografischer und institutionaler Betrachtungsebene weitere Faktoren wie eine größere Strenge der Notengebung in den südlichen Bundesländern und in bestimmten Lehrfächern, Abhängigkeit der Notenvergabe vom Schultypus, von der Gemeindegröße. Hinsichtlich der Interpretationsobjektivität wird bemängelt, dass der Bezugsrahmen für Noten lediglich schulintern und nicht kriteriumsorientiert definiert sei und die Lehrerinnen und Lehrer den gesamten Bereich der Notenskala nicht ausnutzten. Überwiegend unabhängig vom Klassenleistungsniveau erhalten somit die *relativ* Besten die Note „sehr gut“, die Schwächsten hingegen „mangelhaft“ (Köller, Baumert & Kai, 1999).

Da die Objektivität funktional mit der Reliabilität verknüpft ist, wirken sich die oben genannten Probleme auch mindernd auf letztere aus. Baron-Boldt (1989) berichtet demgemäß auch Interraterreliabilitäten von $r = .60$. Noch niedriger und breiter streuend sind die Retestreliabilitäten. Hier ergeben sich Werte von $r = .30$ bis $r = .86$ (Orlik, 1961; Tent, Fingerhut & Langfeld, 1976). Als zusätzliche Fehlerquelle tritt hier neben derjenigen mangelnder Beurteilerübereinstimmung noch die zeitliche Merkmalsfluktuation hinzu. Die Reliabilität der Abiturdurchschnittsnote mit ihrem größeren Aggregationsniveau liegt allerdings gegenüber Einzelfachnoten insgesamt höher, da sich spezifische Fehlereinflüsse einzelner Lehrerurteile und Prüfungsergebnisse ausmitteln (Baron-Boldt, Schuler & Funke, 1988, S. 80, 86).

4.1.2 Prädiktive Validität

Vor dem Hintergrund der berichteten erheblichen psychometrischen Mängel der Abiturnoten ist der vielfach replizierte Befund „umso verblüffender ..., dass die durchschnittliche Abiturnote eines der besten Einzelmaße zur Prognose des Studienerfolgs ist“ (Köller & Baumert, 2002, S. 15). Tabelle 1 gibt einen Überblick über prädiktive Validitäten von Schulnoten aus Literaturübersichten und Metaanalysen.

Tabelle 1: Prädiktive Validität von Schulleistungen (Fortsetzung der Tabelle auf folgender Seite)

Autorenname & Erscheinungsjahr, Art der Studie	**Prädiktor**	**Kriterium**	**Validität**	**Korrekturen**	**Stichprobe/N**
Abschlussnoten als Prädiktor					
Weingardt (1972) Überblicksartikel	Abiturnote	Abschluss- bzw. Zwischennoten	r = .28 bis r = .49	-	7 Studien N = 80 - 551
Trost & Bickel (1979) Überblicksartikel	Abiturnote	Abschluss- bzw. Zwischennoten	r = .02 bis r = .53 Median r = .35	-	54 Korrelationen N = 19 - 3 534
Baron-Boldt, et al. (1988) Metaanalyse	Abiturnote	Abschluss- bzw. Zwischennoten	mittleres r = .46	Minderung in Y	75 Stichproben N = 26 867
Baron-Boldt, et al. (1988) Metaanalyse	Abiturnote	Studienabbruch	mittleres r = .30	Minderung in Y	75 Stichproben N = 26 867
Höppel & Moser (1993)	Abiturnote	Abschluss- bzw. Zwischennoten	r = .54 bzw. r = .56	Minderung in Y Selektionsk.	Agrarwissenschaften N = 920
Giesen und Gold (1996)	Abiturnote	Studiendauer	r = .02 bis r = .15	-	unterschiedliche Studienrichtungen N = 152 - 372
Rindermann & Oubaid (1999)	Abiturnote	Abschluss- bzw. Zwischennoten	r =.28 bis r = .48	-	Überblicksartikel 8 Studien N = 85 - 27 000
Gold und Souvignier (2005)	Abiturnote	Abschlussnote	r = .08 bis r = .31	-	Medizin, Jura, Ingenieur- und Wirtschafts-wissenschaften N = 395
Hell et al. (2005) Metaanalyse	Schulabschluss noten	Abschluss- bzw. Zwischennoten	mittleres r = .46	Minderung in Y	unterschiedliche Studienrichtungen 54 Studien N = 48 178

Einzelnoten als Prädiktor	Prädiktor	Kriterium	Validität	Korrekturen	Stichprobe/N
Weingardt (1972)	Einzel-fach-noten	Abschluss- bzw. Zwischennoten	r = .06 bis r = .31	-	Überblicksartikel 5 Studien N = 35 - 551
Baron-Boldt, Schuler & Funke (1988)		Abschluss- bzw. Zwischennoten	r = .07 bis r = .34	Minderung in Y	Metaanalyse 5 - 18 Stichproben N = 1 542 - 4 244
Rindermann (1996)		Zwischennoten	r = .22 bis r = .31	-	Romanistik N = 85

Anmerkung. GPA:Grade Point Average; Minderung in Y: Minderungskorrektur des Kriteriums; Selektionsk.: Selektionskorrektur

Vorab ist für die Interpretation zu betonen, dass Werte über r = .60 von keinem Einzelprädiktor zu erwarten sind. Der Hochschulerfolg wird neben individuellen Eingangsvoraussetzungen auch von vielen anderen Faktoren beeinflusst, welche erst im Laufe des Studiums wirksam werden und welche die Genauigkeit jedweder Prognose senken. Vor diesem Hintergrund ergibt sich aus den Ergebnissen insgesamt eine vergleichsweise gute Prognosekraft von Schulnoten. Die Abiturdurchschnittsnote als Aggregat verschiedener leistungsrelevanter Personenmerkmale erweist sich dabei Einzelfachnoten stets als überlegen (s. auch Baron-Boldt et al., 1988; Rindermann & Oubaid, 1999, S. 178). Im Mittel liegt ihre prädiktive Validität zwischen r = .30 und r = .40, diejenige von Einzelnoten etwas darunter.

Die Ergebnisse zur durchschnittlichen Abiturleistung korrespondieren auch gut mit solchen aus anderen Ländern wie den USA. Auch hier zeigte sich wiederholt, dass die mittlere Abschlusspunktzahl (Highschool Grade Point Average, HGPA) die Studienleistung sogar noch besser vorhersagen kann als in Deutschland. In den USA liegen die Koeffizienten im Bereich von r = .41 bis r = .53 (Köller & Baumert, 2002, S. 15). Trost (1975, S. 23) verweist in diesem Zusammenhang auf die ähnliche Strukturierung des Studiums an amerikanischen Highschools und Colleges. Die größere „Verschulung“ (im Sinne einer stärkeren Strukturierung des Studiums) fungiert hier als Moderatorvariable und führt zu höheren Zusammenhängen. In diesem Kontext ist auch die in Tabelle 1 und in der Metaanalyse von Baron-Boldt et al. (1988) auffällige starke Streuung der Validitätskoeffizienten zu interpretieren. Für Studiengänge mit einem höheren Strukturiertheitsgrad resultieren im Allgemeinen höhere Koeffizienten (Rindermann & Oubaid, 1999, S. 178). Die an Studienleistungen über verschiedenen Nationen ausgerichtete Metaanalyse von Hell, Trapmann, Weigand, Hirn und Schuler (2005) mit Studien

aus dem Zeitraum 1980 bis 2005 weist die durchschnittliche Schulabschlussnote mit einer mittleren Validität von $r = .46$ auch international betrachtet als guten Prädiktor aus.

Nach Studienfachrelevanz (etwa durch Expertenurteil) gewichtete Einzelfachnoten erbringen keine bessere Vorhersage (s. auch Tabelle 1, Abschnitt „Einzelnoten als Prädiktor“). Beispielsweise korreliert die Studienleistung in Anglistik mit der Abiturdurchschnittsnote höher als mit der Abschlussnote im Schulfach Englisch (Baron-Boldt, 1989). Die Annahme, dass in Einzelnoten die spezifischen Anforderungen und eines entsprechenden Studienfaches besser als die Abiturdurchschnittsnote abbilden, lässt sich nicht unterstützen. Auch in der Metaanalyse von Baron-Boldt et al. (1988, S. 83f.) erzielte die Durchschnittsnote stets höhere mittlere Prognosewerte als Einzelnoten. Als die relativ besten Prädiktoren aufseiten der Einzelfachnoten erwiesen sich die Mathematik- und Physik-Note. Steyer, Yousfi und Würfel (2005) berichten auch für das Studienfach Psychologie geringere Koeffizienten von Einzelfachnoten gegenüber der Abiturdurchschnittsnote. Allerdings können apriorigewichtete Einzelfachnoten neben der Abiturdurchschnittsnote bei einigen Studienfächern einen ergänzenden Prognosebeitrag leisten. Beispielsweise zeigt sich bei einigen naturwissenschaftlichen Studiengängen, dass die Prognosekraft der Abiturdurchschnittsnote durch eine separate Gewichtung von naturwissenschaftlichen Schulfächern noch gesteigert werden kann. (Hell et al., 2005, S. 2).
In Bezug auf die Vorhersage des *Studienabbruchs* erweist sich der Abiturdurchschnitt als der wichtigste Prädiktor vor selbst eingeschätztem Fleiß, Leistungsfähigkeit und sozialer Anerkennung (Gold & Kloft, 1991, zit. nach Rindermann & Oubaid, 1999, S. 178). In der Metaanalyse von Baron-Boldt et al. (1988) erzielt sie eine befriedigende Prognosekraft (s. Tabelle 1). Unwesentlich fallen hingegen die Kriteriumskorrelationen mit Studiendauer aus (s. Tabelle 1). Wie bereits in Kapitel 3.2.4 ausgeführt, hängt diese Kriteriumsvariable jedoch weitaus stärker von institutionellen Rahmenbedingungen als von kognitiven Merkmalen ab.

4.1.3 Zusammenfassung

Die Abiturdurchschnittsnote ist in Bezug auf die Vorhersage des Studienerfolgs in Gestalt von Studiennoten trotz zweifelhafter psychometrischer Qualität ein wichtiger Prädiktor und sollte gemäß den Empfehlungen des Wissenschaftsrats (2004) und anderer Autoren eine zentrale Rolle bei der Entscheidung über die Zulassung zu einem Studium spielen (Köller & Baumert, 2002; Rindermann & Oubaid, 1999). Gleichwohl muss am Abitur als *hinreichendem* Indikator

der Studierfähigkeit aus diversen Gründen gezweifelt werden (s. in diesem Sinne Deidesheimer Kreis, 1997; Konradt, 1997; Rindermann & Oubaid, 1999; Trost, 1975; Trost & Bickel, 1979; Wissenschaftsrat, 2004). Insbesondere die Vergleichbarkeit der Abiturnoten zwischen den einzelnen Bundesländern ist aufgrund des föderalen Bildungssystems in Deutschland nur eingeschränkt gegeben. Die gymnasialen Oberstufen unterscheiden sich hier hinsichtlich Dauer, (Ab-)Wählbarkeit zentraler Fächer sowie der Kombinationsmöglichkeiten von Leistungs- und Grundkursfächern deutlich. Einige Bundesländer führen zentrale Abiturprüfungen durch, andere hingegen dezentrale, was zu deutlichen Schwierigkeitsunterschieden führt (Baumert & Watermann, 2000) und nicht zuletzt auch zu Unfairness in der Zulassungsentscheidung, wird die Abiturdurchschnittsnote als alleiniges Kriterium herangezogen. Demgemäß argumentieren Rindermann und Oubaid (1999), dass die Abiturdurchschnittsnote ein zufriedenstellender Prädiktor der allgemeinen Studierfähigkeit ist, jedoch keiner für *spezifische*. Daher ist zu fordern, dass ergänzende Studierfähigkeitstests eine mindestens befriedigende Korrelation mit Maßen des Studienerfolgs, insbesondere Studiennoten, zeigen. „Diese sollte höher liegen als die Abiturnote. Genauso muss der Einsatz als Abitur-Plus--Konzeption eine inkrementelle Validität bewirken“ (Zimmerhofer, 2003, S. 17).

4.2 Leistungstests

Der Einsatz von Leistungstests neben Schulnoten wird bei der Vorhersage oder Feststellung der Studieneignung von verschiedenen Autoren seit längerer Zeit gefordert. Meist ist diese Forderung mit der Erwartung verknüpft, zum einen über Leistungstests einen Zugewinn an prognostischer Validität zu erzielen, zum anderen sowohl die Objektivität und Reliabilität als auch die Fairness von Selektionsentscheidungen zu optimieren (Konradt, 1997; Rindermann & Oubaid, 1999; Trost & Bickel, 1979; Wissenschaftsrat, 2004). Diese Erwartung lässt sich anhand der Forschungsliteratur untersuchen, da Studieneingangstests mittlerweile neben Schulnoten die bestbeforschten Prädiktoren bei der Hochschulzulassung sind. Es folgt daher nun ein Überblick über die jeweiligen Verfahren nebst einer Darstellung empirischer Befunde zu ihren prädiktiven Validitäten.

Die Untergliederung von Testverfahren ist in der Literatur uneinheitlich (vgl. z.B. Deidesheimer Kreis, 1997; Rindermann & Oubaid, 1999; Trost, 1975, 2003; Trost & Bickel, 1979). Anhand der Literatur bietet sich jedoch das folgende Grundgerüst für eine Einordnung der Verfahren an:

- Intelligenztests
- Kenntnistests
  - Schulfachbezogene Kenntnistests
  - Studienfachspezifische Kenntnistests
- Studierfähigkeitstests
  - Allgemeine Studierfähigkeitstests
  - Spezifische Studierfähigkeitstests

Der Schwerpunkt der folgenden Darstellung soll auf Studierfähigkeitstest liegen, da sie gerade für den empirischen Teil dieser Arbeit besondere Relevanz aufweisen.

4.2.1 Intelligenztests

Untersuchungen zum Zusammenhang von Intelligenztestleistungen und Maßen des schulischen Erfolges liefern konsistent die höchsten Übereinstimmungen in der psychologischen Diagnostik. Amelang und Bartussek (1997, S. 246) geben als groben Schätzwert für die mittlere Korrelation $r = .50$ an. Auch die in den Testmanualen berichteten Koeffizienten liegen meist in diesem Bereich (vgl. z.B. Amthauer, Brocke, Liepmann & Beauducel, 2001; Jäger, Süss & Beauducel, 1997).
Die Hypothese, ähnliche Zusammenhänge mit Maßen des Studienerfolgs zu finden, kann jedoch anhand der Forschungsliteratur nicht gestützt werden. Zwar ergeben sich fast durchgängig Korrelationen in der erwarteten Richtung, doch liegen diese meist vergleichsweise niedrig. Die bei Amelang (1975, S. 121) berichteten Zusammenhänge zwischen Generalfaktorwerten und der (umgepolten) Note im Vorexamen verschiedener Fächer ($N = 547$) liegen im Durchschnitt bei $r = .33$. In einer Untersuchung von Gasch (1971) liegt die an 254 Psychologiestudierenden von fünf westdeutschen Hochschulen ermittelte Korrelation zwischen dem IST-70-Gesamtrohwert und der (umgepolten) Vordiplomnote bei $r = .28$, diejenige mit der Hauptdiplomnote bei $r = .19$. Trost und Bickel (1979) geben in einer Literaturübersicht einen Median von $r = .22$ an. Trost (1975, S. 32f.) findet durchweg niedrige Werte, die durch statistische Gewichtung eine multiple Korrelation von maximal $r = .40$ erreichen.
Einer der Gründe für die insgesamt niedrigen Werte liegt in der Varianzeinschränkung der Testrohwerte durch die stark vorselegierte Stichprobe von Studierenden an Hochschulen. Weiterhin nennen Amelang und Bartussek (1997, S. 246) die besonders niedrige Reliabilität

der Hochschulnoten. Allerdings führt Trost (1975, S. 35; 2003, S. 14) an, dass die in einem Intelligenztest erfassten Faktoren nicht identisch mit denen sind, die zum erfolgreichen Durchlaufen des Studiums notwendig sind und welche demgegenüber in Studierfähigkeitstests erfasst werden. „Diese spezifische ‚akademische' Intelligenz scheint sich in Schulnoten und in spezielleren Tests, deren Anforderungsniveau höher ist und deren Erfassungsbereich - neben manchen Überlappungen - von dem der allgemeinen Intelligenztest abweicht, stärker zu manifestieren" (Trost, 1975, S. 35). Bedenkt man zudem, dass in Intelligenztests meist das logischer Operieren mit relativ kleinen Informationseinheiten erfasst wird, so scheinen die durch sie erfassten Fähigkeiten demnach lediglich eine begrenzte Teilmenge von Fähigkeiten und Fertigkeiten zu sein, welche zum Bewältigen der komplexeren Anforderungen eines Studiums erforderlich sind. In diesem Sinne konstatieren daher Trost und Bickel (1979, S. 26), dass „[n]ach den vorliegenden Befunden ... Leistungen in Intelligenztests also in allenfalls niedrigem Zusammenhang mit Leistungen in Studium [stehen]".

4.2.2 Studienfachspezifische Kenntnistests

Zweck eines Kenntnistests überhaupt ist die Überprüfung des Wissensstandes eines Studienbewerbers „im allgemeinen oder in bestimmten Bereichen, die im Zusammenhang mit seiner Studienwahl stehen" (Deidesheimer Kreis, 1997, S. 84-85). Hier unterscheidet man *schulfachbezogene* und *studienfachbezogene* Kenntnistests. Da die Hochschulrektorenkonferenz allerdings darauf verweist, dass in Auswahlverfahren keine schulfachbezogene Inhalte geprüft werden dürfen (s. u.a. Reich, 2002, S. 305), wird an dieser Stelle nur auf studienfachbezogene Kenntnistests eingegangen.

Kenntnistests zielen in ihren Inhalten auf Fähigkeiten ab, die für das erfolgreiche Absolvieren bestimmter Studiengänge besonders erforderlich sind (s. u.a. Deidesheimer Kreis, 1997, S. 86). Typische Beispiele aus dem englischsprachigen Raum sind der „Medical College Admission Tests" (MCAT), der „Graduate Management Admission Test" (GMAT) und der „Law School Admission Test" (LSAT). Das prominenteste Beispiel aus dem deutschsprachigen Raum ist der „Test of English as a Foreign Language" (TOEFL).
Die Vorhersagekraft von Kenntnistests liegt in den meisten Längsschnittanalysen im befriedigenden Bereich, ist allerdings meist gleich mit demjenigen von Schulleistungen in der Oberstufe. Auch gegenüber Studierfähigkeitstests fällt die Prognosegenauigkeit geringer aus (Deidesheimer Kreis, 1997, S. 87). Problematisch an Kenntnistests ist ihre leichte Trainier-

barkeit und die daraus resultierenden „Begleiterscheinungen“. So ist in Ländern mit starker Verbreitung von Kenntnistests (z.B. Japan, China und Türkei) eine starke Konzentration auf Vermittlung der rein auf reproduktive Fähigkeiten abgestimmten Testinhalte in den letzten Schuljahren zu beobachten und zwar zulasten eines tiefer gehenden Verständnisses und breiten Wissens (Deidesheimer Kreis, 1997, S. 88). Meist geht dies noch einher mit der Etablierung einer „Bildungsindustrie“ in Form von kommerziellen Bildungseinrichtungen neben dem eigentlichen Bildungssystem (Deidesheimer Kreis, 1997, S. 88).

4.2.3 Studierfähigkeitstests

Studierfähigkeitstests sind von den im vorigen Abschnitt behandelten Kenntnistests insofern zu unterscheiden, als dass in ihnen ausschließlich kognitive Fähigkeiten und Fertigkeiten erfasst werden, die für das Bewältigen von Studienanforderungen bedeutsam sind. Erworbene Kenntnisse sind daher nicht Inhalt dieser Tests. Hier muss einschränkend gesagt sein, dass die bekanntesten Studierfähigkeitstests aus den USA, der „Scholastik Aptitude Test“ (SAT) und die „General Record Examination“ (GRE), neben Tests zur allgemeinen Studierfähigkeit auch „Subject Tests“ zu verschiedenen Studienfächern enthalten. Im Falle des SAT hat dies 1993 zu einer Trennung in zwei Testformen geführt. Der SAT I erfasst dabei quantitatives und verbales schlussfolgerndes Denken, der SAT II hingegen ist ein Wissens- und Fertigkeitstest zu verschiedenen curriculumbasierten Themen.

Innerhalb der Gruppe der Studierfähigkeitstest wird wie folgt zwischen allgemeinen und studienfachspezifischen Tests unterschieden (Deidesheimer Kreis, 1997, S. 85ff.):

- Allgemeine Studierfähigkeitstests:
  Sie bestehen meist aus einem verbalen und einem quantitativen Teil. Hierbei können Testteile wiederum in verschiedene Aufgabengruppen untergliedert sein. Die bekanntesten allgemeinen Studierfähigkeitstests aus den USA sind der SAT I: Reasoning Test (College Board, 2005), die GRE (Educational Testing Service, 2005), wobei der SAT I bei der Zulassung zum College-Studium appliziert wird, die GRE hingegen bei der Zulassung zum Master- oder PhD-Studium und schließlich noch der „Admission to College Test“ (ACT, ACT Incorporated, 2005). Beispiele aus dem deutschsprachigen Raum sind der „Auswahltests der Studienstiftung“ (ATS) und der „Test der akademischen Befähigung“ (TAB) . In Schweden findet der „Swedish

Scholastic Aptitude Test“ (SweSAT) (Deidesheimer Kreis, 1997, S. 90) und in Israel der „Psychometric Entrance Test“ (PET) Anwendung.

- Studienfachspezifische Fähigkeitstests:
  Sie erfassen die spezifischen kognitiven Anforderungen der jeweiligen Fächer. Die Aufgaben enthalten überwiegend fachrelevante Informationen, die auf neue Fragestellungen übertragen werden müssen. Durch diesen fachspezifischen Zuschnitt sollen sie genauere Prognosen als die allgemeinen Studierfähigkeitstests erlauben. In Deutschland zählte der von 1986 bis 1996 eingesetzte „Test für medizinische Studiengänge“ (TMS) zu diesen Verfahren, im englischsprachigen Raum der GMAT, LSAT und der MCAT. In den letzten Jahren sind insbesondere an deutschen Fachhochschulen und privaten Universitäten für verschiedene Studienfächer eine ganze Reihe studienfachspezifische Fähigkeitstests entwickelt und evaluiert worden, welche jedoch hier nicht alle dargestellt werden können (für eine Übersicht s. Trost, 2003, S. 17ff.).

Die konzeptuelle Nähe insbesondere der allgemeinen Studierfähigkeitstests mit herkömmlichen Intelligenztests ist offensichtlich und auch empirisch zeigt sich eine Überschneidung der Messbereiche. So berichtet Trost (2003, S. 13) von einer Korrelation zwischen dem TMS und einem allgemeinen Intelligenztest von $r = .70$ und auch Frey und Detterman (2004a; 2004b; 2005) finden einen Zusammenhang des SAT I mit einem g-Faktor-Test von $r = .82$. Die Schlussfolgerung aus korrelativ ermittelten konkurrenten Validitäten, dass die Tests jeweils dasselbe, nämlich Intelligenz messen, ist allerdings rein epistemologisch und hoch problematisch wie Borsboom, Mellenbergh und van Herden (2004) anmerken:

> For instance, suppose one is measuring the presence of thunder. The readings will probably show a perfect correlation with the presence of lightning. The reason is that both are the result of an electrical discharge in the clouds. However, the presence of thunder and the presence of lightning are not the same thing under a different label. They are strongly related - one can be used to find out about the other - and there is a good basis for prediction, but they are not the same thing. When one is validly measuring the presence of thunder, one is not validly measuring the presence of

> lightning for the simple reason that one is not measuring the presence of lightning at all. (S. 1066)

Man mag gegenüber diesem Argument anbringen, es reiche zur Messung von „Elektrizität“ als die beiden Teilphänomenen zugrunde liegende Variable aus, Blitz *oder* Donner zu messen. In Analogie entspräche sodann die Elektrizität der durch beide Tests erfassten Intelligenz. Auf Ebene von *manifesten* Variablen wie Blitz und Donner, bei denen zudem die Kausalstruktur mit Elektrizität *bekannt* ist, mag dies auch stimmen. Im Falle von latenten Variablen wie der Intelligenz ist dies jedoch problematisch, da dies in zirkulärer Weise *voraussetzen* würde, dass beide Tests überhaupt Intelligenz messen, wofür jedoch kein *hinreichender* Grund besteht. Trost (1975, S. 38f.) führt darüber hinaus noch als weitere Unterschiede an, dass Studierfähigkeitstests z.B. hinsichtlich Normen und Schwierigkeitsgrad speziell auf die Bewerber zugeschnitten sind und außerdem spezifische kognitive Anforderungen eines Hochschulstudiums differenzierter abzubilden vermögen. Zudem belegen empirische Untersuchungen die höhere Prognosekraft von Studierfähigkeitstests gegenüber allgemeinen Intelligenztests bezüglich Studienerfolg (Trost, 2003, S. 14).

4.2.3.1 Objektivität und Reliabilität

Die Durchführungsobjektivität ist durch korrekte Testvorgabe und genauer Befolgung der Instruktion bei Studierfähigkeitstest hinlänglich gegeben. Die Auswertungs- und Interpretationsobjektivität wird durch das Multiple-Choice-Format und das maschinelle Einlesen und Auswerten der Antworten in hohem Maße gesichert.
Die Reliabilität bewegt sich im hohen bis sehr hohen Wertebereich. Donlon (1984, zit. nach Trost, 2003, S. 37) berichtet für den SAT I Retestreliabilitäten von $r_{tt} = .89$. Diejenige des TMS lag nach 13 Monaten bei $r_{tt} = .80$ (Fay, zit. nach Trost, 2003, S. 37). Die interne Konsistenz nach der Split-half-Reliabilität liegt bei fast allen Studierfähigkeitstests über .90 (Trost, 2003, S. 37).
Insgesamt kann die psychometrische Qualität von Studierfähigkeitstests hinsichtlich Objektivität und Reliabilität als gut bezeichnet werden.

4.2.3.2 Prädiktive Validität allgemeiner Studierfähigkeitstests

Im folgenden Abschnitt werden die Forschungsergebnisse zur prognostischen Validität von allgemeinen Studierfähigkeitstests referiert und bewertet. Das Hauptgewicht liegt hierbei auf US-amerikanischen Studien, da diese die extensivste Forschungsergebnisse liefern. Tabelle 2 gibt einen Überblick über die Ergebnisse.

Tabelle 2: Übersicht über Ergebnisse zu prädiktiven Validitäten von allgemeinen Studierfähigkeitstests (Fortsetzung der Tabelle auf folgenden Seiten)

Autorenname & Erscheinungsjahr	**Prädiktor**	**Kriterium**	**Korrelation**	**Korrekturen**	**N**
Studien aus dem deutschsprachigen Raum					
Trost (1986, zit. nach Deidesheimer Kreis, 1997)	ATS	Vorexamina	r = .16	-	keine Angabe
Hitpass, Ohlsson und Thomas (1984)	TAB	Zwischenprüfung	r = .51	-	85
Studien aus dem nicht-deutschsprachigen Raum					
Humphreys (1968)	ACT	Durchschnittspunktzahl nach dem 1. Collegejahr	r = .38	-	1 600
		Durchschnittliche Abschlusspunktzahl	r = .17		
Bridgeman, McCamley-Jenkins und Ervin (2000)	SAT I (Reasoning)	Durchschnittspunktzahl nach dem 1. Collegejahr	r = .35	-	48 039
Baron und Norman (1992)	SAT I (Reasoning)	Durchschnittspunktzahl nach dem 1. Collegejahr	r = .20	-	3 816
Crouse und Trusheim (1988, S. 46)	SAT	Durchschnittspunktzahl nach dem 1. Collegejahr	r = .37	-	2 470
Bridgeman, McCamley-Jenkins und Ervin (2000)	SAT	Durchschnittspunktzahl nach dem 1. Collegejahr	r = .34	-	45 100
Burton und Ramist (2001) Überblicksartikel	SAT-V, SAT-M,	Durchschnittspunktzahl nach dem 1. Collegejahr	r = .17 bis r = .56	-	28 bis 4 094
Robbins, Lauver, Davis, Langley und Carlstrom (2004) Metaanalyse	ACT/SAT	Durchschnittspunktzahl nach dem 1. Collegejahr	mittleres r = .36	-	16 648

Autorenname & Erscheinungsjahr	**Prädiktor**	**Kriterium**	**Korrelation**	**Korrekturen**	**N**
Geiser und Studley (2002)	SAT I SAT II	Durchschnittspunktzahl nach dem 1. Collegejahr	mittleres *r* = .35 mittleres *r* = .39	-	77 893
Morrison und Morrison (1995) Metaanalyse	GRE-V, GRE-Q	Durchschnittliche Abschlusspunktzahl	*r* = .22 *r* = .28	-	5 186
House (1998)	GRE	Durchschnittliche Abschlusspunktzahl	*r* = .29	-	5 047
House (1998)	GRE	Punkte in diversen Psychologie-Kursen	*r* = .24 bis *r* = .36	-	275
Kuncel, Hezlett und Ones (2001) Metaanalyse	GRE-V GRE-Q GRE-A	Durchschnittl. Abschlusspunktzahl	*r* = .34 *r* = .32 *r* = .36	Minderung in Y Selektionsk.	1 928 bis 14 425
Kuncel, Hezlett und Ones (2001) Metaanalyse	GRE - Subject Part	Durchschnittl. Abschluspunktzahl	*r* = .41	Minderung in Y Selektionsk.	2 413
Goldberg und Alliger (1992) Metaanalyse	GRE	Punkte in diversen Psychologie-Kursen	*r* = .03 bis *r* = .37 mittleres *r* = .18	-	1 800
Sternberg und Williams (1997)	GRE-V, GRE-Q GRE-A	Durchschnittliche Abschlusspunktzahl in Psychologie	*r* = .17 *r* = .09 *r* = .12	-	155-159
Schneider und Briel (1990)	GRE-V, GRE-Q GRE-A	Durchschnittspunktzahl nach dem 1. Graduiertenjahr	*r* = .29 *r* = .28 *r* = .26	-	9 200
Schneider und Briel (1990)	GRE-V, GRE-Q GRE-A	Fakultätsratings	*r* = .25 *r* = .25 *r* = .21	-	891
Burton und Turner (1983)	GRE-V, GRE-Q GRE-A	Fakultätsratings	*r* = .22 *r* = .24 nicht angegeben	-	
Beller (1993)	PET (Israel)	Durchschnittspunktzahl nach dem 1. Collegejahr	*r* = .38	Selektionsk.	52 985

Anmerkung. ATS: Allgemeiner Test der Studierbefähigung; ACT: American College Test; SAT: Scholastic Apptitude Test; SAT-V: Scholastic Aptitude Test Verbalteil; SAT-M: Scholastic Aptitude Test Mathematischer Teil; GRE: General Record Examination; GRE-V: General Record Examination Verbalteil; GRE-Q: General Record Examination quantitativer Teil; GRE-A: General Record Examination Analytical; PET: Psychometric Entrance Test; GMAT: Graduate Management Admission Test; Minderung in Y: Minderungskorrektur im Kriterium (Y); Minderung in Y, Minderung in X: Minderungskorrektur im Kriterium (Y) und Prädiktor (X).

Die Validitäten der deutschen allgemeinen Studierfähigkeitstests liegen ungefähr im Bereich der international ermittelten. Insgesamt fällt die große Varianz der Validitätskoeffizienten auf. Für den SAT bzw. SAT I liegen die Werte für kurzfristige Prognosen (Durchschnittspunktzahl nach dem ersten Collegejahr) typischerweise um $r = .30$ (s. Trost, 2003, S. 38). Morgan (1990, S. 104) berichtet neben den hier geschilderten Ergebnissen für die Zeitspanne von 1964 bis 1985 Korrelationen des SAT mit der durchschnittlichen Punktzahl im ersten Collegejahr von $r = .37$ bis $r = .45$. Die Korrelationen des High School Grade Point Average (als US-amerikanisches Pendant zur Abiturnote) variierten in diesem Zeitraum von $r = .45$ bis $r = .54$, die Differenz zugunsten des High School Grade Point Average liegt also bei ca. .08 (s. hierzu auch Schönemann, 2005, S. 196; Trost, 2003, S. 39).
Da der SAT als College-Zulassungstest lediglich am Kriterium der Durchschnittspunktzahl nach dem ersten Collegejahr evaluiert wird, liegen für seine langfristigen Prognosen, etwa in Bezug auf den Examensabschluss, keine Studien vor, obgleich dieses Kriterium für *jeglichen* Prädiktor die größere praktische Relevanz aufweist. Die klassische Studie von Humphreys (1968) berichtet eine Validität des ACT mit der durchschnittlichen Abschlusspunktzahl von $r = .17$ und eine des HGPA von $r = .22$. Für beide Prädiktoren erwies sich somit eine Prognose für längerfristige Prognosezeiträume als schwer realisierbar.

Für die GRE als Zulassungstests zu weiterführenden Master- und PhD-Studiengängen liegen die Korrelationen mit den durchschnittlichen Punktwerten aus dem ersten Jahr des Graduiertenstudium insgesamt betrachtet um $r = .30$, wobei auch hier die Werte breit streuen. In Bezug auf das weiter gefasste Erfolgskriterium der Fakultätseinschätzungen liegen die Validitäten um $r = .25$ und damit etwas niedriger. Nach der Metaanalyse von Kuncel, Hezlett und Ones (2001) fällt die Validität für den fachspezifischen Subject-Teil der GRE mit $r = .41$ insgesamt höher aus. In einer nicht in die Tabelle aufgenommenen kürzlich erschienenen Studie von Burton und Wang (2005) zu *langfristigen* Prognosen in Bezug auf den Cumulative Graduate Grade Point Average (CGGPA) und Fakultätseinschätzungen finden die Autoren deutlich höhere Validitäten im Bereich von $r = .38$ bis $r = .62$ für den Gesamtwert der GRE. Dieser Befund muss allerdings äußerst kritisch gesehen werden, da die hohen Korrelationen auf einem Methodenartefakt der von den Autoren durchgeführten Aggregierungsstatistik beruhen (Schönemann, 2006, zur Publikation eingereicht). Die Populationsschätzungen der Validitäten unterliegen hierbei einer systematischen Verzerrung, welche gerade bei niedriger Test-Kriteriums-Korrelation zu einer besonders deutlichen *Über*schätzung führt.

Insgesamt betrachtet liegen die Validitätskoeffizienten für allgemeine Studierfähigkeitstests im moderaten Bereich und unter denen von Schulabschlussnoten. Trost (2003, S. 39) verweist jedoch darauf, dass allgemeine Studierfähigkeitstests in Einzellfällen diese auch bezüglich der Validität übertreffen können. Vor diesem Hintergrund liegt für Anwendungszwecke eine Kombination von Schulnoten mit Studierfähigkeitstests nahe, insofern beide Prädiktoren nicht hoch miteinander korrelieren, und jeder verschiedene wichtige Aspekte des Studienerfolgskriteriums erfasst. Dieser Ansatz, sowie empirische Befunde hierzu, werden eigens im Kapitel 4.4 genauer ausgeführt.

4.2.3.3 Prädiktive Validität spezifischer Studierfähigkeitstests

Der folgenden Abschnitt referiert und evaluiert Forschungsergebnisse zur prognostischen Validität von studienfachspezifischen Studierfähigkeitstests. Für diesen Bereich liegt auch für den deutschsprachigen Raum (anders als für allgemeine Studierfähigkeitstests) eine reichhaltige Forschungsliteratur vor, weshalb der Fokus hierauf gerichtet wird. Tabelle 3 gibt zunächst einen Überblick über die wesentlichen Forschungsergebnisse.

Tabelle 3: Übersicht über Ergebnisse zu prädiktiven Validitäten von spezifischen Studierfähigkeitstests

Autorenname & Erscheinungsjahr	**Prädiktor**	**Kriterium**	**Korrelation**	**Korrekturen**	**N**
Studien aus dem deutschsprachigen Raum					
Trost und Piel (1991)	Auswahltest der WHU	Zwischennote	r = .15 bis r = .71	Selektionsk.	37 bis 47
Trost und Freitag (1991)	SFTs	Zwischennote	r = .11 bis r = .46	-	51 bis 512
Hänsgen und Spicher (2001, zit nach Trost, 2003, S. 38)	TMS (Schweiz)	Vorprüfung Humanmedizin	r = .50 bis r = .66	Selektionsk.	-
Klieme und Nauels (1996)	TMS	Vorprüfung Zahnmedizin	r = .24 bis r = .39	Selektionsk.	1277 bis 2807
Nauels und Stumpf (1997)	TMS	Vorprüfung Humanmedizin	r = .34	-	2037
			r = .34	-	2037
		subjektiver Studienerfolg	r = .14 bis r = .15	-	2037

Autorenname & Erscheinungsjahr		Prädiktor	Kriterium	Korrelation	Korrekturen
Stumpf und Nauels (1988)	TMS	Vorprüfung Humanmedizin	r = .29 bis r = .51	Selektionsk.	1 119 bis 2 329
Trost, Blum, Fay, Klieme, Maichle, Meyer und Nauels (1998)		Vor- und Hauptprüfungen in Human-, Tier- und Zahnmedizin	r = .24 bis r = .51	Selektionsk.	564 bis 25 876
Studien aus dem nicht-deutschsprachigen Raum					
Paolillo (1982)	GMAT		r = .26	-	220

Anmerkung. Studien*noten* wurden umgepolt, sodass höhere Werte für bessere Leistungen stehen.
WHU: Wissenschaftliche Hochschule für Unternehmensführung Koblenz; SFTs: Studienfeldbezogene Tests zur Beratung von Studierwilligen;
TMS: Test für medizinische Studiengänge; GMAT: Graduate Management Admission Test.
Selektionsk.: Selektionskorrigierte Korrelation

Die Koeffizienten liegen tendenziell höher als diejenigen der allgemeinen Studierfähigeitstests (vgl. Tabelle 2). In diesem Sinne ist den Forderungen nach speziell auf die Studienfächer bezogenen Studierfähigkeitstests zuzustimmen (s. in diesem Sinne Rindermann & Oubaid, 1999; Trost, 2003). Gleichwohl muss man auch für diese Tests eine große Streuung der Validitätskoeffizienten feststellen. Für den subjektiv eingeschätzten Studienerfolg ergeben sich keine praktisch bedeutsamen Zusammenhänge. Höhere Korrelationen zeigen sich nach einer Selektionskorrektur, welche die empirische Korrelation bei eingeschränkter Varianz in der selektionierten Bewerberstichprobe auf die gesamte Bewerberpopulation aufwertet bzw. schätzt.

4.2.3.4 Exkurs: Probleme bei der Anwendung von Selektionskorrekturen

An dieser Stelle sind einige kritische Anmerkungen zur Methode der Selektionskorrektur angebracht, weil sie mittlerweile ein durchaus gängiges, aber leider selten genug kritisch hinterfragtes Verfahren bei der Evaluation von Studierfähigkeitstests darstellt. Üblicherweise wird diese mit dem Argument durchgeführt, dass ein Test nicht an allen Bewerbern validiert wird, sondern lediglich an den durch den Test selektionierten. Die dadurch entstehende Varianzeinschränkung führt in aller Regel zu einer Minderung der Kriteriumskorrelation und lässt den Test schlechter wirken als bei einer Validierung an der gesamten Bewerberstichprobe. Die zu diesem Zweck eingesetzten Selektionskorrekturformeln korrigieren daher auch die

Kriteriumskorrelationen um +.10 bis +.15 (s. z.B. Willingham, Lewis, Morgan & Ramist, 1990, S. 125). Zwei Aspekte sind hieran problematisch. Erstens werden die statistischen Voraussetzungen der Korrekturformeln meist nicht überprüft, obgleich sie sehr sensitiv auf Voraussetzungsverletzungen reagieren (J. P. Campbell, 1976; Heckman, 1976, 1979; Humphreys, 1968; Linn, 1968). Die hierbei problematischste Annahme ist, dass die Varianzeinschränkung lediglich ein Effekt des Testtrennwertes ist. In der Praxis jedoch werden die Studienplatzbewerber neben dem Testwert auch anhand von Schulleistungen, Wartezeiten, Selbstelektion etc. vorselegiert. In der Folge sind die Korrelationschätzungen durch die Selektionskorrektur mehr oder weniger stark verzerrt, je nachdem, wie viele und wie stark weitere Faktoren neben dem Testergebnis auf den Selektionsprozess Einfluss nahmen (s. auch Crouse & Trusheim, 1988, S. 45f.). So konnte Rothstein (2002) zeigen, dass eine Verletzung der Voraussetzung der Varianzeinschränkung einzig durch den Testtrennwert u.a. zu einer *Über*schätzung der inkrementellen Validität des SAT gegenüber dem High School Grade Point Average von gut 7% führt. Zweitens kann aber auch gefragt werden, ob eine Selektionskorrektur aus *pragmatisch-inhaltlicher* Sicht sinnvoll ist, da der Test nicht für die gesamte Bewerberpopulation konzipiert wurde und für diese prädiktiv sein soll, sondern für die ausgewählte Stichprobe. Die Spezifikation der Zielpopulation ist bei diesen Korrekturen meist falsch vorgenommen, da die Testergebnisse lediglich in der Population der zugelassenen Bewerber verwertet werden. Anders liegt die Sache, wenn beispielsweise bei Berufsberatungstests auf die Gesamtpopulation geschlossen werden soll, da tatsächlich alle Personen potenziellen Zugang zu Beratung haben. Schönemann (2006, zur Publikation eingereicht) führt daher aus:

> Though widely embraced, this logic can be questioned on pragmatic grounds: From a decision point of view it matters little how high the validity "really" might have been if the test were given to a sample from the unselected population – since it virtually never is. If the test performs poorly in the validation sample, then it will probably also perform poorly in predicting academic success in the subpopulation of applicants to graduate school. Thus, some may argue, the problem is not so much whether the uncorrected validities make the test "look like a poorer predictor of graduate school outcomes than it really is", but whether the corrected validities make it look better than it really is when it comes to making decisions. (S. 11)

In jedem Fall muss man bei der Interpretation und vor allem vor der Anwendung einer Selektionskorrektur diese beiden Kritikpunkte stets bedenken. Vor dem Hintergrund verzerrter Populationsschätzungen wegen Verletzungen der Voraussetzungen der Selektionskorrektur und der inhaltlich kritisierbaren Logik erscheint es ohnehin am sinnvollsten, den Studierfähigkeitstest schon bei der Konstruktion in seinen Anforderungen und seinem Schwierigkeitsgrad möglichst genau auch auf die hypothetisch erforderlichen Fähigkeiten der Zielpopulation, also der zu selegierenden Personen, zu konzipieren. Dies würde in der Konsequenz verstärkt eine kriteriumorientierte Testung verlangen, würde dann aber sicherlich auch zu einem Mehr an inhaltlichen Überlegungen über die Anforderungen eines spezifischen Studiums führen, anstatt die Bewerber lediglich in eine Rangreihe zu bringen.

4.2.4 Zusammenfassung

Die geschilderten Befunde weisen für Studierfähigkeitstest generell Validitäten aus, die im mittleren Bereich und in der Differenz um ca. $r = .08$ unter derjenigen der Schulabschlussnote liegen. Trost (2003, S. 39) verweist darauf, dass Studierfähigkeitstests in ihrer Validität damit unmittelbar hinter Schulabschlussnoten rangieren. Eine Kombination beider Prädiktoren liegt daher nahe, um die Erfolgsprognose insgesamt zu verbessern. Im Falle des TMS konnten Befunde diese Hypothese auch unterstützen. Der *generellen* Untersuchung dieser Fragestellung ist allerdings ein separates Teilkapitel gewidmet.

Die Frage, ob eher allgemeine oder studienfachspezifische Studierfähigkeitstests eingesetzt werden sollen, wird in der Literatur kontrovers diskutiert (s. z.B. Burton & Ramist, 2001; Deidesheimer Kreis, 1997; Kuncel, Hezlett & Ones, 2001; Rindermann & Oubaid, 1999; Trost, 1975, 2003). Autoren aus dem deutschsprachigen Raum favorisieren meist studienfachspezifische Studierfähigkeitstests (s. besonders Deidesheimer Kreis, S. 105), da die Validitäten von letzteren nach der bisherigen Befundlage leicht über denen von allgemeinen liegen (vgl. auch Tabelle 2 und Tabelle 3). Allerdings konnten Hell et al. (2005) in einer Metaanalyse keinen Unterschied zwischen allgemeinen und spezifischen Studierfähigkeitstests ausmachen. Beide wiesen eine durchschnittliche Validität von $r = .43$ auf, wobei jedoch die Validitätskoeffizienten der allgemeinen Studierfähigkeitstests breiter streuten. Zur weiteren Klärung dieser Frage bedarf es daher noch weiterer Forschung.

4.3 Persönlichkeitsmerkmale

Die im Folgenden behandelten Persönlichkeitsmerkmale sollen sich auf solche im „engeren Sinne“ (Amelang & Zielinski, 1997, S. 252) beschränken. Gewohnheiten, Einstellungen und Lebensdaten sollen, obgleich sie zum weiter gefassten Bereich von Persönlichkeitsmerkmalen gehören, ausgespart bleiben. Vielmehr werden hier Dimensionen des *emotionalen, motivationalen und sozialen Verhaltens* dargestellt, denen gemeinsam ist, dass sie keine normative Wertigkeit beinhalten, „sondern die Richtung sowie Art und Weise, in der dieses [emotionale, motivationale und soziale Verhalten] geschieht“ (Amelang & Zielinksi, 1997, S. 252) von Bedeutung ist.
Bislang wurden Verfahren zur expliziten Erfassung von Persönlichkeitsmerkmalen aus den oben genannten Bereichen nicht direkt in Auswahlverfahren integriert, obgleich ihr indirekter Einfluss auf die Beurteilung z.B. in Assessment Centern und Auswahlgesprächen belegt ist (Scholz & Schuler, zit. nach Schuler, 1992, S. 259, sowie Schuler, 2002, S. 247). Zum überwiegenden Teil begründet sich der Verzicht auf Persönlichkeitsmerkmale in Auswahlverfahren auf ihrer sehr leichten Verfälschbarkeit bzw. Trainierbarkeit im Sinne sozial erwünschter Antworten (Trost, 2005, S. 139). Hierauf stützen sich auch rechtliche Beschränkungen, denn Verfahren, welche sich als hochgradig verfälschbar und trainierbar erwiesen, würden sehr rasch von Gerichten für unzulässig erklärt werden (Trost, 2005, S. 139). Auch widerspricht ihr Einsatz dem Hochschulrahmengesetz, da dieses „fachspezifische Studierfähigkeitstest[s]“ (Reich, 2004, S. 305) verlangt. Jedoch verweist der Wissenschaftsrat (2004, S. 50) darauf, dass u.a. Persönlichkeitstests „gleichwohl – gegebenenfalls im Rahmen von Modellversuchen – intensiver erprobt werden [sollten]“. Wesentlich scheint in diesem Zusammenhang also nicht so sehr die Frage danach, *ob* Fragebogendaten verfälschbar sind, sondern, *wie stark* sich dies insbesondere auf ihre prädiktive Validität auswirkt.

Im Folgenden werden zunächst Forschungsergebnisse zur Objektivität, Reliabilität und insbesondere zur prädiktiven Validität von Fragebogendaten referiert. Der Forschungsstand zu ihrer Validität unter Verfälschbarkeitsbedingungen wird in einem separaten Teilkapitel des Methodenteils berichtet, da diese Fragestellung ein empirischer Bestandteil dieser Arbeit ist und erst im Kontext des Untersuchungsdesigns verständlicher wird.

4.3.1 Objektivität und Reliabilität

Nach Amelang und Zielinski (1997, S. 254) können Persönlichkeitsfragebögen in dem Sinne als objektiv gelten, als dass durch die Invarianz der Itemformulierung und des Antwortformates sowie der weitgehenden Reduktion der Probanden-Testleiter-Interaktion eine Durchführungs-, Auswertungs- und Interpretationsobjektivität gegeben ist. Da es jedoch bei der Beantwortung der Items dem Probanden überlassen bleibt, welche Referenzkategorien, Ereignisse o.ä. er zur Beantwortung heranzieht, sind Fragebögen in diesem Sinne auch als subjektive Verfahren zu bezeichnen, „die aber wegen ihrer psychometrischen Objektivität eine Sonderstellung innehaben“ (Amelang & Zielinski, 1997, S. 254). Aufgrund dieser starken subjektiven Komponente, welche durch eine ganze Reihe von Subprozessen wie vielschichtiger Urteilsprozesse und (idiosynkratisch) variierenden Bezugssysteme (etwa eigene Verhaltensstichproben gegenüber normativer Relativierung) gekennzeichnet sind, liegen insbesondere die Reliabilitäten der Verfahren meist unter denen von Leistungstests. Die internen Konsistenzen variieren zwischen .60 und .80, Retestreliabilitäten zwischen .50 und .70 (Amelang & Zielinski, 1997, S. 255). Bei letzteren bleibt zudem nach Janke (1973, zit. nach Amelang & Zielinksi, 1997, S. 255) fraglich, ob die Retestreliabilitäten für konstantes Urteilsverhalten, Gedächtniseffekte oder tatsächlich für die Konstanz des Verhaltens stehen. Alleine wegen der geringen Reliabilitäten ist mit niedrigeren Validitätskoeffizienten als bei Leistungstests zu rechnen.

4.3.2 Prädiktive Validität

Die meiste Forschung zur prädiktiven Validität von Persönlichkeitsfragebögen stammt aus dem Bereich der beruflichen Eignungsdiagnostik. Die nachfolgende Übersicht (Tabelle 4) berichtet dagegen überblicksartig Studien aus dem für diese Arbeit relevanten Bereich des Studienerfolgs.

Tabelle 4: Prädiktive Validitäten von Fragebogendaten in Bezug auf Studienerfolg
(Fortsetzung der Tabelle auf folgender Seite)

Autorenname & Erscheinungsjahr, Studientypus	**Prädiktor**	**Kriterium**	**Korrelation**	**Korrekturen**	**Stichprobe/N**
Schmidt-Atzert (2005)	Leistungs-motivation	Zwischen-prüfungsnote	r = .25	-	57
Friday (2005)	Verträglichkeit	Zwischen- bzw. Abschluss-prüfungsnoten	r = .17 bis r = .28	-	79
Robbins, Lauver, Davis, Langeley und Carlstrom (2004) Metaanalyse	Akademische Selbstwirksamkeit Generelles Selbstkonzept	Durchschnitts-punktzahl nach dem 1. Collegejahr	r = .37 r = .03	Minderung in Y	9 598 9 621
Chamorro-Premuzic und Furnham (2003a)	Gewissenhaftigkeit Neurotizismus	Zwischen- bzw. Abschluss-prüfungsnoten	r = .36 r = -.16	-	542
Trapmann, Hell, Schuler (2005b)	Gewissenhaftigkeit Offenheit Introversion Verträglichkeit Emotionale Stabilität	Zwischen- bzw. Abschluss-prüfungsnoten	r = .25 r = .16 r = .06 r = .06 n.s.	Minderung in Y, Minderung in X	247
Trapmann, et al. (2005a)	Selbstdisziplin Kompetenz Leistungsstreben Pflichtbewusstsein Offenheit für Ideen Ordnungsliebe	Zwischen- bzw. Abschluss-prüfungsnoten	r = .31 r = .27 r = .24 r = .22 r = .15 r = .07	Minderung in Y, Minderung in X	10 011 bis 14 942
Trapmann, et al. (2005a)	Emotionale Stabilität Introversion Offenheit	Studienabbruch	r = .06 r = .02 r = .03		1 681 bis 13 474
Trapmann, et al. (2005a)	Selbstdisziplin Emotionale Stabilität Leistungsstreben Durchsetzungsfähigkeit Extraversion	Studien-zufriedenheit	r = .38 r = .37 r = .31 r = .28 r = .10		2 194 bis 3 916
Giesen, Gold, Hummer und Jansen (1986)	Klasse nicht-intellektueller Merkmale[1]	Studien-zufriedenheit	mittleres R = .48	-	85 bis 357; verschiedene Fachrichtungen

Autorenname & Erscheinungsjahr, Studientypus	Prädiktor	Kriterium	Korrelation	Korrekturen	Stichprobe/N
Schiefele, Krapp und Schreyer (1993, zit. nach Rindermann & Oubaid)	Studieninteressentest	Zwischen-prüfungsnote	mittleres *r* = .33		keine Angaben

Anmerkung. Studien*noten* wurden umgepolt, sodass höhere Werte für bessere Leistungen stehen. Minderung in Y, Minderung in X: Minderungskorrektur im Kriterium (Y) und Prädiktor (X). [1]: Fleiß, fachliches Selbstbewusstsein, emotionale Stabilität, positive Umweltbeurteilung

Die Validitätskoeffizienten liegen für das Kriterium Studienleistungen im niedrigen bis maximal moderaten Bereich. Etwas höher fallen die Werte aus, wenn spezifisch auf die Arbeitshaltung bezogene Konstrukte wie Selbstdisziplin, selbst eingeschätzte Kompetenz, Interessen u.ä. gegenüber breiteren Persönlichkeitsdimensionen wie Neurotizismus, Extraversion, Offenheit etc. erfragt werden (vgl. etwa die entsprechenden Koeffizienten in der Metaanalyse von Trapmann et al., 2005a und in der Studie von Schiefele et al., 1993). Studienabbruch lässt sich anhand der verwendeten Persönlichkeitskonstrukte praktisch nicht vorhersagen, wie die Metaanalyse von Trapmann et al. (2005a) zeigt. Höhere Koeffizienten ergeben sich auch, wenn das „weichere" Kriterium der Studienzufriedenheit herangezogen wird, was besonders in der Studie von Giesen et al. (1986) deutlich wird. Heise, Westermann, Spies und Schiffler (1997, zit. nach Rindermann & Oubaid, 1999, S. 181) konnten zudem für verschiedene Studienfächer zeigen, dass fachspezifische Interessen den bedeutsamsten Prädiktor für die Studienzufriedenheit mit den Studienfachinhalten darstellen. Rindermann und Oubaid (1999, S. 181f.) interpretieren die Forschungsbefunde daher auch insgesamt betrachtet dahingehend, dass Persönlichkeitsmerkmale und Interessen sich insbesondere zur Vorhersage der Studienzufriedenheit eignen und ihren Zweck außerhalb des Selektionskontextes erfüllen könnten.

4.3.3 Zusammenfassung

Insgesamt betrachtet zeigen durch Fragebögen erfasste Persönlichkeitsmerkmale allenfalls moderate Zusammenhänge zu Studienleistungen. Höhere Koeffizienten ergeben sich für das Kriterium der Studienzufriedenheit. Allerdings beschränken sich alle hier berichteten Ergebnisse auf Studien außerhalb des Selektionskontextes. Wie groß sich der Einfluss von

Antwortverfälschungen in Richtung sozialer Erwünschtheit auf die Kriteriumsvaliditäten von Fragebogendaten auswirkt, bleibt bei diesen gewissermaßen „idealen“, um nicht zu sagen „idealistischen“ Erhebungsdesigns unbeantwortet. Zwar existiert bereits eine umfangreiche Literatur zu den Auswirkungen von Verfälschbarkeit von Fragebogenantworten auf deren kriteriumsbezogene Validität (s. z.B. Viswesvaran & Ones, 1999), allerdings beziehen sich deren Ergebnisse in überwiegendem Maße auf die Personalauswahl im Berufskontext. Wenig Forschung wurde bislang in Bezug auf den Einsatz von Fragebögen im Rahmen von Studierendenauswahlverfahren betrieben (s. z.B. Peeters & Lievens, 2005). Die Übertragbarkeit der Ergebnisse zur Verfälschbarkeit von Fragebogenantworten aus dem Berufskontext auf den Studienkontext ist daher eine weiterhin *empirisch* zu untersuchende Frage und soll auch in der vorliegenden Arbeit Forschungsgegenstand sein. Auch wenn der Wissenschaftsrat (2004, S. 50) Forschung zum Einsatz von Fragebogendaten im Kontext von Studierendenauswahlverfahren fordert, so soll angesichts der berichteten vergleichsweise geringen Zusammenhänge mit Studienerfolgsmerkmalen an dieser Stelle jedoch betont werden, dass der Forschungsfokus hierbei stärker grundlagenwissenschaftlich als anwendungsbezogen ausgerichtet ist. Die grundlagenwissenschaftliche Perspektive ist insbesondere deshalb von Interesse, um sowohl im Hinblick auf unterschiedliche Leistungsbeurteilungskontexte (beruflich gegenüber universitär) als auch in Bezug auf unterschiedliche Zielpopulationen (Berufs- gegenüber Studienbewerbern) zu übertragbaren oder aber zu differenzierten Aussagen der Effekte von Antwortverfälschbarkeit auf Fragebogenvaliditäten zu gelangen.
Eine Übersicht über Untersuchungsdesigns, bisheriger Befunde und Rationale der Analyse zur Antwortverfälschbarkeit in dieser Arbeit gibt Kapitel 8.1.5.7.

4.4 Kombination von Prädiktoren

Die bisherige Darstellung beschränkte sich bislang auf die prädiktive Validität einzelner Prädiktoren. Es liegt aufgrund der Befundlage insbesondere von Schulabschlussnoten und allgemeiner wie spezifischer Studierfähigkeitstests nahe, hierbei nach der inkrementellen Validität als dem Zuwachs an prognostischer Validität durch Hinzunahme von Auswahltests neben der Schulabschlussnote zu fragen (Trost, 1975, S. 46). Da Studierfähigkeitstests in ihren Validitäten an zweiter Stelle hinter Schulabschlussnoten rangieren (s. Trost, 2005, S. 39 sowie Kapitel 4.2.3) besteht durch sie die beste Aussicht auf einen Zuwachs an Vorhersagegenauigkeit für Studiennoten. Der Einsatz von Studierfähigkeitstests wäre allerdings nur dann

zu rechtfertigen, wenn sich durch sie ein signifikantes Validitätsinkrement ergäbe. Wie *groß* das Inkrement ausfallen muss, um nicht nur eine statistisch, sondern ebenso praktisch bedeutsame Verbesserung in der Vorhersage zu erzielen, ist letztlich eine Frage der Selektionsrate (als Anzahl der Bewerber zur Anzahl an Studienplätzen) und der Basisrate (als Anteil der ohnehin also auch ohne Selektion erfolgreichen Bewerber) (Meehl & Rosen, 1955; Schönemann, 1997b; Taylor & Russell, 1938, 1939). In diesem Zusammenhang merkt Stemmler (2005, S. 126) unter Bezugnahme auf die Arbeiten von Taylor und Russel (1938, 1939) an, dass Studierendenauswahlverfahren zur Senkung der Abbruchquoten in verschiedenen Fächern an deutschen Universitäten eine Mindestvalidität von $r = .50$ aufweisen müssten, in Humanmedizin sogar eine von geringstenfalls $r = .70$. Zur Beantwortung des praktischen Aspektes können daher statistische Signifikanzen oder Effektstärken keine alleinige Antwort geben; es müssen stets auch diese Variablen und nicht zuletzt kostenökonomische Analysen in die Überlegung miteinbezogen werden, ob sich eine die Anwendung eines Studierfähigkeitstests lohnt (zu kostenökonomischen Analysen von Auswahlverfahren siehe u.a. Amelang & Zielinski, 1997, S. 369ff.).

Eine notwendige (aber keine hinreichende) Bedingung für inkrementelle Validitäten stellt die weitgehende Unkorreliertheit der Prädiktoren dar. Eine hohe Korrelation zwischen diesen würde auf eine weitgehende Überschneidung des Messbereichs hinweisen, wodurch die Wahrscheinlichkeit, einen eigenständigen Vorhersagebeitrag durch ein Testverfahren zu identifizieren, deutlich verringert wäre.
Die in der Literatur berichteten Korrelationen verschiedener Studierfähigkeitstest mit Schulabschlussnoten weisen tatsächlich auch auf eine weitgehende Unabhängigkeit beider Prädiktoren hin. So lag die durchschnittliche Korrelation des TMS mit der (umgepolten) Abiturdurchschnittsnote bei $r = .39$ (Deidesheimer Kreis, S. 113), diejenige des mittlerweile in der Schweiz durchgeführten TMS (dort Eignungstest für Medizinische Studiengänge, EMS, genannt) beträgt $r = .48$ (Hänsgen, 2006). Mit gemeinsam geteilten Varianzanteilen von 15% bzw. 23% erweist sich daher die Überlappung der Messbereiche als relativ gering. Im TMS kamen also überwiegend Fähigkeiten zum Ausdruck, die in der schulischen Leistungsbewertung keinen Niederschlag fanden, die jedoch wichtig für den Erfolg in der Ärztlichen Vorprüfung waren, wie die Validitätskoeffizienten belegen (s. Tabelle 3).
Die Korrelation zwischen dem SAT und dem ACT mit dem High school grade Point average liegt nach einer Metaanalyse von Robbins, Lauver, Davis, Langley und Carlstrom

(2004, S. 272) bei $r = .46$. Auch hier kann also von einer weitgehenden Überschneidungsfreiheit der Messbereiche von Schulabschlussleistungen und Studierfähigkeitstests ausgegangen werden.

Im Folgenden werden die empirisch ermittelten inkrementellen Validitäten allgemeiner und spezifischer Studierfähigkeitstests berichtet. Im Anschluss daran werden einige Studien zu praktischen Implikationen einer Testapplikation zusätzlich zur Schulabschlussleistung berichtet. Die umfassendsten Studien hierzu stammen aus den USA, weshalb hierauf der Schwerpunkt der Darstellung liegt. Für den deutschsprachigen Raum stehen die Ergebnisse aus der Forschung um den TMS im Vordergrund, da für ihn die Forschungsliteratur am ergiebigsten ist.
Tabelle 5 gibt eine Übersicht über die Ergebnisse von Validitäten einzelner Prädiktoren im Vergleich zu Prädiktorkombinationen.

Tabelle 5: Prognostische Validität von Einzelprädiktoren und deren Kombinationen in Bezug auf Studienleistungen (Fortsetzung der Tabelle auf folgender Seite)

Autorenname & Erscheinungsjahr; Studientypus	**Prädiktor**	**Kriterium**	**Korrelation**	**Korrekturen**	***N***
Kuncel, Hezlett und Ones (2001) Metaanalyse	GRE-Gesamtwert unit weighted composite nach Nunally (1978)	Abschlussnoten	mittleres $r = .50$		879 bis 14 425
	Durchschnittliche Zwischenpunktzahl an der Hochschule		nicht angegeben	Minderung in Y, Selektionsk.	
	Kombination (unit-weighted composite nach Nunally, 1978)		mittleres $\boldsymbol{r = .54}$		
Burton und Ramist (2001) Metaanalyse	HGPA	Durchschnittspunktzahl nach dem 1. Collegejahr	mittleres $r = .42$	-	HGPA: $N = 25\,175$ SAT I: $N = 16\,995$ Kombi.: $N = 17\,649$
	SAT I (Reasoning)		mittleres $r = .36$		
	Kombination		mittleres $\boldsymbol{R = .52}$		
Geiser und Studley (2002)	HGPA	Durchschnittspunktzahl nach dem 1. Collegejahr	$r = .39$	-	77 893 (Jahrgänge 1996-1999)
	SAT I (Reasoning)		$r = .36$		
	Kombination		$\boldsymbol{R = .45}$		

Autorenname & Erscheinungsjahr; Studientypus	**Prädiktor**	**Kriterium**	**Korrelation**	**Korrekturen**	***N***
Baron und Norman (1992)	HGPA	Durchschnitts-punktzahl nach fünf Jahren	*r* = .30	-	4 170
	SAT (Reasoning)		*r* = .20		
	Kombination		***R* = .33**		
Beller (1993)	Schulabschlussnote	Durchschnittspunktzahl nach dem 1. Collegejahr	*r* = .38	Selektionsk.	52 985
	PET		*r* = .32		
	Kombination (1:1)		***R* = .48**		
Crouse und Trusheim (1988)	Schulabschlussnote	Bachelor Degree Attainment	*r* = .23	-	2 470
	SAT (Reasoning)		*r* = .22		
	Kombination		***R* = .26**		
Trost, Blum, Fay, Klieme, Maichle, Meyer und Nauels (1998)	Abiturnote	Zwischenprüfung in Humanmedizin	*r* = .47	Selektionsk.	27 876
	TMS		*r* = .45		
	Kombination (55:45)		***r* = .54**		
Trost et al. (1998)	Abiturnote	Zwischenprüfung Tiermedizin	*r* = .51	Selektionsk.	2 580
	TMS		*r* = .41		
	Kombination (55:45)		***r* = .54**		
Trost et al. (1998)	Abiturnote	Zwischenprüfung Zahnmedizin	*r* = .37	Selektionsk.	5 533
	TMS		*r* = .32		
	Kombination (55:45)		***r* = .41**		
	Gesamttest		*r* = .28		
	Kombination		***R* = .44**		

Anmerkung. GRE: General Record Examination; HGPA: High School Grade Point Average; TMS: Test für medizinische Studiengänge; SAT I (Reasoning): Scholastic Aptitude Test Reasoning Part; PET: Psychometric Entrance Test.
Minderung in Y: Minderungskorrektur im Kriterium (Y);
Minderung in Y, Minderung in X: Minderungskorrektur im Kriterium (Y) und Prädiktor (X).

In allen Studien kommt es zu einer Erhöhung der Prognosekraft durch zusätzliche Aufnahme von Studierfähigkeitstest als Prädiktoren der Studienleistung. Im Falle des TMS wurde auf eine zufallskritische Absicherung dieser inkrementellen Validität verzichtet, da die Testteilnahme für alle Studienplatzbewerber obligatorisch war und somit stets Populationskennwerte

vorlagen. In den anderen Studien wurden keine derartigen Signifikanztestungen berichtet; allerdings ist bei den durchweg sehr großen Stichproben von der Signifikanz des Inkrements auszugehen. Betrachtet man den Bereich der inkrementellen Validität, so variiert diese zwischen 3% und 10% (vgl. Inkrement hinsichtlich Zwischenprüfung in Tiermedizin bei Trost et al, 1998 sowie die Studie von Bellar, 1993). Hell et al. (2005) konnten zudem in ihrer Meta-analyse eine mittlere durchschnittliche Validität der Kombination aus Schulnoten und Test-leistungen von $r = .52$ feststellen.

Crouse und Trusheim (1988, S. 40ff.) stellen die Frage nach den *praktischen Implikationen* der inkrementellen Validität des SAT zum HGPA, indem sie darauf hinweisen, dass „Admission officials make correct admissions forecasts when they accept applicants who earn satisfactory freshman grades and eventually graduate" (Crouse & Trousheim, 1988, S. 54). Ein Argument übrigens, welches auch von Rindermann und Oubaid (1999, S. 186) vorgetragen wird, wenn sie gegen die Vorselektion zur Testteilnahme anhand der Abiturnote plädieren, da „in diesem Falle ... jedoch schwache Schulleistungen nicht mehr durch überzeugende Ergebnisse im Auswahlverfahren ausgleichbar [wären]".
Crouse und Trusheim (1988, S. 54) vergleichen daher anhand einer Stichprobe von $N = 2\,470$ High-School-Seniors die Verbesserung der *Anzahl korrekt als geeignet identifizierter Personen* („Correct admissions", S. 54) bei jeweils variierenden Kriteriumstrennwerten durch Hinzunahme der SAT-Punktwerte zum HGPA. Als Trennwert für Erfolg gegenüber Misserfolg legen sie den Mittelwert der Punktwerteverteilung nach dem ersten College-Jahr fest ($M = 2.5$ bei einem Skalen-Range von 1 bis 4). Unter alleiniger Verwendung des HGPA als Prädiktor dieses Hochschulzulassungskriteriums resultieren 62.2% korrekte Zulassungen. Dies ent-spricht bei 100 Bewerbern einer Verbesserung in der Anzahl korrekter Zulassungen von 9.2 verglichen mit einer reinen Zufallsauswahl aus der gesamten Bewerberstichprobe. Die Hinzunahme der SAT-Punktwerte verbesserte die Quote auf 64.6%, was 11.9 als geeignet identifizierter Personen bei 100 Bewerbern gegenüber einer Zufallsauswahl entspricht. Konkreter gesprochen: durch die zusätzliche Verwendung des SAT zum HGPA würden bei 100 Bewerbern 2.7 mehr richtige Zulassungen resultieren. Setzte man den Kriteriumswert um 0.5 Einheiten höher an, sinkt dieser Wert leicht auf 2.2 (Crouse & Trusheim, 1988, S. 55). Auch für Trennwerte unterhalb des Kriteriumsmittelwertes resultieren Werte korrekter Zulassungen von ca. zwei.
In gleicher Weise analysierten die Autoren die Verbesserung korrekter Zulassungen durch Hinzunahmen des SAT hinsichtlich des Kriteriums eines Bachelor-Abschlusses. Die Anzahl als geeignet identifizierter Personen würde sich dadurch bei 100 Bewerbern auf 0.1 verbessern.

Auch eine Verschärfung des Kriteriumstrennwertes von 2.5 auf 3.0 würde diesen Wert lediglich auf 0.2 erhöhen. Crouse und Trusheim (S. 68) schlussfolgern aus diesen Ergebnissen: „The result is that colleges that admit applicants according to their academic eligibility now benefit little from use of the SAT over high school record alone".

Baird (1985) ging der Frage nach, ob Universitäten durch zusätzliche Beachtung der Testergebnisse aus Studierfähigkeitstests *Bewerber mit voraussichtlich exzellenten Studienleistungen besser identifizieren* können; ein Aspekt von Studierendenauswahlverfahren also, der vor dem Hintergrund der Exzellenzinitiative (Kultusministerkonferenz, 2004) auch hierzulande diskutiert wird. Baird (1985) verglich in seiner Untersuchung die akademischen Leistungen von mehr als 14 000 Studierenden, welche aufgrund ihres exzellenten Abschneidens im ACT für ein Elite-Studien-Programm ausgewählt worden waren (mit einem ACT-Durchschnitt dieser Stichprobe von 25.5 Punkten bei 30 maximal erreichbaren) mit denen von 10 000 Studierenden, welche wegen einer mittelmäßigen Leistung von im Schnitt 18 ACT-Punkten nicht zugelassen worden waren. Baird (1985, S. 52) führt zu den Ergebnissen aus: „The distribution of the number of accomplishments were very similar in both groups, although there were small significant differences favoring the high-ability group in writing, leadership, and science", woraus Baird (1985, S. 52) schloss: „…in no case did academic ability (ACT scores) account for as much a 1% of the variance in accomplishment".
Zu ähnlichen Ergebnissen gelangen auch Crouse und Trusheim (1988, S. 60), als sie die Auswirkungen der zusätzlichen Verwendung des SAT auf die durchschnittliche Punktzahl nach dem ersten College-Jahr untersuchten. Sie schließen aus den Analysen: „The gain from adding the SAT is therefore only an increase in average freshman-grades of 0.02 on a four-point scale…"

Crouse und Trusheim (1988, S. 51f.) untersuchten weiterhin, *ob es durch die zusätzliche Verwendung des SAT zum HGPA zu anderen Zulassungsentscheidungen kommt* als bei alleiniger Beachtung des HGPA. Der SAT könnte nämlich als „second chance" (Linn, 2005, S. 151) für Bewerber fungieren, die etwa wegen schlechterer schulischer Lernumgebungen o.ä. durch alleinige Beachtung des HGPA bei der Zulassung benachteiligt wären. Bei ihrer Analyse verglichen die Autoren (Crouse & Trusheim, 1988, S. 51) die jeweiligen Zulassungsentscheidungen für verschieden strenge Kriteriumstrennwerte bei alleiniger Beachtung des HGPA und bei einer Kombination aus HGPA und SAT-Ergebnis. Als Minimalwert identischer Zulassungsentscheidungen resultierte 83.8%, als Maximalwert 98.4%, weshalb die Autoren schlussfolgerten: „Colleges therefore make identical admissions decisions, either to admit or

reject, on a great majority of their applicants whether they use the SAT along with the high school record, or the high school record alone" (Crouse & Trusheim, 1988, S. 53).
Bridgeman, Burton und Cline (2001, 2004) analysierten von diesen Befunden ausgehend und in gleicher Weise, ob es durch die Kombination des HGPA mit dem SAT II gegenüber einer mit dem SAT I zu anderen Auswahlentscheidungen kommt. Linn (2005) fasst diese Befunde dahingehend zusammen, indem er konstatiert:

> [they] found that the students who would be selected using high-school grades and the SAT II test average would be mostly the same as those who would be selected using high-school grades and the SAT I ... Overall, however, prediction does not provide a strong basis for preferring one test to another. (S. 147)

Diese Befunde sind insofern beachtlich, als dass auch bei einem Wechsel vom *allgemeinen* Studierfähigkeitstest (SAT I) zu einem *spezifischen* (SAT II) weitgehend identische Zulassungsentscheidungen resultieren.

Der Gewinn an Vorhersagegenauigkeit durch *Kombination von mehreren Studierfähigkeitstests* zum HGPA untersuchten Bridgeman, Burton und Cline (2001, 2004) sowie Kobrin, Camara und Milewski (2006) in Bezug auf die zusätzliche Aufnahme des SAT II zur Kombination von HGPA und SAT I. Bridgeman, Burton und Cline (2001, 2004) fanden jeweils nur geringe Zuwächse an prognostischer Validität. Auch bei Geiser und Studley (2001, S. 6) liegt der aufgeklärte Varianzanteil durch Aufnahme des SAT II bei 22.3% gegenüber einem von 22.2% für die Kombination aus HGPA und SAT I.
Diesbezüglich ist der Äußerung Fishmans und Pasanellas (1960, zit. nach Trost, 1975, S. 46-47) zuzustimmen, wenn sie bemerken: „It seems useless ... to employ more than two or three intellective predictors, from both point of view of practicality and that of efficiency".

4.4.1.Zusammenfassung

Allgemein lässt sich aus den referierten Befunden erkennen, dass die Prognose des Studienerfolgs aufgrund der Kombination aus Schulabschlussleistungen und Studierfähigkeitstests besser gelingt als aufgrund eines Prädiktors alleine. Trost (1975) merkt jedoch an:

> Allerdings ist die Validität der kombinierten Prädiktoren noch immer zu niedrig, um individuelle Entscheidungen – etwa im Rahmen einer verbesserten Studienberatung – alleine auf dieser Basis zu rechtfertigen .... Auch für institutionelle Entscheidungen – etwa bei der Hochschulzulassung oder der Aufnahme in ein Stipendium – ist eine höhere Validität der Prädiktoren wünschenswert. (S. 46)

Dieser Einschätzung ist gerade vor dem Hintergrund der dargestellten sehr geringen *praktischen* Implikationen auch signifikanter inkrementeller Validitäten zuzustimmen. Wohl auch vor diesem Hintergrund plädieren daher Rinderman und Oubaid (1999, S. 185ff.) für ein Auswahlverfahren mit mehreren *multimodalen* Komponenten, welches sowohl die Abiturdurchschnittsnote als auch Eignungstests und Auswahlgespräche umfasst. Dieses sogenannte Abitur-Test-Interview-Modell (ATIM) soll zudem sequenziell angelegt sein: 40% der Bewerber erhalten einen Studienplatz aufgrund ihrer Abiturdurchschnittsnote; weitere 40% werden auf der Grundlage einer Kombination aus Abiturdurchschnittsnote und studienfachspezifischem Studierfähigkeitstest verteil; die übrigen 20% der Studienplätze werden über Interviews zugewiesen, deren Inhalte berufliche Motivation, Interessen, Studienerwartungen und spezifische Erfahrungen sind. Die angegebenen Quoten können allerdings den Umständen des jeweiligen Studiengangs angepasst werden. Zur empirischen Bewährung dieser Art von Prädiktorkombination liegen allerdings noch keine Befunde vor, da im deutschsprachigem Raum lediglich die Universität Freiburg ein Pilotprojekt zur Implementierung eines derartigen Auswahlverfahrens begonnen hat (Pixner, Zapf & Schüpbach, 2005) und an der Universität Heidelberg begrenzte Vorarbeiten angebahnt sind (s. hierzu Amelang & Funke, 2005).

Als Konsequenz aus den geschilderten Befunden verstehen sich die in dieser Arbeit entwickelten Testverfahren lediglich als ein Baustein eines breiter konzeptualisierten sequenziellen und multimodalen Auswahlverfahrens im Sinne Rindermann und Oubaids (1999, S. 185ff.) bzw. eines spezifisch für die Heidelberger Universität entworfenen nach Amelang und Funke (2005). In der gedanklichen Weiterführung obigen Zitates von Trost (1975, S. 46) muss man gerade im Lichte der bisherigen Ergebnisse zu Studierendenauswahlverfahren feststellen, dass mindestens befriedigend präzise individuelle Studieneignungsdiagnosen von nicht-sequenziellen und unimodalen Auswahlstrategien nicht zu erwarten sind. Studienerfolg wird von zu vielen individuellen und institutionellen Bedingungsfaktoren modelliert, als dass sich eine Studieneignungsdiagnose auf eine sehr

begrenzte Prädiktorauswahl beschränken sollte. Wie bereits Brigham, einer der „Urväter“ des SAT, bemerkte, sollten Studierfähigkeitstest viel eher als „a ready method of interview“ denn als „a measure of some mysterious power“ (Brigham, zit. nach Lemann, 1999, S. 34) angesehen werden. Allgemeine oder spezifische Studierfähigkeitstests sollten daher nicht alleine mit Schulabschlussnoten kombiniert werden und zusammen mit diesen über Zulassung oder Ablehnung eines Studienplatzbewerbers entscheiden, sondern Bestandteil eines sequenziellen und multimodalen Auswahlverfahrens sein.

4.5 Allgemeine Zusammenfassung und abschließende Anmerkungen

Auf Basis aller hier dargestellten Befunde zur prognostischen Validität von Studierenden-auswahlverfahren lassen sich folgende Aussagen treffen:

- Die höchste prädiktive Validität in Bezug auf Studiennoten erzielt die Schulabschlussnote. Die Korrelationen liegen überwiegend im Intervall von $r = .40$ bis $r = .50$.
- An zweiter Stelle liegen Studierfähigkeitstests. Unklar ist bislang noch, ob allgemeine oder studienfachspezifische Studierfähigkeitstests bessere prädiktive Validitäten aufweisen; dies kann nach dem derzeitigen Stand der Forschung noch nicht abschließend beurteilt werden. Die Metaanalyse von Hell et al. (2005) findet keine Unterschiede (mittlere Kriteriumskorrelation jwls. $r = .46$), wobei die Varianz der Validitäts-koeffizienten von allgemeinen Studierfähigkeitstests gegenüber spezifischen größer ausfällt.
- Persönlichkeitsmerkmale zeigen lediglich niedrige bis moderate Zusammenhänge mit Studienerfolgsmaßen. Moderate Korrelationen ergeben sich hierbei meist mit „weicheren“ Studienerfolgsmaßen wie der Studienzufriedenheit.
- Eine Kombination aus Schulabschlussnote und Ergebnis in einem Studierfähigkeitstest ergibt in den meisten Fällen eine Erhöhung der Vorhersagegenauigkeit. Die praktischen Implikationen der berichteten inkrementellen Validitäten sind nach US-amerikanischen Studien unter sehr verschiedenen Beurteilungsaspekten allerdings als sehr gering einzustufen. Auch eine Kombination mit mehr als einem Studierfähigkeitstest erbringt lediglich sehr geringe Zuwächse an prognostischer Validität und kann daher auch nicht zu einer deutlichen praktischen Verbesserung führen.

Möglicherweise liegt in der Kombination von multimodalen Methoden in einem sequenziellen Auswahlverfahren im Sinne von Rindermann und Oubaid (1999) die Chance auf eine deutliche Verbesserung der Vorhersage, zumal hierbei auch nicht-kognitive, aber relevante Prädiktoren des Studienerfolgs etwa über Selbst-selektionsprozesse durch ausgedehnte Vorabinformationen über das Studium sowie strukturierte oder teilstrukturierte Interviews miteinbezogen werden könnten.

5. Zielsetzung der Arbeit

Die Kapitel 4 bis 4.5 stellten die bisherige Befundlage zur Vorhersage des Studienerfolgs dar. Der höchste prognostische Wert unter den bislang verwendeten Prädiktoren ergibt sich für die Abiturdurchschnittsnote und kann als befriedigend angesehen werden. Jedoch muss angesichts der mangelnden Vergleichbarkeit von Schulabschlussnoten zwischen den Bundesländern mit daraus resultierenden unterschiedlichen Fähigkeitsniveaus (Prenzel et al., 2005) die alleinige Verwendung dieses Prädiktors als Zulassungskriterium kritisch gesehen werden, werden die bestehenden Mittelwertsunterschiede nicht durch ein Bonus- oder Malussystem ausgeglichen. Darüber hinaus streuen die Validitätskoeffizienten der Abiturdurchschnittsnote je nach Strukturierungsgrad des Studiums relativ breit (Rindermann & Oubaid, 1999, S. 178), wodurch die Vorhersage für ein spezifisches Studienfach durch die Abiturdurchschnittsnote nicht immer hinreichend ist. Daher scheint die zusätzliche Beachtung weiterer Informationsquellen über studienerfolgsrelevante Personenmerkmale angeraten, um eine breitere und solidere Basis für Zulassungsentscheidungen zu erhalten.

Die Kombination mehrere Informationsquellen kann jedoch nur dann eine verlässliche Basis von Auswahlentscheidungen darstellen, wenn die psychometrische Qualität der eingesetzten Verfahren gesichert ist. Dabei ist zu bedenken, dass diese nicht einfach auf der Basis von bereits bestehenden Verfahren oder Studienbefunden als gegeben angesehen werden kann. Insbesondere eine Validitätsgeneralisierung aus früheren Befunden setzt eine weitgehende Äquivalenz der Fragestellungen, Untersuchungsgruppen, Verfahren und Kriterien mit dem neuen Auswahlverfahrenskontext voraus, wie die DIN 33430 betont (Deutsches Institut für Normung, 2002, S. 16). Jedes neue Auswahlverfahren bedarf daher einer eingehenden psychometrischen Fundierung, beginnend mit der Spezifizierung der für den Studienerfolg eines Studienfaches hypothetisch relevanten Personenmerkmalen, über die Aufgabenkonstruktion bis hin zur Kriteriumsvalidierung. Der Fokus der vorliegenden Arbeit wird daher auf folgenden Punkten liegen:

- Ableitung von hypothetisch studienerfolgsrelevanten Personenmerkmalen für das Studienfach Psychologie
- Operationalisierung der Personenmerkmale in einer studienfachspezifischen Testbatterie
- Evaluation des Verfahrens:
  - Überprüfung der psychometrischen Güte der Tests nach Maßgabe von Rasch-Modellen
  - Überprüfung der psychometrischen Güte einer Prädiktorkombination aus Abiturdurchschnittsnote und Testleistungen hinsichtlich Studienleistungen.

6. Ableitung der Fragestellung und Hypothesen

Zentrales Anliegen dieses Kapitels ist die Ableitung der Fragestellung sowie entsprechender Hypothesen aus den in den vorigen Kapiteln referierten Befunden und aus der allgemeinen Zielsetzung dieser Arbeit.

Wie bereits in Kapitel 3.1 genannt, besteht sowohl national als auch international insgesamt ein Mangel an theoriegeleiteter Konstruktion fachspezifischer Studierfähigkeitstests, wobei manche Autoren sogar hierin eine Hauptursache für die bislang eher moderaten Zusammenhänge zwischen Studierfähigkeitstests, insbesondere allgemeinen und Maßen des Studienerfolges sehen (Sternberg & Expertise, 2004). Entsprechend den Empfehlungen des Wissenschaftsrates (Wissenschaftsrat, S. 86) und der Deutschen Gesellschaft für Psychologie (Weber & Schmidt-Atzert, 2005) soll deshalb in dieser Arbeit ein auf den Studiengang Psychologie zugeschnittenes Auswahlverfahren konstruiert und evaluiert werden. In Anlehnung an die Methode der kritischen Ereignisse (Flanagan, 1954) soll über eine Anforderungsanalyse eine solide Basis spezifischer Merkmale des Psychologiestudiums geschaffen werden, auf der geeignete Testverfahren konstruiert werden können.

Aus den bestehenden Forschungsbefunden lässt sich zunächst als Basis zur Ableitung weiterer Hypothesen feststellen, dass Schulabschlussnoten einen substanziellen Zusammenhang mit Studienleistungen zeigen werden, demzufolge *Hypothese I* lautet:

Es bestehen substanzielle Zusammenhänge zwischen Schulabschlussnoten und Studienerfolgsmaßen.

Da der Einsatz eines spezifischen Studierfähigkeitstests nur notwendig (wenngleich nicht hinreichend) gerechtfertigt werden kann, wenn auch er bedeutsame Zusammenhänge mit Studienleistungen aufweist, ergibt sich *Hypothese II*:

Zwischen den Testergebnissen und den Maßen des Studienerfolges bestehen substanzielle Zusammenhänge.

Die zusätzliche Anwendung eines Studierfähigkeitstests lässt sich allerdings nicht alleine über einen empirisch nachgewiesenen Zusammenhang mit Studienerfolgsmaßen rechtfertigen. Zusätzlich sollte sich durch seine Berücksichtigung die prognostische Validität des gesamten Auswahlverfahrens inkrementell erhöhen, woraus *Hypothese III* folgt:

Durch Berücksichtigung der Testergebnisse wird die kriteriumsbezogene Validität des Auswahlverfahrens signifikant erhöht.

Da die Konstruktion eines studienfachspezifschen Auswahlverfahrens für Psychologie intendiert ist, sollte dieses auch zwischen Studienanfängern verschiedener Fächer zu differenzieren vermögen. Hierbei sollten die Studienanfänger in Psychologie signifikant bessere Leistungen als solche anderer Fächer zeigen. Denn somit würden die spezifisch auf die Anforderungen des Psychologiestudiums ausgerichteten Testkomponenten typische Leistungsmerkmale und Interessensstrukturen von Bewerbern in diesem Fach erfassen. *Hypothese IV* lautet daher:

Das spezifisch für Psychologie konstruierte Auswahlverfahren vermag zwischen Studienanfängern der Psychologie und solchen anderer Fächer im Sinne eines typischen Profils zu differenzieren. Es wird hierbei erwartet, dass Studienanfänger der Psychologie im Durchschnitt signifikant bessere Leistungen in den Tests erzielen. Diese Unterschiede sollten auch bei statistischer Kontrolle der Abiturdurchschnittsnote bestehen bleiben, daher kein alleiniger Effekt schulischer Leistungsfähigkeit sein.

Es ist weiterhin möglich, dass sich über eine Anforderungsanalyse auch für das Psychologiestudium erfolgskritische *Persönlichkeitsdimensionen* ableiten lassen. In diesem Fall soll deren Einsatzmöglichkeit in Auswahlverfahren im Fragebogenmodus unter

experimenteller Induktion von sozial erwünschtem Antwortverhalten analysiert werden. Die hieraus abgeleitete *Fragestellung* lautet daher:

Behalten hypothetisch für den Erfolg im Psychologiestudium relevante Persönlichkeitsdimensionen ihre Kriteriumsvalidität auch unter experimentell induziertem sozial erwünschtem Antwortverhalten?

7. Anlage der Studie

Studierendenauswahlverfahren, gerade solche aus dem US-amerikanischen Raum, blicken auf eine nunmehr lange Forschungsgeschichte zurück, welche ihren Beginn in den Anfängen des 20. Jahrhunderts nahm und bis heute andauert. Wenngleich die Psychologie hierdurch auf eine reichhaltige Forschungsliteratur zu Studierendenauswahlverfahren und Studienerfolg zurückblickt, so mangelt es gerade in Deutschland an der Konstruktion und Evaluation *fachspezifischer* Auswahlverfahren an öffentlichen Hochschulen (für einen Überblick über fachspezifische Auswahlverfahren an privaten Hochschulen in Deutschland s. Trost, 2003). Die Explizierung fachspezifischer Anforderungen für die Konstruktion von Auswahlverfahren wird nicht zuletzt vom Wissenschaftsrat im Rahmen der Profilbildung von deutschen Hochschulen gefordert (vgl. Wissenschaftsrat, 2004, S. 86), wobei dieser betont, dass hierdurch „für vergleichbare Studienfächer an unterschiedlichen Hochschulen zunehmend auch unterschiedliche Qualifikationsprofile und Kenntnisniveaus zur Voraussetzung eines erfolgreichen Studiums [werden] .... Auch die mit der Hochschulzugangsberechtigung nachgewiesenen Einzelleistungen lassen sich immer weniger mit den Anforderungen einzelner Studiengänge zuverlässig in Beziehung setzen“ (Wissenschaftsrat, 2004, S. 30f). Hieraus schlussfolgert der Wissenschaftsrat schließlich: „...die Notwendigkeit einer bundesweiten Zulassungsbeschränkung [sollte] nicht nur mit Blick auf das Verhältnis von Bewerbern zu vorhandenen Kapazitäten, sondern auch mit Blick auf die Profile der einzelnen betroffenen Studiengänge geprüft werden“ (2004, S. 53).

Diese Forderung aufgreifend, wurde daher am Heidelberger Psychologischen Institut im Frühjahr 2004 beschlossen, ein Pilotprojekt zur Studierendenauswahl zu beginnen und auf Basis einer Anforderungsanalyse ein studienfachspezifisches Auswahlverfahren zu konstruieren und zu evaluieren. Das konkrete Ziel hierbei war die möglichst anforderungsnahe Ableitung von hypothetisch relevanten Personenmerkmalen für den Studienerfolg in

Psychologie an der Universität Heidelberg und der Bestimmung ihres prädiktiven Beitrags in Bezug auf Studiennoten.
In den folgenden Kapiteln dieser Arbeit werden daher die einzelnen Konstruktionsschritte und Evaluationsergebnisse dieses Pilotprojektes dargestellt.

7.1 Prädiktorgewinnung über eine Anforderungsanalyse

Ausgangspunkt aller Entwicklungsarbeiten bildete eine im April 2004 durchgeführte Anforderungsanalyse, die zwei Zielen diente. Zum einen sollten hierüber hypothesengeleitet Prädiktoren gewonnen werden, welche die Basis für die Entwicklung eines studienfach-spezifischen Studierfähigkeitstests für Psychologie darstellen sollten. Zum anderen war beabsichtigt, die abzuleitenden Prädiktoren als Grundlage zur Konkretisierung des vom Wissenschaftsrat (2004) verlangten fachspezifischen Anforderungsprofils zu nehmen, welches die spezifischen Forschungsschwerpunkte des Heidelberger Psychologischen Instituts widerspiegelt, um künftigen Bewerbern als Orientierungsgrundlage für die relevanten Anforderungen zu dienen.

7.1.1 Details der Methode

Um für die anforderungsnahe Ableitung von erfolgskritischen Personenmerkmalen einen möglichst erschöpfenden Prädiktor- und Kriteriumsraum zu schaffen, wurden alle Professoren nebst zwei bis drei Mitarbeitern und die Mitglieder der studentischen Fachschaft in einer Einladung über den allgemeinen Hintergrund des Vorhabens informiert und zur Teilnahme eingeladen. Zum Termin erschien schließlich eine Gruppe von 21 Personen, die sich aus 15 Institutsmitarbeitern und sechs studentischen Vertretern zusammensetzte.
Die Anforderungsanalyse wurde in Anlehnung an die Critical-Incident-Technik von Flanagan (1954) durchgeführt. Die Grundstruktur dieses Vorgehens sieht zunächst die Sammlung erfolgskritischer Ereignisse bzw. Situationen des Psychologiestudiums vor. In einem darauf folgenden Schritt wird nach denjenigen Verhaltensweisen gefragt, welche zu einer Bewältigung des jeweiligen Ereignisses führen bzw. worin sich hierbei erfolgreiche von weniger erfolgreichen Personen unterscheiden. Die sich hieraus ergebenden Verhaltensweisen lassen sich über Expertenurteile oder statistische Verfahren zu Klassen von Arbeitsverhaltensweisen gruppieren, aus denen in einem letzten Schritt über Expertenurteil erfolgskritische Personenmerkmale als höchste Abstraktionsstufe abgeleitet werden.

Die Teilnehmer erhielten zunächst die Instruktion „Nennen Sie typische und erfolgskritische Situationen des Psychologiestudiums“, um zunächst solche fachspezifischen Studiensituationen zu identifizieren, welche als besonders indikativ für den Studienerfolg angesehen werden können. Hierauf notierten die Teilnehmer in Kleingruppen zu je zwei Personen typische und erfolgskritische Anforderungssituationen des Psychologiestudiums. Die dadurch gewonnenen erfolgskritischen Einzelsituationen wurden im Anschluss in Zusammenarbeit aller Teilnehmer nach Oberbegriffen an einer sogenannten Metawand geordnet. In einem weiteren Schritt, welcher der eigentlichen hypothesengeleiteten Prädiktorgenerierung diente, wurden diese typischen und erfolgskritischen Situationen als Ausgangsbasis zur Ableitung derjenigen kognitiven wie nicht-kognitiven Personenmerkmale genutzt, die zur effizienten Erfüllung der aus den Anforderungssituationen erwachsenen Aufgaben des Psychologiestudiums hypothetisch gegeben sein müssten. Auch hier diente die Methode der Kartenabfrage mit anschließender Zuordnung der gefundenen Personenmerkmale zu allgemeineren Begriffen der Strukturierung der Ergebnisse. Die mit diesem Vorgehen gewonnenen Ergebnisse sowohl der typischen und erfolgskritischen Situationen und der daraus abgeleiteten Personenmerkmale wurden in Form einer tabellarischen Übersicht (s. Anhang A und Anhang B) zusammengefasst und nach dem Workshop den Teilnehmern zusammen mit einem Fragebogen (s. Anhang C) über E-Mail zugesandt. Dieser Fragebogen diente zum einen dazu, die abgeleiteten Personenmerkmale hinsichtlich ihrer Relevanz auf einer vierstufigen Ratingskala („unwichtig“ bis „sehr wichtig“) bezüglich verschiedener einzelner Studienerfolgsmerkmale sowie auf ihre generelle Bedeutsamkeit hin zu bewerten. Die Auswahl der Merkmale orientierte sich im Wesentlichen am Studienerfolgsmodell von Rindermann und Oubaid (1999) wie im Folgenden aufgeführt:

a) Studienabschlussnote
b) Weniger Studienabbrüche
c) Kürzere Studiendauer
d) Studienzufriedenheit
e) Allgemeine Wichtigkeit für das Psychologiestudium

Weiterhin wurden die Teilnehmer gebeten, diagnostische Verfahren oder Aufgabentypen zum Erfassen des jeweiligen Eignungsmerkmals zu nennen sowie weitere mögliche Validierungskriterien der gewonnenen Prädiktoren. Fünfzehn Fragebögen wurden an den Autor zurückgesandt und ausgewertet.

7.2 Ergebnisse der Anforderungsanalyse

Eine Zusammenfassung nach Oberbegriffen der Kartenabfrageergebnisse zu typischen und erfolgskritischen *Situationen* des Psychologiestudiums durch die Workshop-Teilnehmer ergab folgende, als wesentlich erachtete Anforderungssituationen (für einen ausführlichen Überblick einzelner Kartenabfragen der Teilnehmer s. Anhang A):

1) Selbstständiger Umgang mit wissenschaftlichen Fragestellungen
2) Literaturarbeit
3) Methodenkenntnisse erwerben und anwenden
4) Umgang mit englischer Literatur
5) Transfer und Integration von Informationen
6) Prüfungsmanagement
7) Studienorganisation
8) Präsentation
9) Organisation von Ressourcen
10) Teamarbeit

Die Ableitung erfolgskritischer kognitiver wie nicht-kognitiver Anforderungs*merkmale* aus obigen typischen Anforderungssituationen als dem eigentlichen Ziel, der Prädiktoren-gewinnung und deren Gliederung nach Oberbegriffen, ergab folgendes Ergebnis (für einen ausführlichen Überblick einzelner Kartenabfragen der Teilnehmer s. Anhang B):

1) Intelligenzfaktoren
2) Instrumentelle Intelligenz
3) Argumentationskompetenz
4) Problemsensitivität
5) Leistungsmotivation
6) Divergentes Denken und Kreativität
7) Persistenz
8) Stabile Persönlichkeit
9) Soziale Kompetenz
10) Selbstständigkeit und Kooperation

In einem weiteren Schritt wurden, wie bereits in Kapitel 7.1.1 beschrieben, die Ergebnisse vom Autor zusammengefasst und den Teilnehmern nebst eines Fragebogens zur Einschätzung der Anforderungsmerkmale hinsichtlich der in Kapitel 3.2 genannten Studienerfolgsmerkmale sowie ihrer generellen Bedeutsamkeit für das Psychologiestudium zugesandt. Ziel hierbei war es, zwischen den Prädiktorendimensionen hypothetische Bedeutsamkeitsunterschiede für die Kriteriumsdimensionen zu identifizieren und über die Einschätzung genereller Wichtigkeit für das Psychologiestudium zu einer empirisch überprüfbaren Bedeutsamkeitsbewertung der Prädiktoren zu gelangen. Die Ergebnisse hierzu zeigen Abbildung 2 bis Abbildung 7.

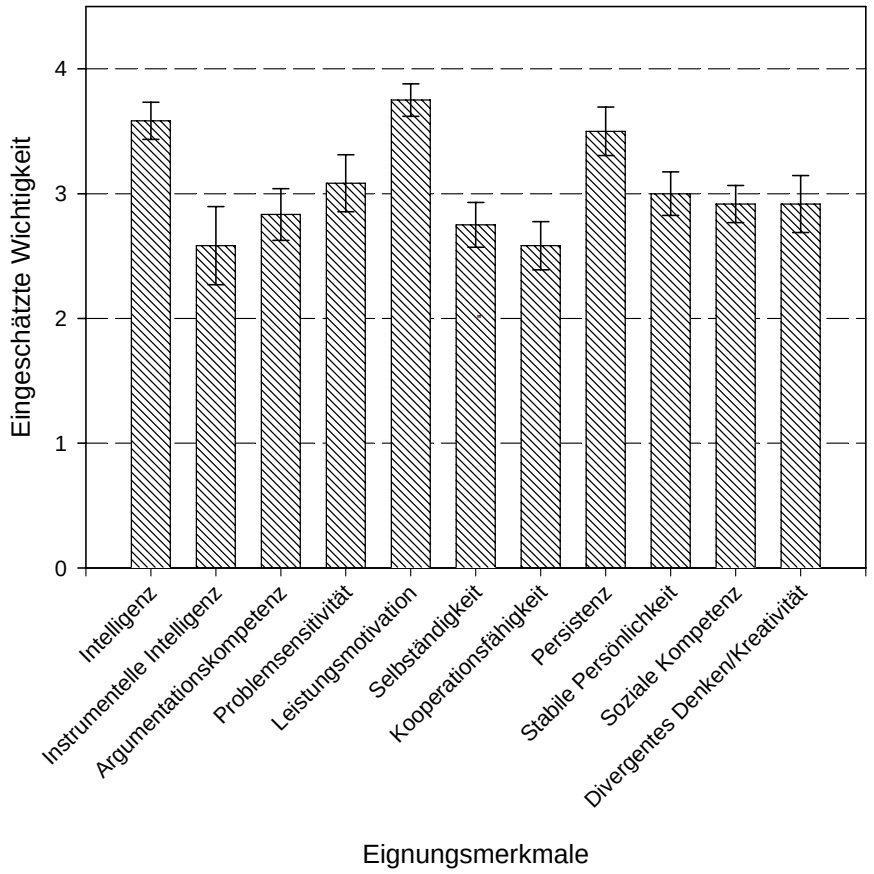


Abbildung 2: Mittelwerte und Standardfehler des eingeschätzten Einflusses der Eignungsmerkmale auf die Abschlussnote. $N = 15$

Bezüglich des Kriteriums Abschlussnote wurden Leistungsmotivation, Intelligenz und Persistenz als maßgeblichste und zugleich gleichbedeutende Einflussgrößen eingeschätzt, legt man den Standardfehler des Mittelwertes als grobes Maß eines bedeutsamen Unterschiedes zugrunde. Die übrigen Prädiktoren zeichnen (unter Einbezug des Standardfehlers) ein nahezu gleichwertiges Bild eingeschätzterWichtigkeit. Erwartungsgemäß wird den leistungsbezo-

genen kognitiven wie nicht-kognitiven Merkmalen hierbei der größte Einfluss zugesprochen. Ohnehin wird allen abgeleiteten Prädiktoren Wichtigkeit bezüglich der Abschlussnote zugemessen (mittlere Einschätzungen durchweg größer als zwei), was für die Wesentlichkeit der abgeleiteten Dimensionen spricht.

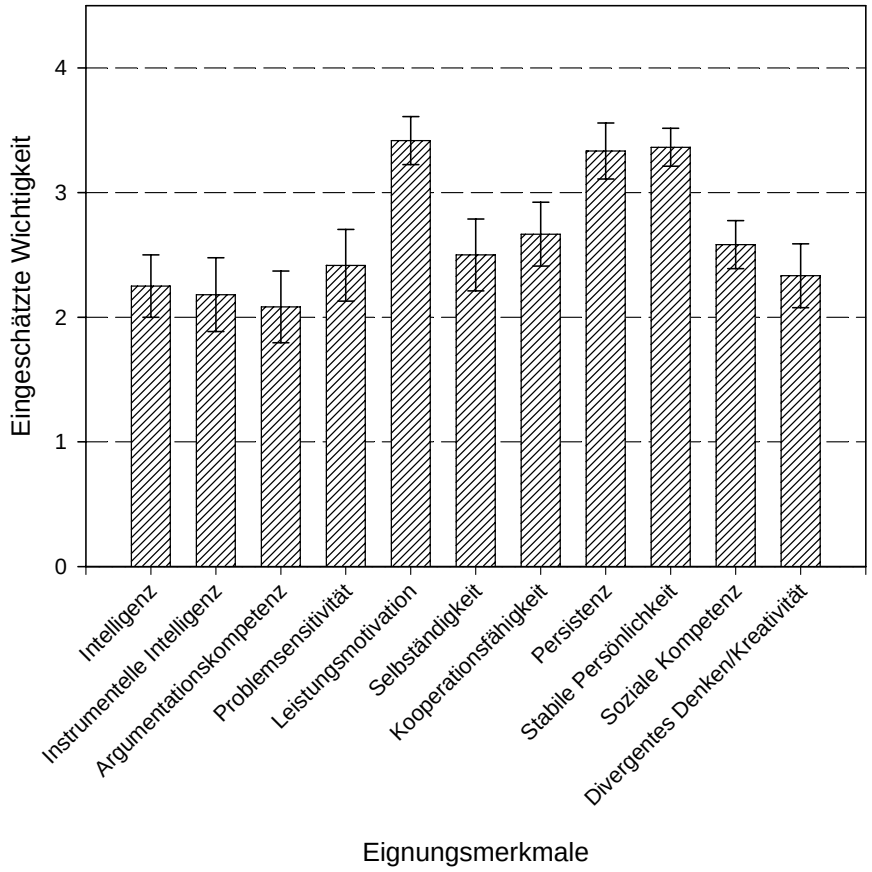


Abbildung 3: Mittelwerte und Standardfehler des eingeschätzten Einflusses der Eignungsmerkmale auf die Abbruchquote. $N = 15$

Für das Kriterium geringerer Abbruchquote ergibt sich eine deutliche und gleichrangige Bedeutsamkeitsgewichtung zugunsten der Persönlichkeitsvariablen Leistungsmotivation, Persistenz und stabile Persönlichkeit. Die übrigen Merkmale zeichnen im Rahmen ihrer Standardfehler ein undifferenziertes Bild. Die Prägnanz der Persönlichkeitsvariablen im Zusammenhang mit geringerem Studienabbruch kann wohl am ehesten vor dem Hintergrund weniger spezifischer Studienanforderungen, als vielmehr allgemeiner „Studienwidrigkeiten“ (etwa Neuartigkeit des Lernstoffes, der Lernbedingungen, Lebensumgebung u.ä.) gesehen

werden. Man kann vermuten, dass bei diesen weitaus weniger kognitive Merkmale, wie schlussfolgerndem Denken, unerlässlich sind, sondern vielmehr Selbstregulations-mechanismen im Zusammenhang einer stabilen Persönlichkeit. Vermutlich wird diese in Verbindung mit einer Persistenz bei der Zielverfolgung und einer allgemeinen Leistungs-bereitschaft als eine notwendige (wenn auch nicht hinreichende) Voraussetzung der Studienbewältigung angesehen.

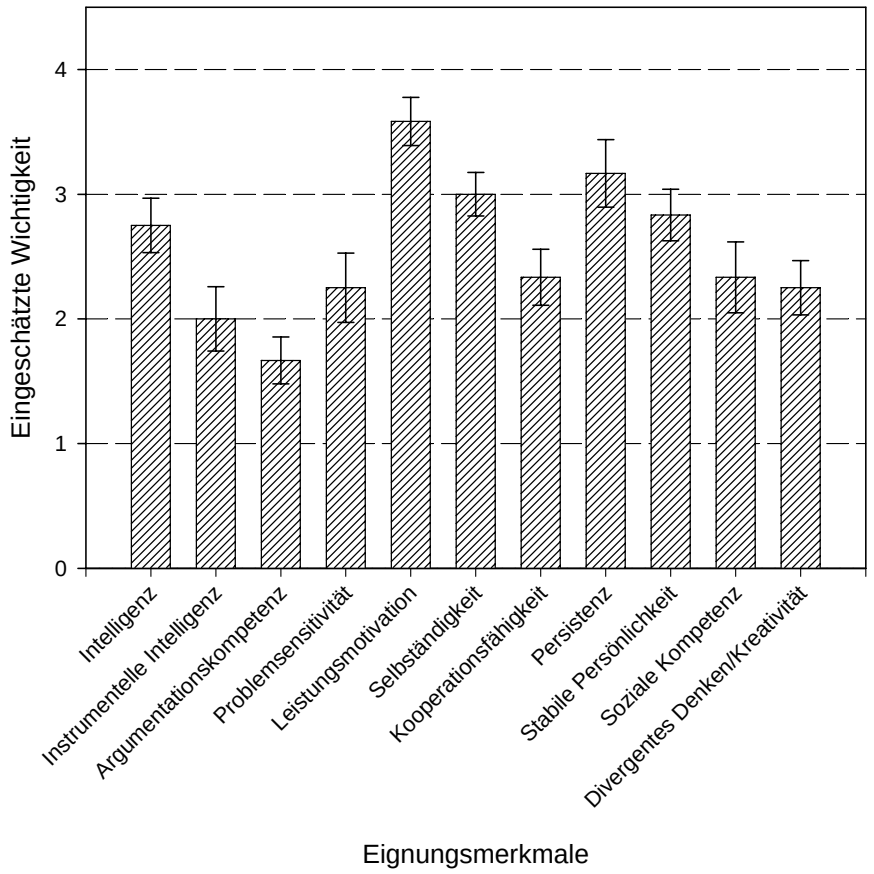


Abbildung 4: Mittelwerte und Standardfehler des eingeschätzten Einflusses der Eignungsmerkmale auf kürzere Studiendauer. $N = 15$

Mit deutlicher Akzentuierung der Leistungsmotivation fällt die Einschätzung des Einflusses auf eine kürzere Studiendauer aus, gefolgt von Persistenz, Selbstständigkeit, stabiler Persönlichkeit und Intelligenz, welche als nahezu gleich wichtig betrachtet werden. Deutlich geringer liegen die Bewertungen für instrumentelle Intelligenz und besonders Argumentationskompetenz, die beide wohl eher bei der Bewältigung spezifischerer Anforderungen als bei einer kürzeren Studiendauer zum tragen kommen.

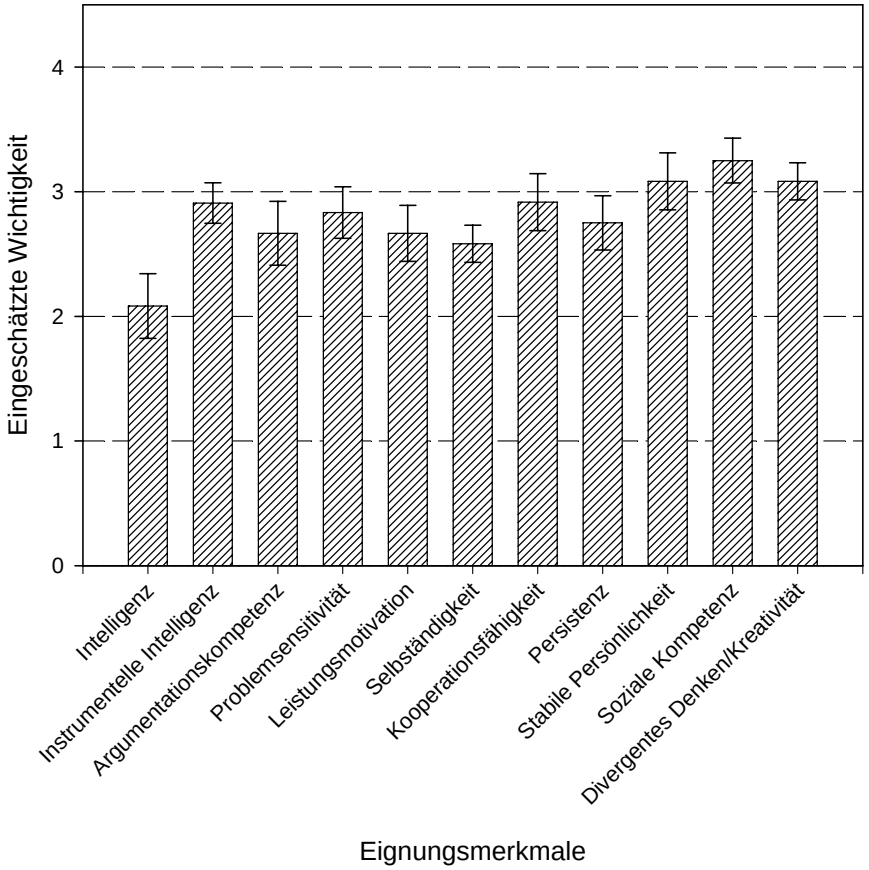


Abbildung 5: Mittelwerte und Standardfehler des eingeschätzten Einflusses der Eignungsmerkmale auf die Studienzufriedenheit. $N = 15$

Ein undifferenziertes Bild ergibt sich für das Kriterium Studienzufriedenheit. Nahezu alle Prädiktoren bis auf Intelligenz werden als wichtig eingeschätzt. Tendenziell erhalten hier die Dimensionen soziale Kompetenz, stabile Persönlichkeit und divergentes Denken/Kreativität die höchsten Beurteilungen. Möglicherweise versteckt sich hinter diesem Ergebnis die Annahme, dass Studienzufriedenheit deutlich mehrfaktorieller bedingt ist als die eher „harten" Kriterien, wie z. B. Studienabschlussnoten. Nicht zuletzt, da Studienzufriedenheit auch eine Frage der Lehrqualität und Betreuungssituation im Studium ist.

Einen Gesamtüberblick über die eingeschätzte Wichtigkeit der Eignungsmerkmale zur Identifizierung derjenigen Merkmale, welche am besten *zwischen* den Kriterien differenzieren, gibt Abbildung 6:

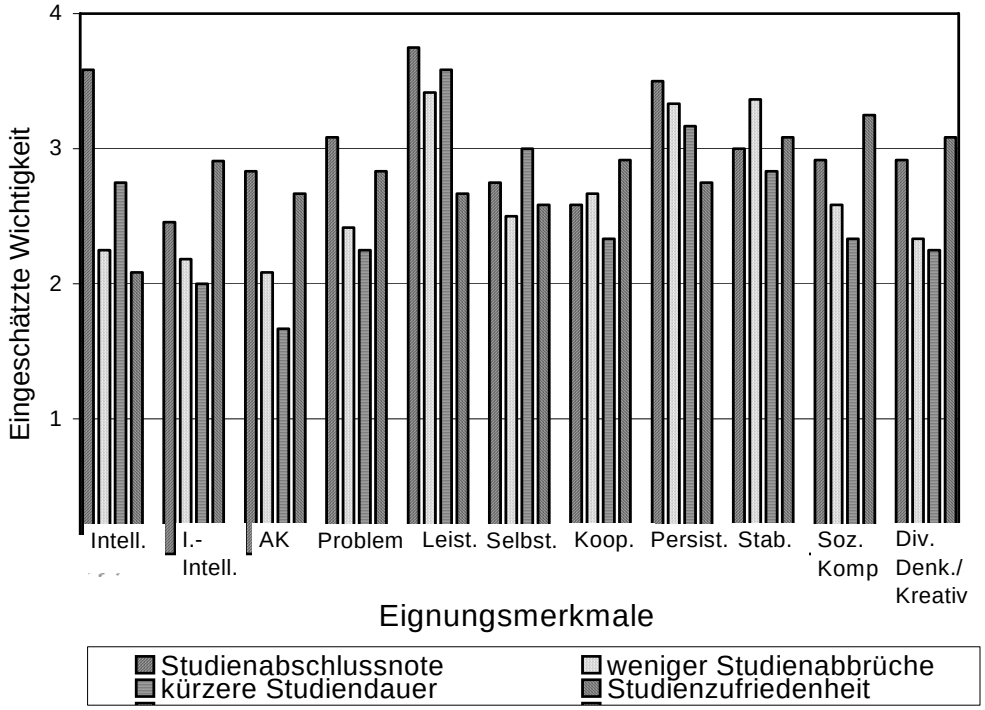


Abbildung 6: Mittelwerte der eingeschätzten Wichtigkeit aller Eignungsmerkmale auf Studienerfolgsmaße.
Anmerkung. Intell.: Intelligenz; I.-Intell.: Instrumentelle Intelligenz; AK: Argumentationskompetenz; Problem.: Problemsensitivität; Leist.: Leistungsmotivation; Selbst.: Selbstständigkeit; Koop.: Kooperationsfähigkeit; Persist.: Persistenz; Stab.: Stabile Persönlichkeit; Soz. Komp.: Soziale Kompetenz; Div. Denk./Kreativ.: Divergentes Denken/Kreativität. $N = 15$

Am differenziertesten fällt die eingeschätzte Wichtigkeit auf Ebene der Fähigkeitsdimensionen aus. Besonders ausgeprägte Profile ergeben sich für Argumentationskompetenz, gefolgt von Intelligenz, Instrumenteller Intelligenz, Problemsensitivität und divergentem Denken/Kreativität. Deutlich undifferenzierter zeigen sich hingegen die Wichtigkeits-einschätzungen der stärker die Persönlichkeit betonenden Eignungsmerkmale. Hier weist lediglich soziale Kompetenz ein differenzierteres Profil auf, die übrigen Persönlichkeits-merkmale (Leistungsmotivation, Selbstständigkeit, Kooperationsfähigkeit, Persistenz und stabile Persönlichkeit) zeichnen ein vergleichsweise weniger prägnantes Bild. Womöglich werden gerade diese stärker auf die Persönlichkeit fokussierenden Dimensionen deutlicher als allgemeine denn spezifische Studieneignungsmerkmale wahrgenommen und erhalten daher in allen Kriterien höhere Wichtigkeitseinschätzungen.

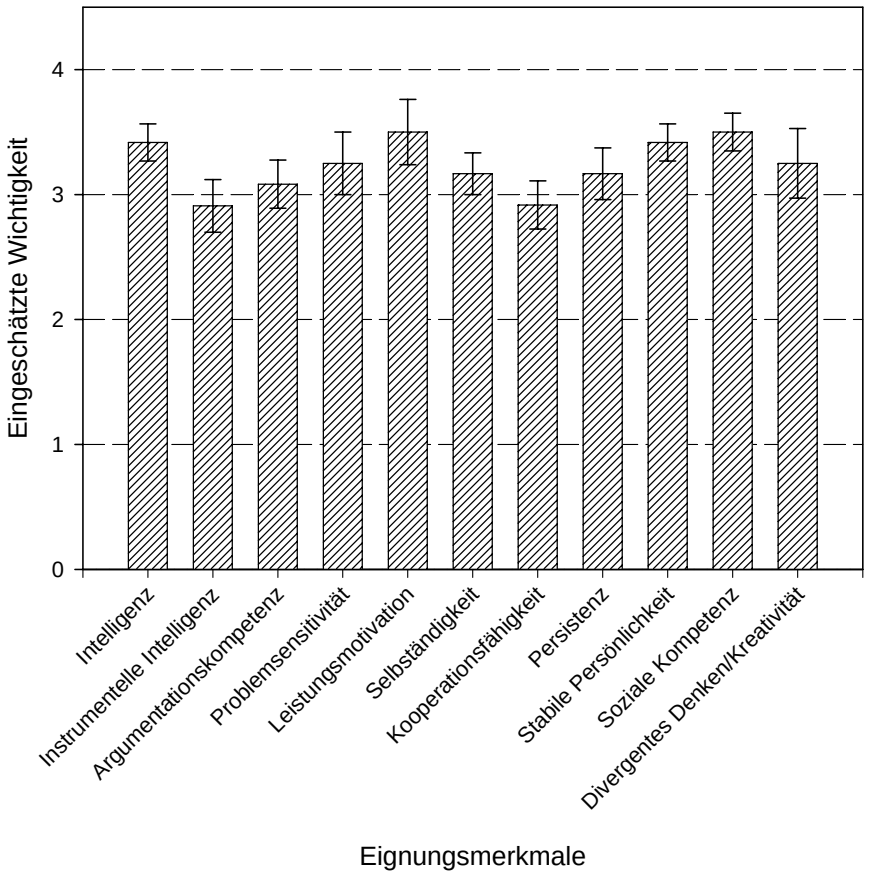


Abbildung 7: Mittelwerte und Standardfehler der eingeschätzten generellen Bedeutsamkeit für das Psychologiestudium. $N = 15$

In der abschließenden Einschätzung der generellen Bedeutsamkeit erhielten wiederum alle abgeleiteten Prädiktoren hohe Bewertungen, was nicht verwundert, war es doch das erklärte Ziel der Anforderungsanalyse, hypothetisch besonders gewichtige Dimensionen des Studienerfolgs zu identifizieren. Zudem spricht (wie schon in den Ausführungen zu Abbildung 2 dargelegt) dieses Ergebnis aus der Fragebogenabfrage für die Wesentlichkeit der abgeleiteten. Zur abschließenden deskriptiven Beurteilung der generellen Wichtigkeit der Dimensionen für das Psychologiestudium wurden die Mittelwerte aus Abbildung 7 rangtransformiert. Die Ergebnisse zeigt Tabelle 6.

Tabelle 6: Rangplätze eingeschätzter genereller Bedeutsamkeit für das Psychologiestudium ($N = 15$)

Eignungsmerkmal	**Rang**
Leistungsmotivation	1
Soziale Kompetenz	1
Intelligenz	2
Stabile Persönlichkeit	2
Divergentes Denken	3
Problemsensitivität	3
Selbstständigkeit	4
Persistenz	4
Argumentationskompetenz	5
Kooperationsfähigkeit	6
Instrumentelle Intelligenz	7

Anmerkung. Rangbindungen wurden nicht aufgelöst

7.3 Einordnung und Diskussion der Ergebnisse

Man mag an der Einzigartigkeit der abgeleiteten Personenmerkmale für das Psychologiestudium berechtigte Zweifel haben, deren geringe Trennschärfe bemängeln und fragen, ob nicht genau diese Eigenschaften bspw. ebenso gut für einen Studienbewerber der Humanmedizin, der Rechtswissenschaften und anderer Studienfächer gelten könnten. Anders ausgedrückt muss eingewendet werden, dass mit der Anforderungsanalyse keine vollständige Abbildung der studienfachspezifischen Anforderungen gewährleistet ist und somit zwar notwendige, aber keineswegs hinreichende Prädiktoren des Studienerfolgs in Psychologie identifiziert wurden. Gründe dafür, dass überwiegend sehr universelle Studieneignungsmerkmale extrahiert wurden, mögen zum einen in dem durch einen einzigen Workshop zur Anforderungsanalyse sehr begrenzten zeitlichen Rahmen liegen, wodurch wahrscheinlich lediglich besonders saliente und daher allgemeine Merkmale genannt wurden. Zum anderen erfolgte die Kategorienbildung der erfolgskritischen Situationen und Personenmerkmale nach einem rein erfahrungsbasiert-intuitivem Vorgehen der am Workshop beteiligten Personen und nicht über eine objektivere quantitativ-analytische Gruppierung der Merkmale in Oberkategorien bspw. über clusteranalytische Verfahren. Zugleich muss allerdings die Kritik am

erfahrungsbasiert-intuitivem Vorgehen eingeschränkt werden, da es das erklärte Ziel der Anforderungsanalyse war, die relevanten Eignungsmerkmale *a priori hypothetisch* zu ermitteln und nicht statistisch-induktiv.

Bei aller Kritik an der Ermittlung von überwiegend allgemeinen Studieneignungsmerkmalen darf nicht außer acht gelassen werden, dass sich die Universalität der Eignungsmerkmale zumindest zum Teil auch als „kleinster gemeinsamer Nenner" eines inhaltlich breiten und heterogenen Studienfaches wie Psychologie ergeben haben mag. Denkt man etwa nur an die Unterschiedlichkeit der Anforderungen z. B. psychosozialer Kompetenz in der klinischen Psychologie gegenüber überwiegend solchen der Formalisierungsfähigkeit der methodisch--theoretisch orientierten Psychologie, so ist diese Vermutung als weiterer Erklärungsansatz plausibel (vgl. hierzu auch Rahn et al. 1976, S. 167).
Darüber hinaus ließen sich innerhalb dieser zentralen Studieneignungsmerkmale fachspezifische Gewichtungen vornehmen, um ein typisches Anforderungsprofil des jeweiligen Studienfaches zu charakterisieren.

Insgesamt wäre es nötig, die hier dargestellten Ergebnisse mit solchen zu vergleichen, die ebenfalls den Anspruch hatten, wesentliche Prädiktoren für das Psychologiestudium zu gewinnen. Dies würde einen qualitativen Vergleich im Sinne konvergenter Validität ermöglichen, um Personenmerkmale zu identifizieren, welchen invariant eine ausgeprägte Bedeutsamkeit für den Erfolg im Psychologiestudium zugemessen wird. Bislang besteht allerdings in der Literatur insgesamt noch ein Mangel speziell an Anforderungsanalysen für Studienfächer, wobei Ergebnisse älterer und neuerer Studien zur Ableitung von studienfachspezifischen Anforderungsmerkmalen existieren. Die früheste Studie stammt hierbei von Hitpass (1975), die zugleich Teil-Grundlage für ein eine Reihe von Symposien zu Anforderungsmerkmalen in verschiedenen Studienfeldern waren, die Rahn et al. (1976) durchführten. Aus neuerer Zeit stammt eine Anforderungsanalyse für das Fach Psychologie von Wetzenstein et al. (2004) an der Humboldt Universität zu Berlin. Allerdings unterscheiden sich die genannten Ansätze sowohl hinsichtlich ihrer genauen Zielsetzung als auch in der Methodik stark voneinander. Im Falle der Studie von Hitpass (1975) und Rahn et al. (1976) sollte das Gesamt *wünschenswerter* Personenmerkmale über Expertenurteile eruiert werden. Bei Wetzenstein et al. (2004) wurde zwar, wie im vorliegenden Fall, in Anlehnung an die Critical Incident Technique (Flanagan, 1954) vorgegangen, jedoch nicht in Form eines Workshops mit überwiegend Dozenten, sondern als Fragebogenstudie an weitaus mehr Studierenden als Dozenten. Hinzu tritt, dass die erfolgskritischen Personenmerkmale wiederum von den

spezifischen Anforderungen einer „Massenuniversität“ wie der Humboldt Universität moderiert werden (Wetzenstein, 2004, S. 29) und einen Vergleich mit den Ergebnissen aus Heidelberg sehr erschweren. Selbst im Falle von übereinstimmenden Personenmerkmalen könnte zudem kritisch gefragt werden, ob die jeweiligen Definitionen allgemein so verständlich formuliert waren, dass alle Befragten dieselbe Interpretationsbasis nutzten. Ein Vergleich der Ergebnisse aus der Heidelberger Anforderungsanalyse mit denjenigen aus Studien mit zumindest ähnlichem Hintergrund wird daher aus den genannten Problemen nicht vorgenommen.

8. Entwicklung von Test-Skalen unter Bezugnahme auf die Ergebnisse der Anforderungsanalyse

Maßgeblich für die Itemgenerierung war die möglichst reliable, objektive und inhalts- und konstruktvalide sowie ökonomische psychometrische Umsetzung der in der Anforderungsanalyse (s. Kap. 7.2) abgeleiteten Personenmerkmale in Testverfahren. Diese Zielsetzung erforderte allerdings auch eine spezifische Auswahl aus den anforderungsanalytisch extrahierten Dimensionen, da diese Dimensionen nicht alle in standardisierte Verfahren hätten überführt werden können, sollte den oben genannten Testgütekriterien entsprochen werden. Daher wurden in einem ersten Schritt die Dimensionen soziale Kompetenz, Selbstständigkeit und Kooperation sowie instrumentelle Intelligenz ausgeschlossen. Für die Dimensionen divergentes Denken und Kreativität erfolgte wegen der großen Konstruktbreite eine Beschränkung auf die Kreativitätsfacetten Ideenflüssigkeit und Ideenflexibilität. Aus demselben Grund wurde für das Konstrukt Problemsensitivität auf die Facette „Offenheit für neue Erfahrungen“ fokussiert.

8.1 Überblick und Erläuterungen zu den eingesetzten Testverfahren

Zunächst werden an dieser Stelle die in dieser Arbeit eingesetzten Konstruktionsprinzipen der Intelligenztestteile detailliert dargestellt, um anschließend einen Überblick über die eingesetzten Standardverfahren zu geben.

8.1.1 Intelligenzdimensionen

Die Auswahl zu messender Intelligenzdimensionen erfolgte nach rationalen Gesichtspunkten. Weil das Psychologiestudium einerseits große Anteile verbaler Anforderungen enthält, andererseits aber den Studierenden im Rahmen der psychologischen Methodenlehre numerische Fähigkeiten abverlangt, sollten die Intelligenzdimensionen verbale und numerische Intelligenz durch jeweils zwei Subtests reliabel erfasst werden. Darüber hinaus wurde fluide Intelligenz über einen Matrizentest operationalisiert, um hierüber eine generelle kognitive Leistungsfähigkeit zu erfassen. Die Konstruktionsprinzipien der Subtests werden im Folgenden dargestellt.

8.1.1.1 Konstruktionsprinzipien Subtest „verbale Analogien“

Sämtliche Items des Subtests für verbale Analogien entstanden aus Übersetzungen von 20 verbalen Analogieaufgaben aus dem Vorbereitungsbuch für die General Record Examination (GRE) „How to prepare for the GRE test“ (Weiner & Wolf, 2003, S. 59f.). Die Auswahl geeigneter Analogien wurde anhand einer möglichst repräsentativen Stichprobe der in der GRE verwendeten Analogie-Regeln (Weiner & Wolf, 2003, S. 57f.) vorgenommen, wie in Tabelle 7 aufgeführt.

Tabelle 7: Überblick über die bei Weiner & Wolf (2003) verwendeten Analogieitem-Konstruktionsregeln und die für das Originalverfahren verwendeten (Fortsetzung der Tabelle auf folgender Seite)

Verbale Analogie-Regel	**Beispiel**	**Anwendung der Regeln in Testitems im Originalverfahren (+/-)**
Definitorisch	Taxonomist : Klassifizieren	-
definierende Charakteristika	Bienenstock : Biene	-
Klasse und Klassenelement	Amphibien : Salamander	+
Antonyme	Heiß : kalt	+
Synonyme	grandios : großartig	+
Intensitätsgrad	eilig : hastig	+
Teil zum Ganzen	Insel : Inselgruppe	+
Funktion	Ballast : Stabilität	+
Art und Weise	Nuscheln : sprechen	-

Verbale Analogie-Regel	**Beispiel**	**Anwendung der Regeln in Testitems im Originalverfahren (+/-)**
Aktion und Bedeutung	zucken : erschrecken	+
Erschaffer und Werkzeug	Maler : Pinsel	+
Erschaffer und Handlung	Akrobat : Salto	+
Arbeiter und Arbeitsplatz	Musikstudent : Musikhochschule	-
Werkzeuge und deren Aktion	Bohrer : drehen	-
Ursache und Wirkung	Schlafmittel : Schläfrigkeit	+
Geschlecht	Hengst : Stute	-
Alter	Fohlen : Hengst	-
Zeitliche Folge	Krönung : Herrschaft	+
Räumlich Folge	Dach : Fundament	+
Symbol und seine Bedeutung	Taube : Friede	+

Anmerkung. +: Regel in Items des Subtests „verbale Analogien“ enthalten;
-: Regel in Items des Subtests „verbale Analogien“ nicht enthalten

Als Beispiel für die Umsetzung dieses Aufgabentyps seien an dieser Stelle drei Originalaufgaben des Testverfahrens aufgeführt:

1) Krönung : Herrschaft = ?

a) Krone : Zepter
b) reich : arm
c) Kindergarten : Schule
d) Diamant : hart

Lösung: c); Regel: zeitliche Abfolge

2) Asyl : Schutz = ?

a) Fischer : Netz
b) Ballast : Stabilität
c) Tiger : Fleischfresser
d) Tenor : Arie

Lösung: b); Regel: Funktion

3) lässig : Vorbedacht = ?

a) aufrichtig : Integrität
b) umgekehrt : Richtung
c) fundamental : Grundlage
d) ehrlich : List

Lösung: d); Regel: Antonym-Variante

Die Übersetzungen wurden zunächst vom Autor vorgenommen, im Weiteren von einem Englischlehrer einer gymnasialen Oberstufe überprüft und, wenn nötig, korrigiert oder verbessert. In der Testdurchführung wurde den Probanden der Aufgabentypus anhand zweier Beispiele in einem Einleitungstext erläutert (s. Anhang F). Die Testzeitbegrenzung wurde durch Analyse der in Angriff genommenen Items in einem Vortest mit fünf Studierenden des Psychologiestudiums von ursprünglich ad hoc geschätzten sieben auf fünf Minuten reduziert.

8.1.1.2 Konstruktionsprinzipien Subtest „Odd-One-Out-verbal“

Allgemein besteht ein Odd-One-Out-Item aus einer Anzahl von Elementen, bei denen ein Element gemäß einer Regel nicht zu den anderen passt. Die Aufgabe der Probanden besteht darin, dieses Element korrekt zu identifizieren. Im vorliegenden Fall wurden 20 Items mit je fünf Elementen erstellt. Dabei wurde versucht, die Items so zu konstruieren, dass sie in ihrem Ambiguitätsgehalt variierten, da es plausibel erschien, dass dieser die Itemschwierigkeit maßgeblich mitbestimmen würde. So ist z. B. Element 3) in Item 1) der folgenden Beispielaufgaben aus dem Originalverfahren oberflächlich betrachtet ein nicht passendes Element, da eine Tür kein Einrichtungsgegenstand ist. Allerdings stellt ein Kissen ebenso kein *typisches* Element der Begriffsklasse „Einrichtungsgegenstände“ dar und es ist eine weitere Suchoperation nach dem unterscheidenden Merkmal nötig, in diesem Falle also „Materialeigenschaft“.
Beispielhaft für seien die drei folgenden Originalaufgaben des Subtests Odd-One-Out-verbal dargestellt:

1) **Welcher Begriff passt nicht zu den anderen?**

1) Schrank
2) Kissen
3) Tür
4) Stuhl
5) Tisch

Lösung: 2); Regel: kein hartes Material

2) **Welcher Begriff passt nicht zu den anderen?**

1) Milch
2) Fleisch
3) Käse
4) Butter
5) Brot

Lösung: 1); Regel: flüssig

3) **Welcher Begriff passt nicht zu den anderen?**

1) Reihe
2) Wirkung
3) Konsequenz
4) Folge
5) Resultat

Lösung: 1); Regel: Keine streng logische Relation

In der Testdurchführung wurde den Probanden der Aufgabentypus in einem Einleitungstext anhand zweier Beispiele erläutert (s. Anhang G). Die Testzeitbegrenzung wurde durch Analyse der in Angriff genommenen Items in einem Vortest mit fünf Studierenden des Psychologiestudiums von ursprünglich ad hoc geschätzten fünf auf drei Minuten reduziert.

8.1.1.3 Konstruktionsprinzipien Subtest „Zahlenreihen“

Prinzipiell orientierte sich die Aufgabenkonstruktion am Aufgabentypus, den bereits Thurstone im „Chicago Test of Primary Mental Abilities“ (zit. nach Amelang & Bartussek, 1997, S. 208) verwendet hatte und wie er in zahlreichen gängigen Intelligenztests enthalten ist (z. B. I-S-T-2000 R; Amthauer, Brocke, Liepmann & Beauducel, 2001). Hierbei müssen nach einer bestimmten Regel aufgestellte vorgegebene Zahlenreihen fortgesetzt werden. Für eine reliable Erfassung der Personenfähigkeit wurden 20 Items regelgeleitet konstruiert, um eine zunächst hypothetische aufsteigende Aufgabenschwierigkeit zu erzielen. Als schwierigkeits-bestimmende Itemkomponenten wurden systematisch der Operatorentypus (Addition, Subtraktion, Multiplikation, Division, Quadrierung) variiert, die Kombination dieser Operatoren, die Operatorenschwierigkeit innerhalb einer Kombination und schließlich die Kombinationsmenge der Operatoren. Diese Itemkomponenten stellen also erforderliche kognitive Operationen zum Lösen einer Aufgabe dar. Damit die Probanden sich zunächst an den Aufgabentypus gewöhnen konnten, wurden in einem Einleitungstext zwei erläuterte Beispiele dargeboten (s. Anhang G). Im anschließenden Subtest waren die ersten vier Items bewusst relativ leicht gehalten („Warming-up-Items“), was sich in einer Konstruktion der Regel mit leichteren Operatoren (Addition, Subtraktion und Multiplikation) und ohne deren Kombination niederschlug. Originalaufgabe a) verdeutlicht dies:

a) 2 3 5 6 8 9 ?
(Lösung: 11, Regel: +1, +2, +1, +2).

Im folgenden Konstruktionsschritt wurde eine Kombination von zwei Operatoren eingeführt, wie Originalaufgabe b) zeigt:

b) 3 8 15 44 87 ?
(Lösung: 260; Regel: Abwechselnd *3-1, *2-1).

Im Weiteren wurde zusätzlich die Operatorenschwierigkeit in einer Kombinationsregel variiert, wie man an Originalaufgabe c) sieht:

c) 1 3 7 15 31 ?
(Lösung: 63; Regel: 2^1-1, 2^2-1, 2^3-1, 2^4-1, 2^5-1, 2^6-1)

Eine letzte Schwierigkeitssteigerung sollte mit der Kombination von mehr als zwei Operatoren erzielt werden wie Originalaufgabe d) verdeutlicht:

d) 6 10 5 7 12 ?
(Lösung: 8; Regel: Abwechselnd *2-2, :2+2

Die Testzeitbegrenzung wurde durch Analyse der in Angriff genommenen Items in einem Vortest mit fünf Studierenden des Psychologiestudiums von ursprünglich ad hoc geschätzten zehn auf acht Minuten reduziert.

8.1.1.4 Konstruktionsprinzipien Subtest „Zahlenmatrizen“

Bei den Aufgaben dieses Subtests handelte es sich nach Kenntnis des Autors um einen bislang neuen Itemtypus zur numerischen Intelligenz. Die Aufgabenstellung bestand darin, in einer nach einer bestimmten Regel aufgebauten 3×3-Zahlenmatrix eine fehlende Zahl zu ergänzen. Wie bereits bei den Zahlenreihen-Items beschrieben, wurden auch diese Aufgaben regelgeleitet mit hypothetisch schwierigkeitsbestimmenden Itemkomponenten konstruiert. Bei diesen handelte es sich zunächst ebenso wie beim Subtest Zahlenreihen um die Komponente Operatorentypus (Addition, Subtraktion, Multiplikation, Division , Quadrierung), die

Kombination dieser Operatoren, die Operatorenschwierigkeit innerhalb einer solchen Kombination und die Operatoren-Kombinationsmenge. Als zusätzliche Komponenten traten hinzu die Orientierung der Regel (zeilen- oder spaltenweise) und jwls. deren Richtung (von rechts nach links oder umgekehrt, bzw. von oben nach unten oder umgekehrt), deren schwierigkeitsbestimmender Einfluss für Matrizen-Items zur fluiden Intelligenz bekannt ist (s. z. B. Forman, 1973; Hornke & Habon, 1986). Die Originalaufgaben a) bis d) verdeutlichen die beschriebenen Konstruktionsregeln:

a)

18	**45**	**67**
34	**34**	**26**
52	**?**	**93**

Lösung: 79; Regel: 45 + 34 = 79

b)

?	**33**	**8**
48	**42**	**5**
22	**12**	**9**

Lösung: 42; Regel: (8 + 33) + 1 = 42

c)

13	**6**	**18**
7	**11**	**14**
19	**1**	**?**

Lösung: 22; Regel: 18 × 2 - 14 = 22

d)

12	**30**	**?**
4	**24**	**6**
64	**36**	**25**

Lösung: 11; Regel: (Quadratwurzel aus 25) + 6 = 11

Die Testzeitbegrenzung wurde durch Analyse der in Angriff genommenen Items in einem Vortest mit fünf Studierenden des Psychologiestudiums von ursprünglich ad hoc geschätzten 12 auf 13 Minuten erhöht.

8.1.1.5 Konstruktionsprinzipien Subtest „Matrizen“

Wie bereits bei den Subtests „Zahlenreihen“ und „Zahlenmatrizen“ wurden auch die Matrizen-Items weitestgehend regelgeleitet konstruiert. Allerdings war es aus zeittechnischen Gründen lediglich möglich, 14 der 20 Items in dieser Weise zu konstruieren, die übrigen sechs wurden mit den für diesen Zweck nötigen Modifikationen aus den Raven Advanced Progressive Matrices (Raven, Raven & Court, 1998) entnommen. Neben Item-Layout-Eigenschaften musste hierbei insbesondere die Distraktorenmenge auf fünf reduziert werden, da der

Test zeitbegrenzt durchgeführt werden musste. Zudem wurden die ersten drei Aufgaben als 2×2-Matrizen konstruiert, um als „Warming-up-Items“ zu fungieren, die weiteren im üblichen 3×3-Design. Als Grundlage für die regelgeleitete Itemkonstruktion der 14 selbst konstruierten Items diente eine Arbeit von Hornke und Rettig (1988), deren Grundprinzipien im Folgenden dargestellt werden.

Wesentlich an diesem Ansatz ist die Itemkonstruktion nach kognitionspsychologischen Überlegungen bezüglich der bei der Itemlösung erforderlichen kognitiven Operationen. Die hierdurch abgeleiteten schwierigkeitsbestimmenden Komponenten ermöglichen eine regelgeleitete Itemkonstruktion, indem systematisch einzelne Regeln kombiniert werden. Zur Überprüfung der Hypothese, dass diese Konstruktionsregeln geeignet sind, die empirischen Itemschwierigkeiten vorherzusagen, kann das linear-logistische Testmodell (LLTM) von Fischer (1973) herangezogen werden. Das LLTM stellt einen Spezialfall des dichotomen Rasch-Modells (Rasch, 1960) dar und zerlegt dessen Itemschwierigkeiten in eine Linearkombination sogenannter Basisparameter. Dieses stellen die präexperimentelle Hypothese über die bei der Aufgabe beteiligten kognitiven Komponenten dar. Folglich ergäbe sich bei Modellgeltung eine Übereinstimmung der Itemschwierigkeiten des dichotomen Rasch-Modells mit denen nach dem LLTM, da erstere sich aus der Linearkombination der hypothetischen Basisparameter des LLTM ergäben.
In ihrer Terminologie schwierigkeitsbestimmender Faktoren unterscheiden Hornke und Rettig (1988) zwischen Repräsentanten als den in den Matrizen-Quadranten enthaltenen Einzelfiguren sowie Komponenten als die Kombination aus mehreren schwierigkeitsbestimmenden Merkmalen. Das hierzu abgeleitete Regelsystem beinhaltet die folgenden Bestimmungsmerkmale:

I. Mögliche Beziehungen zwischen den Einzelkomponenten einer Aufgabe
   1) *Identität:* Dreimalige Anordnung eines Repräsentanten.
   2) *Addition:* Bildung eines Repräsentanten einer Komponente durch Übereinanderlegung der übrigen in derselben Zeile oder Spalte stehenden Repräsentanten.
   3) *Subtraktion:* Entstehung eines Repräsentanten einer Komponente durch Ausblendung einer Teilfigur aus einer vollständigen Figur (in der 1. oder 2. Spaltenposition).
   4) *Disjunktion:* Konstruktion eines Repräsentanten einer Komponente dadurch, dass von den übrigen in derselben Zeile oder Spalte stehenden Repräsentanten nur solche Elemente übertragen werden, die in beiden vorkommen.

5) *Einzelelemente-Addition:* Entstehung eines Repräsentanten einer Komponente dadurch, dass von den beiden übrigen in derselben Zeile oder Spalte angeordneten Repräsentanten nur solche Elemente übernommen werden, die jeweils in der einen, jedoch nicht in der anderen Figur vorkommen. Allerdings bestehen in beiden Repräsentanten gemeinsame Elemente.

6) *Seriation:* Entstehung eines Repräsentanten einer Komponente dadurch, dass dieselbe Veränderungsregel zur Umwandlung der ersten in die zweite Figur (bspw. Verschiebung in der Ebene oder im Raum, Addition oder Subtraktion) auf die zweite Figur angewendet wird, um die dritte Figur zu erzeugen.

7) *Variieren:* Variation der Reihenfolge von drei Repräsentanten. Hierbei kann es sich um a) geschlossene Gestalten (bspw. Kreis, Dreiecke, Mehrecke) oder b) offene Gestalten (bspw. Pfeile, Linien) handeln.

II. Eine Regel gilt entweder
1) zeilenweise
2) spaltenweise
3) sowohl zeilenweise als auch spaltenweise

III. Ergänzung des Konstruktionsregelansatzes durch Realisierungsmodi

1) *Separierte Komponenten:* Es existieren konstante Repräsentantenmerkmale (wie z. B. groß vs. klein, Vorder- vs. Hintergrund), die zur Unterscheidbarkeit der Komponenten führt. Dabei führt die Abwesenheit ambivalenter Stimuli in einem Item also zu einer leichten Unterscheidbarkeit der figuralen Elemente.

2) *Integrierte Komponenten:* „Das relevante Repräsentantenmerkmal einer Komponente ist ein nicht-konstantes irrelevantes Repräsentantenmerkmal einer anderen Komponente“ (Hornke & Rettig, 1988, S. 142). D.h., wegen der Integration von zwei Komponenten, von denen nur eine zur Lösung führt, kommt es zu einem Mehrfachbezug. Um diesen aufzulösen, ist eine Suchoperation notwendig, um das relevante Merkmal aus den störenden, auf den falschen Lösungsweg führenden Komponenten, hervorzuheben.

3) *Integrierte Repräsentanten:* lässt sich durch das Zeichenprinzip „Verschmelzung“ beschreiben. Eine dadurch entstandene geschlossene Gestalt muss durch eine Suchoperation nach der richtigen Regel „aufgelöst“, desintegriert werden.

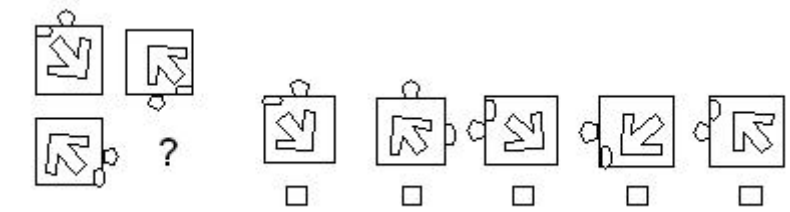
Aus Zeitgründen konnten nicht alle oben erläuterten Konstruktionsregeln mitsamt einer zufälligen Aufgabenauswahl aus allen möglichen Kombinationen (also dem vollkommen spezifizierten Itemuniversum) realisiert werden. Um einen Überblick über die angewendeten Itemkonstruktionsregeln zu geben, werden im Folgenden beispielhaft sieben Originalaufgaben nebst Nennung ihrer Konstruktionsregeln dargestellt:

Beispielaufgaben aus dem Originalverfahren zu den angewendeten Matrizenkonstruktions-regeln (verkleinerte Abbildungen):

a)

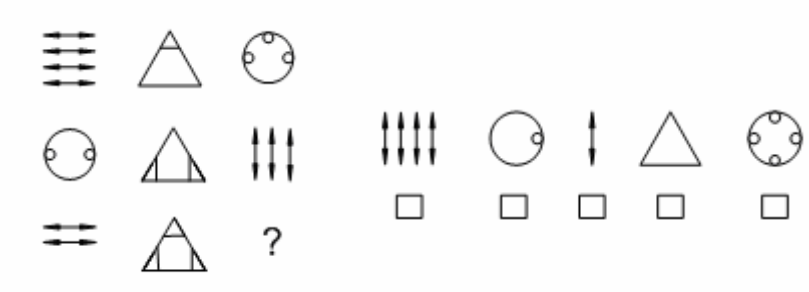
Variation geschlossener Gestalten (Lösung: 3. Alternative)

b)



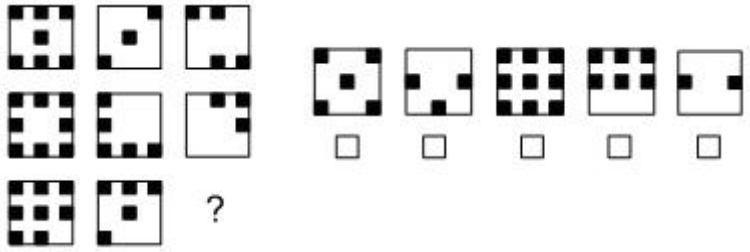
Separierte Komponenten (Lösung: 3. Wahlalternative)

c)



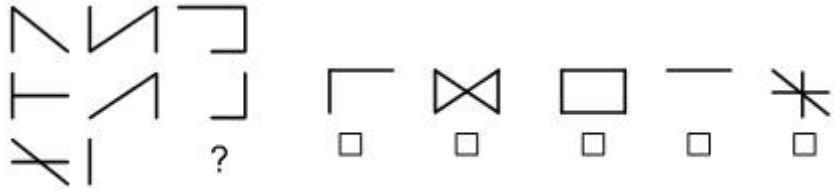
Integrierte Komponenten (Lösung: 2. Wahlalternative)

d)



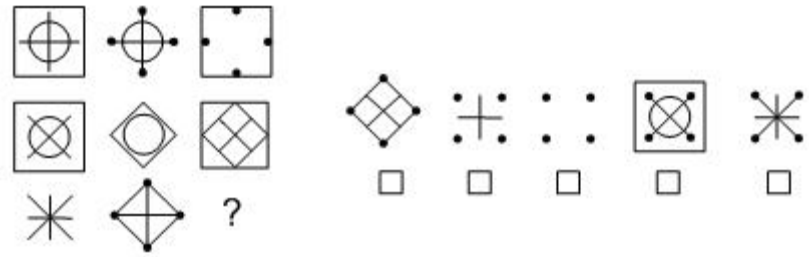
Subtraktion in Zeilen (Lösung: 2. Wahlalternative)

e)



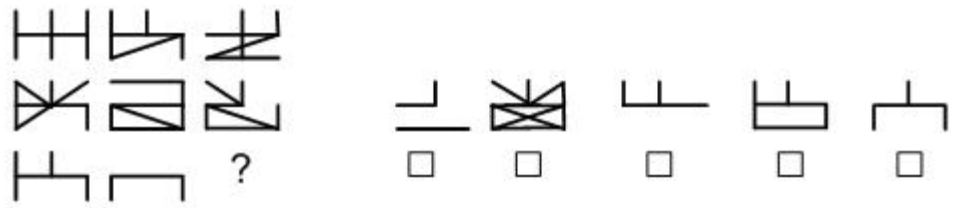
Integrierte Repräsentanten in Spalten (Lösung: 4. Wahlalternative)

f)



Einzelelemente-Addition in Zeilen (Lösung: 1. Wahlalternative)

g)



Disjunktion in Spalten (Lösung: 1. Wahlalternative)

8.1.2 Skala zur Erfassung der passiven argumentativ-rhetorischen Kompetenz (SPARK)

Zur Messung der Fähigkeitsdimension Argumentationskompetenz diente die Skala zur Erfassung der passiven argumentativ-rhetorischen Kompetenz (SPARK) von Flender, Christmann, Groeben und Mlynski (1996). Hierbei handelt es sich um ein Szenario-basiertes Verfahren zur Erfassung der „Fähigkeit zur Identifikation von argumentativen, rhetorischen, argumentationslogischen und interaktiven Auffälligkeiten in einer Argumentation“ (Flender, Christmann & Groeben, 1999, S. 309). Die vier Bereiche der passiven argumentativ--rhetorischen Kompetenz sowie die damit verbundenen Aufgabenstellungen lassen sich wie folgt beschreiben (nach Flender, Christmann & Groeben, 1999, S. 312):

1) Passive *argumentative* Kompetenz: Identifikation von argumentativen Regelverletzungen. Angabe, ob im vorgelegten Szenario (Diskussion zweier Personen) eine argumentative Auffälligkeit vorhanden ist und – falls zutreffend – Nennung der Zeile, in der diese steht. Anschließend ist unter neun Vorgaben im Mutiple-choice-Format die korrekte Auffälligkeit zu benennen (Umkehrschluss, eigene Sicht als gesicherte Tatsache darstellen, Strohmänner aufbauen, Kompetenz absprechen, Handlungsdruck erzeugen) mit Angabe der Urteilssicherheit. Weitere Auffälligkeiten können in freier Antwort benannt werden.
2) Passive *rhetorische* Kompetenz: Identifikation von rhetorischen Stilfiguren. Identifizierung von maximal zwei vorhandenen rhetorischen Stilfiguren (Ironie, Metapher, rhetorische Frage, Litotes, Oxymoron etc.) im Multiple-Choice-Format mit Angabe der Urteilssicherheit. Zwei weitere rhetorische Auffälligkeiten können im freien Antwortformat benannt werden.
3) Passive *argumentationslogische* Kompetenz (Verständnis für die Logik einer Argumentation): Einschätzen weiterer Argumente und Angabe der Urteilssicherheit. Drei weitere Argumente, welche die jeweils vorgetragene Position untermauern, unterminieren oder im betreffenden Diskussionskontext irrelevant sind, werden vorgegeben. Die Untersuchungsperson soll auf dem Hintergrund einer strittigen These eines der Diskussionspartner beurteilen, ob die Argumente jeweils als Pro- oder Contra-Argument oder irrelevantes Argument einzustufen (Multiple-Choice- Format) sind.
4) Passive *interaktive* Kompetenz: Identifikation von konfrontativen Äußerungen und geteilten Aussagen mit Angabe der Urteilssicherheit. Gemeinsam geteilte Aussagen

stellen szenariospezifische Abstraktionen vorgetragener Positionen dar, welche nicht unmittelbar Gegenstand des Gesprächs waren und daher erschlossen werden müssen (Multiple-Choice-Format).

SPARK liegt in zwei inhaltlich parallelen Formen (Versionen A und B) mit jeweils fünf Szenarien vor. Das Rahmenthema von Version A besteht in der Diskussion zur Frage „Macht Fernsehen aggressiv?“, dasjenige von Version B in einer zur Fragestellung „Ist Intelligenz angeboren?“. In jedem der fünf Szenarien befinden sich eine argumentative, zwei rhetorische Auffälligkeiten sowie eine konfrontative Äußerung und müssen von den Probanden jeweils nach den oben genannten Aufgabenstellungen 1) - 4) bearbeitet werden.

Da es im Rahmen des Gesamtverfahrens aus Zumutbarkeitsgründen nicht möglich gewesen wäre, die gesamte SPARK durchzuführen, wurde diese zunächst auf die Szenarien „Tagung für Sozialarbeiter/innen“ (Version A) und „Fachhochschulausbildung“ (Version B) verkürzt. Die Auswahl speziell dieser Szenarien orientierte sich an den bei Flender et al. (1996, S. 39) angegebenen Trennschärfen und internen Konsistenzen nach Cronbachs Alpha. Für das Szenario „Tagung für Sozialarbeiter/innen“ sind diese mit $r_{it} = .23$ und $\alpha = .32$ angegeben, für das Szenario „Fachhochschulausbildung“ liegen diese bei $r_{it} = .28$ und $\alpha = .24$. Insgesamt liegen diese Werte nicht im zufriedenstellendem Bereich, diejenigen anderer Szenarien fallen allerdings noch niedriger aus. Zudem wurde auf die Teilaufgabenstellung zur passiven *rhetorischen* Kompetenz verzichtet, da dieser Aspekt speziell für die Anforderungen des Psychologiestudiums nicht zentral erschien. Ebenso erfolgte keine Erfassung der individuellen Urteilssicherheit, da diese nach Flender et al. (1996, S. 44) in der vollständigen Version lediglich zu $r = .17$ mit dem Testgesamtwert korrelierte und somit keine hohe zusätzlich bedeutsame Informationsquelle für die Ermittlung der Personenfähigkeit darstellte. Die Testzeitbegrenzung wurde durch Analyse der in Angriff genommenen Items in einem Vortest mit fünf Studierenden des Psychologiestudiums auf 17 Minuten festgelegt.

8.1.3 Subtests Kreativitätsfacetten Ideenflüssigkeit und Ideenflexibilität

Die Kreativitätsfacetten Ideenflüssigkeit und Ideenflexibilität wurden über die Kreativitäts-subtests des Berliner Intelligenz Strukturtest (BIS) (Jäger, Süss & Beauducel, 1997) erfasst. Tabelle 8 gibt eine Übersicht über die Aufgabenstellung und den Auswertungsmodus.

Tabelle 8: Überblick über Aufgabenstellungen zu den Kreativitätskomponenten Ideenflüssigkeit und Ideenflexibilität des BIS (Jaeger, Suess & Beauducel, 1997)

Aufgabenstellung	**Auswertungsmodus**
Aufschreiben möglichst vieler deutscher Wörter mit der Vorsilbe „Fern“. *Zeitvorgabe: 1:30 Min.*	Entfällt, da Aufwärmaufgabe
Aufgabenstellung	**Auswertungsmodus**
Ergänzung einer vorgegebenen einfachen Figur so, dass daraus reale Gegenstände entstehen. *Zeitvorgabe: 2:30 Min.*	*Ideenflüssigkei*t: Menge der Lösungen (Anzahl der instruktionsgemäßen Lösungen) *Ideenflexibilität*: Vielfalt der Lösungen (Anzahl der unterschiedlichen Kategorien)
Bildung von möglichst vielen Sätzen mit möglichst verschiedenem Inhalt aus drei vorgegebenen Hauptwörtern (Pluralbildung nicht erlaubt) *Zeitvorgabe: 2:00 Min*	*Ideenflüssigkeit*: Menge der Lösungen (Anzahl der instruktionsgemäßen Lösungen)
Nennung möglichst vieler verschiedener Verwendungsmöglichkeiten für einen vorgegebenen Gegenstand *Zeitvorgabe: 2:00 Min.*	*Ideenflüssigkeit*: Menge der Lösungen (Anzahl der instruktionsgemäßen Lösungen) *Ideenflexibilität*: Vielfalt der Lösungen (Anzahl der unterschiedlichen Kategorien)

Die Beurteilung der Testprotokolle erfolgte gemäß des standardisierten Auswertungsschemas doppelt und unabhängig voneinander durch den Autor und eine studentische Hilfskraft.

8.1.4 Subtest „empiriebezogenes Denken“

Zusätzlich zu den anforderungsanalytisch abgeleiteten Dimensionen wurde ein davon distinktes Leistungsmerkmal in Form einer frei zu beantwortenden Fragestellung zum erfahrungswissenschaftlichen, empiriebezogenen Denken konstruiert. Wenn auch diese Dimension nicht explizit in den anforderungsanalytisch ermittelten Personenmerkmalen enthalten war, so schien seine Aufnahme in die Testbatterie doch wichtig, da sich dieses

Merkmal zum großen Teil in der erfolgskritischen *Situation* „Selbstständiger Umgang mit wissenschaftlichen Fragestellungen“ widerspiegelte (s. Anhang A). Ziel war es hierbei, eine Aufgabe zu schaffen, welche bereits von Studienanfängern gelöst werden konnte, ohne dass fachspezifisches methodisches Vorwissen nötig war. Hierbei bot es sich an, eine Fragestellung aus dem Alltagsbereich zu konstruieren. Die vom Autor entworfene Aufgabenstellung lautete wie folgt:

Ein Getränkemittelhersteller behauptet, durch Konsum eines von ihm neu entwickelten „Energy-Drinks“ lasse sich die Konzentrationsleistung stärker steigern im Vergleich zu anderen „Energy-Drinks“.
a) Wie würden Sie vorgehen, um diese Behauptung **empirisch** zu überprüfen? Schildern Sie **kurz** ihr Vorgehen! b) Welche Vorkehrungen würden Sie hierbei treffen, um die Verallgemeinerbarkeit Ihrer Ergebnisse zu gewährleisten? **Nennen** Sie wesentliche Punkte!

Um zu einer möglichst standardisierten Auswertung durch zwei Beurteiler zu gelangen, wurde vom Autor ein Bewertungsschema zur Beurteilung der Teilaufgaben a) und b) entworfen und zwei Professoren des Psychologischen Instituts Heidelberg zur Korrektur und Erweiterung unabhängig voneinander vorgelegt. Hieraus resultierte ein Bewertungsschema mit den folgenden Kategorien je Teilaufgabe nach Tabelle 9

Tabelle 9: Überblick über Beurteilungsdimensionen des Tests zum „empiriebezogenem Denken“

Teilaufgabe	**Dimensionen der Aufgabenbeurteilung**
a) Generelles empirisches Vorgehen	-Prinzip des Gruppenvergleichs genannt und skizziert -Vorkehrungen zur Störvariablenkontrolle und Fehlervarianzminimierung vorgenommen
b) Verallgemeinerbarkeit der Ergebnisse	-Prinzipien der Verallgemeinerbarkeit auf verschiedene Personen und Variablen genannt -Absicherung der Ergebnisse gegenüber dem Zufall genannt und skizziert -Externe Validität der Ergebnisse gesichert

Die Beurteilung der Dimensionen erfolgte anhand einer fünfstufigen Antwortskala (von 1 „ungenügend“ bis 5 „sehr gut“) durch den Autor und einer angeleiteten studentischen Hilfskraft doppelt und unabhängig voneinander. (Für einen detaillierten Überblick über das Beurteilungsschema s. Anhang D und Anhang E). Vor der Bewertungsprozedur aller Aufgabenbearbeitungen wurde das Kategorienschema in einem Testdurchlauf anhand von zehn zufällig ausgewählten Aufgabenbearbeitungen vom Autor und der angewiesenen studentischen Hilfskraft auf Verständlichkeit der Beurteilungsdimensionen sowie der Beurteilerübereinstimmung hin überprüft. Als deskriptives Maß der Übereinstimmung diente Kendalls Tau-b (s. z. B. Lienert, 1978, S. 45f.). Die Koeffizienten für den Testdurchlauf lagen mit Ausnahme der Dimensionen „Verallgemeinerbarkeit auf verschiedene Personen und Variablen“ und „Externe Validität“ im befriedigenden bis sehr guten Bereich (Prinzip des Gruppenvergleichs: $\tau_b = .83$, Störvariablenkontrolle und Fehlervarianzminimierung: $\tau_b = .88$, Verallgemeinerbarkeit auf verschiedene Personen und Variablen: $\tau_b = .45$, Absicherung gegenüber dem Zufall: $\tau_b = 1$, Externe Validität $\tau_b = .48$). Die geringe Übereinstimmung in den Beurteilungsdimensionen zur Verallgemeinerbarkeit der Ergebnisse und in der externen Validität ging nach den Ergebnissen einer Nachbesprechung auf Unklarheiten in der Definition zurück. Daher wurden diese klarer und prägnanter reformuliert und eine nochmalige Beurteilung dieser Beurteilungsdimensionen anhand anderer zehn zufällig ausgewählter Testprotokolle durchgeführt. Die Kennwerte der Beurteilerübereinstimmung lagen danach für beide Dimensionen im befriedigenden Bereich (Verallgemeinerbarkeit auf verschiedene Personen und Variablen: $\tau_b = .78$, externe Validität: $\tau_b = .80$). Die Auswertungsobjektivität des Bewertungsschemas konnte somit als insgesamt befriedigend angesehen werden und wurde daher in dieser Form (s. Anhang D) zur Beurteilung der Testprotokolle herangezogen.

8.1.5 Persönlichkeitsfragebögen

Zur Erfassung der Persönlichkeitsdimensionen wurden verschiedene standardisierte Fragebögen verwendet, deren inhaltliche Hintergründe im Folgenden überblicksartig erläutert und deren interne Konsistenzen nach Cronbachs Alpha für die jwl. Normierungsstichproben angegeben werden. Folgende Skalen kamen hierbei zum Einsatz:

- Leistungsmotivationsinventar Kurzversion (LMI-K) von Schuler und Prochaska (2001)
- Skala „Hartnäckige Zielverfolgung“ (HZV) und Skala „Flexible Zielanpassung“ (FZA) von Brandtstädter und Renner (1988)
- Skala „Offenheit für Erfahrungen“ aus dem NEO-FFI von Borkenau und Ostendorf (1993)
- Skala „Gewissenhaftigkeit“ aus dem NEO-FFI von Borkenau und Ostendorf (1993)
- Skala „Neurotizismus“ aus dem NEO-FFI von Borkenau und Ostendorf (1993)
- Lügen- und Leugnungsskala von Ling (zit. nach Amelang & Bartussek, 1970)

Die Beschreibung einschlägiger Skalen wie denjenigen aus der deutschen Übersetzung des NEO-FFI (Borkenau & Ostendorf, 1993) und des Leistungsmotivationsinventars in der Kurzversion (Schuler & Prochaska, 2001) soll auf die wesentlichsten beschreibenden Merkmale fokussieren. Bei weniger bekannten Skalen, insbesondere denjenigen zur Hartnäckigen Zielverfolgung und Flexiblen Zielanpassung (Brandstädter & Renner, 1988) sowie den Lügen- und Leugnungsskalen von Ling (zit. nach Amelang & Bartussek, 1970), soll die Beschreibung hingegen eingehender ausfallen.

8.1.5.1 Leistungsmotivation

Für diese Dimension wurde aus Gründen der Testökonomie auf die Kurzversion (LMI-K) des Leistungsmotivationsinventars (LMI) von Schuler und Prochaska (2001) zurückgegriffen. Das LMI-K (Schuler & Prochaska, 2001) beinhaltet 30 der jeweils trennschärfsten Items aus siebzehn Skalen des LMI. Leistungsmotivation wird hierbei als mehrdimensionales Konstrukt verstanden mit den Facetten Beharrlichkeit, Dominanz, Engagement, Erfolgszuversicht, Flexibilität, Flow, Furchtlosigkeit, Internalität, Kompensatorische Anstrengung, Leistungsstolz, Lernbereitschaft, Schwierigkeitspräferenz, Selbstständigkeit, Selbstkontrolle, Statusorientierung, Wettbewerbsorientierung und Zielsetzung. Zur theoretischen Einbettung des Verfahrens schreiben Schuler & Prochaska (2001):

> Das LMI wurde unter Nutzung vorliegender theoretischer und empirischer Arbeiten zur Leistungsmotivation sowie allgemeiner persönlichkeitstheoretischer Ansätze und Messverfahren sowie von Ergebnissen der Leistungsmotivationsforschung im eignungsdiagnostischen Kontext entwickelt. Ziel war die Formulierung eines breiten Konzepts berufsbezogener Leistungsmotivation. Persönlichkeitstheoretischer Hintergrund ist das Verständnis von Leistungsmotivation als Ausrichtung weiter Anteile der Persönlichkeit auf die Leistungsthematik. (S. 5)

Die interne Konsistenz des LMI-K liegt bei $\alpha = .94$.

8 1.5.2 Hartnäckige Zielverfolgung und Flexible Zielanpassung

Zur Erfassung der Dimension Persistenz wurde auf die Skala „Hartnäckige Zielverfolgung“ (HZV), für „Flexible Zielanpassung“ (FZA) die gleichnamige Skala aus dem „Fragebogen zur hartnäckigen Zielverfolgung und flexiblen Zielanpassung“ von Brandstädter und Renner (1988) zurückgegriffen. Weil im Zwei-Prozess-Modell der Zielverfolgung und Zielanpassung von Brandtstädter (Bak & Brandtstädter, 1998; Brandstädter, 1984, 2002; Brandstädter & Rothermund, 2002; Rothermund & Brandstädter, 2003) beide Dimensionen untrennbar verbunden sind, wurde die Skala FZA schon aus rein theoretischen Gründen miteinbezogen, was jedoch im Folgenden weiter begründet werden soll.

Brandstädter umschreibt in seinem Zwei-Prozess-Modell der Entwicklungsregulation die zwei Persönlichkeitsvariablen Zielverfolgung und Zielanpassung als getrennte Dimensionen eines Stabilitäts-Flexibilitätsdilemmas (Brandstädter & Rothermund, 2002):

> Action regulation in the pursuit of goals and plans faces a basic dilemma: It must be sufficiently stable and closed to stay focused on the goal and resist distractive influences; at the same time, it has to be open and flexible enough so that plans and priorities can be adjusted to new and unexpected circumstances (S. 120).

Weiterhin werden vor diesem Hintergrund assimilative und akkomodative Aktivitäten unterschieden. Assimilative Aktivitäten stellen Bestrebungen des Individuums dar, aktuelle Anforderungen in eigene Ziele zu integrieren, etwa über eigene Anstrengungen: „To achieve some identity goal, the person may try to acquire relevant knowledge and skills“ (Brandstädter & Rothermund, 2002, S. 121). Hiermit fest verknüpft ist die Dimension der hartnäckigen Zielverfolgung. Demgegenüber sind akkomodative Aktivitäten durch Anpassungsprozesse des Individuums an nicht erreichbare Ziele gekennzeichnet: „Typical outcomes of accomodative processes include the rescaling of aspirations, the dissolving of barren attachments, and the channeling of assimilative energies toward new, feasible goals“ (Brandstädter & Rothermund, 2002, S. 123). Klar erkennbar ist hierin die Dimension der flexiblen Zielanpassung. Beide Variablen werden jedoch nicht als kompetitiv angesehen, sondern vielmehr in den Rahmen eines synergistischen Ansatzes gestellt: „...assimilative and accomodative processes (...) may operate synergistically and complement each other in concrete episodes of coping. (…) life events typically involve a plurality of adaptive tasks that may call to various degress for assimilative persistence of accommodative flexibility” (Brandstädter & Rothermund, 2002, S. 123). Das Zwei-Prozessmodel auch auf die Bewältigung von Studienanforderungen anzuwenden, lag also gerade aus theoretischer Sicht nahe, bedenkt man, dass ein Studierender im Laufe seines Studiums mit verschiedenen Zielen und Anforderungen konfrontiert wird und hierbei immer wieder Entscheidungen über weiteres Investment oder aber Neuorientierungen und Neuadjustierungen anstehen. Durch die Hinzunahme der Skala „Flexible Zielanpassung“ wurde zwar etwas vom rein anforderungsanalytischen Vorgehen abgewichen, allerdings nur zur Erweiterung um eine theoretisch (und womöglich praktisch) relevante Dimension.
Die internen Konsistenzen werden für hartnäckige Zielverfolgung mit $\alpha = .80$ und für flexible Zielanpassung mit $\alpha = .83$ angegeben.

Für die Dimensionen Offenheit für Erfahrungen, Gewissenhaftigkeit und Neurotizismus als negativem Pol emotionaler Stabilität (zur Erfassung der anforderungsanalytischen Dimension „stabile Persönlichkeit“) dienten die gleichnamigen Skalen aus dem Neo-Fünf-Faktoren-Inventar (NEO-FFI) (Borkenau & Ostendorf, 1993), deren Konstruktbeschreibungen das Testhandbuch (Borkenau & Ostendorf, 1993, S. 28f.) wie folgt wiedergibt.

8.1.5.3 Offenheit für Erfahrungen

> Die Skala erfasst das Interesse an, und das Ausmaß der Beschäftigung mit neuen Erfahrungen, Erlebnissen und Eindrücken. Personen mit hohen Punktwerten geben häufig an, dass sie ein reges Phantasieleben besitzen, ihre eigenen Gefühle, positive wie negative, akzentuiert wahrnehmen und an vielen persönlichen und öffentlichen Vorgängen interessiert sind. Sie beschreiben sich als wissbegierig, intellektuell, phantasievoll, experimentierfreudig, und künstlerisch interessiert. Sie sind eher bereit, bestehende Normen kritisch zu hinterfragen und auf neuartige soziale, ethische und politische Wertvorstellungen einzugehen. Sie sind unabhängig in ihrem Urteil, verhalten sich häufig unkonventionell, erproben neue Handlungsweisen und bevorzugen Abwechslung ... (S. 28).

Die interne Konsistenz liegt bei $\alpha = 71$.

8.1.5.4 Gewissenhaftigkeit

„Personen mit hohen Punktwerten in der Skala beschreiben sich als zielstrebig, ehrgeizig, fleißig, ausdauernd, systematisch, willensstark, diszipliniert, zuverlässig pünktlich, ordentlich, genau und penibel.“ (S. 28)
Die interne Konsistenz liegt bei $\alpha = .85$.

8.1.5.5 Neurotizismus

> Die Skala erfasst individuelle Unterschiede in der emotionalen Stabilität und der emotionalen Labilität (*Neurotizismus*) von Personen. (....). Der Kern der Dimension liegt in der Art und Weise, wie Emotionen, vor allem negative Emotionen, erlebt werden. Personen mir hohen Ausprägungen in Neurotizismus geben häufiger an, sie seien leicht aus dem seelischen Gleichgewicht zu bringen. Im Vergleich zu emotional stabilen Menschen berichten sie häufiger negative Gefühlszustände zu erleben und von diesen manchmal geradezu überwältigt zu werden. Sie berichten viele Sorgen und

geben häufig an, z. B. erschüttert, betroffen, beschämt , unsicher, verlegen, nervös, ängstlich und traurig zu reagieren (S. 27).

Die interne Konsistenz liegt bei α =.85.

8.1.5.6 Lügen- und Leugnungsskalen

Zur weiteren Analyse der Verfälschbarkeit der Persönlichkeitsfragebögen in Selektionskontexten hinsichtlich Mittelwerts- und Varianzunterschieden sowie Kriteriumsvalidierungen wurden die Lügen- und Leugnungsskalen von Ling in das Verfahren aufgenommen.
Die Lügenskala besteht aus 14, die Leugnungsskala aus 18 Items im dichotomen Antwortformat, die nach der rationalen Methode konstruiert wurden. Die Scores beider Subskalen können zur eigentlichen Lügenskala als Gesamtwert der Tendenz, sozial erwünschte Antworten zu geben, aufsummiert werden.
Ling (zit. nach Amelang & Bartussek, 1970, S. 105f.) konnte in einer Faktorenanalyse mit einer Stichprobe von 288 Versuchspersonen zwei Faktoren extrahieren, die zu r = .51 interkorrelierten. Der erste Faktor bestand überwiegend aus „stimmt-Items“ und bezeichnet daher die „Tendenz, sich in ein günstiges Licht zu setzen, indem man sich sozial erwünschte Verhaltensweisen zuschreibt, ohne dass diese mit dem tatsächlichen Verhalten übereinstimmen“ (Ling, 1967, zit. nach Amelang & Bartussek, 1970, S. 106). Beispielhafte Items für die daraus entstandene *Lügensskala* sind: „Ich fälle niemals ein Urteil über andere Menschen, ehe ich nicht genau die Tatsachen kenne“ (Item 2) oder „Ich habe niemals das Gefühl gehabt, grundlos bestraft worden zu sein“ (Item 3). Auf dem zweiten Faktor luden überwiegend „stimmt nicht-Items“, weshalb Ling (Ling, 1967, zit. nach Amelang & Bartussek, 1970, S. 106) diese Items zu einer *Leugnungsskala* zusammenfasste, da sie die Tendenz erfasst, „sich in ein gutes Licht zu setzen, indem man sozial unerwünschte Verhaltensweisen nicht zugibt, obgleich man sie aufweist“ (Ling, 1967, zit. nach Amelang & Bartussek, 1970, S. 106). Typische Items dieser Skala sind: „Ich sage nicht immer die Wahrheit“ (Item 6) und „Ich werde manchmal ärgerlich, wenn mich andere um einen Gefallen bitten (Item 8). Amelang & Bartussek (1970) konnten in einer Studie an N = 198 Versuchspersonen die Validität beider Skalen nachweisen. Sie stellten hierbei die Hypothese auf, dass sozial erwünschte Antworten ein relativ eindeutiges Kriterium zur Beantwortung von Fragebogenitems darstellen und die Beantwortung demgemäß leichter fallen müsse als eine ehrliche Beantwortung, welche ungleich mehr komplexe Vergleichs- und Entscheidungsprozesse verlange. Durch diese Vereinfachung von Entscheidungen sollten sich bei wiederholter Erhebung im Wesentlichen

identische Ergebnisse ergeben. Hypothesenkonform zeigten sich höhere Retestreliabilitäten bei Personen mit steigender Tendenz, sozial erwünschte Antworten zu geben.
Für die Skalen werden bei Amelang und Bartussek (1970) lediglich Retest-Reliabilitäten, bezogen auf ein dreiwöchiges Intervall, von .87 (Lügenskala) bzw. .85 (Leugnungsskala) angegeben. Angaben zur internen Konsistenz nach Cronbachs Alpha fehlen.

Um für die eingesetzten Skalen den Einfluss von Verfälschbarkeit im Selektionskontext hinsichtlich verschiedener statistischer Parameter abzuschätzen, wurde die Hälfte der Testhefte mit einer Normalinstruktion zum Beantworten der Persönlichkeitsfragebögen versehen, die andere hingegen mit einer Faking-Good-Instruktion. Die beiden Instruktionen lauteten wie folgt:

Normalinstruktion:

Mit dem vorliegenden Fragebogen werden Ihre Einstellungen gegenüber Arbeit, Beruf, Leistung, persönlichen Zielen und Einstellungen erfasst. Nur Ihre persönliche Meinung zählt – **es gibt weder richtige noch falsche Antworten.**
Bitte lesen Sie jede Aussage genau durch und geben Sie an, inwieweit sie **auf Sie persönlich** zutrifft.

Faking-Good-Instruktion:

Mit dem vorliegenden Fragebogen werden Ihre Einstellungen gegenüber Arbeit, Beruf, Leistung, persönlichen Zielen und Einstellungen erfasst.
Achtung!: Stellen Sie sich beim Beantworten bitte vor, Sie müssten den Fragebogen im Rahmen einer Bewerbung auf einen Studienplatz ausfüllen und wollten einen Ihrer Meinung nach möglichst guten Eindruck hinterlassen!

Die Hintergründe und Rationale für dieses Vorgehen sollen im folgenden Exkurs ausgeführt werden.

8.1.5.7 Exkurs: Theoretischer Hintergrund und Rationale der Fragebogenanalyse unter einer Faking-good- und einer Normalinstruktion

Die Validität von Persönlichkeitsfragebögen bezüglich der Vorhersage von leistungsrelevanten Kriterien ist in einer Vielzahl von Studien und Metaanalysen gut belegt. Besonders stark beforscht ist dies im beruflichen Bereich wie z. B. genereller Leistungsfähigkeit (Barrick & Mount, 1991; Barrick, Mount & Strauss, 1993), in Beratungssituationen (McCrae & Costa, 1991) oder hinsichtlich kontraproduktiven Verhaltens in Organisationen (Moser, Schwörer, Eisele & Haefele, 1998; Robie, Born & Schmit, 2001). Insbesondere stellten sich hierbei immer wieder die Dimensionen Verträglichkeit, Gewissenhaftigkeit und Leistungsmotivation als prädiktiv für Berufserfolg heraus (Tett, Jackson & Rothstein, 1991; van den Berg & Feij, 2003). Auch im Kontext von Persönlichkeitsfragebögen konnte die Vorhersagevalidität nachgewiesen werden. Insbesondere zeigen sich hier die Persönlichkeitsdimensionen Verträglichkeit, Offenheit für Erfahrungen und Gewissenhaftigkeit des Big Five Personality Inventory (Costa & McCrae, 1992) als nützliche Prädiktoren (Chamorro-Premuzic & Furnham, 2003b; Farsides & Woodfield, 2003; Hair & Graziano, 2003) wie auch Leistungsmotivation und Selbst-Disziplin (Chamorro-Premuzic & Furnham, 2003b).
Diesen positiven Befunden steht allerdings das Problem potenzieller Verfälschbarkeit von Persönlichkeitsinventaren gegenüber. Verfälschbarkeit wird hierbei als mögliche Verzerrung von Antworten im Sinne sozialer Erwünschtheit definiert (s. hierzu etwa Ones & Viswesvaran, 1998). Zur Prüfung der Effekte sozialer Erwünschtheit lassen sich vier Forschungsrichtungen unterscheiden (Viswesvaran & Ones, 1999). Die erste Richtung (Bowen, Martin & Hunt, 2002; Cowles, Darling & Skanes, 1992) umfasst experimentelle Studien, bei denen die Instruktion variiert wurde (Faking-Bedingung vs. Ehrlich-Bedingung) und deren Effekte hinsichtlich verschiedener statistischer Parameter verglichen werden. Im allgemeinen dient dieses Vorgehen dazu, das Maximum an Verfälschbarkeit auszuloten (Viswesvaran & Ones, 1999). Hierbei unterscheidet man Innerhalb-Subjekt- oder Zwischen-Subjekt-Designs (Furnham, 1986). D.h., entweder bearbeitet dieselbe Stichprobe die Fragebögen zweimal, aber mit zwei verschiedenen Instruktionstypen, oder zwei verschiedene Stichproben bearbeiten diese mit verschiedenen Instruktionen. Beide Ansätze haben jwls. statistische wie auch inhaltliche Vorzüge und Nachteile (Cook & Campbell, 1979). Es besteht weitgehende Übereinstimmung dahingehend, dass Personen ihre Antworten in Richtung sozialer Erwünschtheit verfälschen können, wenn sie dazu angehalten werden. So konnten Viswesvaran and Ones (1999) in einer Metaanalyse zeigen, dass Rohwerte mittels Faking-Instruktionen um bis zu einer halben

Standardabweichung verändert werden können. Dies bedeutet allerdings noch nicht, dass Personen in realistischen Auswahlverfahren tatsächlich sozial erwünscht antworten (Rosse, Stecher, Miller & Levin, 1998). Daher verfolgt die zweite Forschungsrichtung (z. B. D. B. Smith & Ellingson, 2002) die Analyse der Effekte sozialer Erwünschtheit im Rahmen von Feldstudien. Studien dieser Forschungsrichtung belegen geringere Effektstärken (d = .30) der Verfälschbarkeit als die experimentell induzierten (d = .73) (Smith & Ellingson, 2002, zit. nach Peeters & Lievens, 2005), was nahe legt, dass es bei den experimentellen Designs zu „Worst-Case-Effekten“ kommt (Griffith, 1998; Rosse, Stecher, Miller & Levin, 1998). Auch wenn diese Effekte geringer sein mögen, als die experimentell induzierten, so besteht dennoch keine weitgehende Einigkeit hinsichtlich der praktischen Konsequenzen für den Selektionskontext. Einige Studien berichten, dass Bewerber nicht sozial erwünscht antworten und, dass selbst wenn sie dies tun, dies keine negativen Konsequenzen für die Validität zeigt (Ellingson, Smith & Sackett, 2001; Marcus, 2003; Ones & Viswesvaran, 1998). Andere Studien hingegen können einen Effekt sozialer Erwünschtheit im Selektionskontext nachweisen mit der Folge, dass sich die Kriteriumsvalidität verringert (Rosse, Stecher, Miller & Levin, 1998; Schmit & Ryan, 1993; Zickar, 1997) und die Antwortverfälschung dazu beiträgt, welcher Bewerber zugelassen wird (Rosse, Stecher, Miller & Levin, 1998).

Eine dritte Forschungsrichtung beschäftigt sich mit der Erfassung sozialer Erwünschtheit über eigens hierzu konstruierter Skalen (SE-Skalen). Die Annahme besteht darin, dass Personen mit hohen Werten auf diesen Skalen nicht ehrlich antworten und vice versa (Paulhus, 1991). Die Hauptkritik an dieser Hypothese ist allerdings, dass soziale Erwünschtheit nicht gleichgesetzt werden kann mit der Tendenz, falsche Antworten zu geben, sondern vielmehr einen für die Prognose von Berufserfolg wichtigen Trait darstellt, der, wenn er kontrolliert wird, die Validität verringert (s. hierzu etwa Nicholson & Hogan, 1990). Demzufolge dürften Personen mit hohen Punktwerten auf SE-Skalen nicht ungerechtfertigterweise bestraft werden.

In diesem Zusammenhang ist auch die Studie von Amelang, Schäfer und Yousfi (2002) zu sehen, in welcher die Verfälschbarkeit von verbalen und non-verbalen Persönlichkeitsfragebögen untersucht wurde. Als Hauptbefund zeigte sich unter der Faking-good-Bedingung zunächst für beide Fragebogenformen ein Abfall der Validität, welche die Autoren über Peerbeurteilungen operationalisierten. Die Koeffizienten lagen allerdings immer noch um .40. Im Kontext der oben genannten Hypothese, dass soziale Erwünschtheit nicht gleichzusetzen ist mit der Tendenz, in Skalen zu deren Erfassung falsche Antworten zu geben, erwies sich ein Nebenbefund dieser Studie als zentral. Es ergab sich hierbei, dass SE-Skalen positiv mit den Peerbeurteilungen korrelierten. Dieser Befund lässt in der Tat vermuten, dass diese Skalen

neben Varianz sozialer Erwünschtheit *kriteriumsrelevante* Varianz enthalten. Um das Ausmaß der Antwortverfälschung in den selbstberichteten Persönlichkeitsdimensionen abschätzen zu können, partialisierten die Autoren die kriteriumsrelevante Varianz der SE-Skalen aus deren Korrelationen mit den Selbstbeurteilungen aus. Es resultierte ein lediglich geringer Abfall der Korrelation zwischen SE-Skalen und den Selbstbeurteilungen, wobei die Höhe dieser Korrelationen vergleichbar mit denen der Peerbeurteilungen bezüglich der Persönlichkeits- und SE-Skalen war. Dies unterstützt die Hypothese dieser Forschungsrichtung, da die SE-Skalen offenbar neben Varianz zulasten sozialer Erwünschtheit auch Varianz enthalten, welche für die Vorhersage eines Kriteriums prädiktiv ist.

Die vierte Forschungsrichtung fokussiert auf Moderatoren der Situation und der Persönlichkeit, welche zu Verfälschungstendenzen führen (s. hierzu etwa McFarland & Ryan, 2000).

Der in der vorliegenden Arbeit durchgeführte experimentelle Ansatz mit einem Between--subject-Design zur Überprüfung der Effekte des Fakings hinsichtlich verschiedener diagnostisch relevanter Parameter wie Mittelwerts- Varianz- und Validitätsdifferenzen ergab sich aus der Einbettung der Fragebögen in die große Anzahl weiterer Verfahren. Ein Within-subject-Design wäre gerade aus der Perspektive einer Kontrolle der intraindividuellen Antwortvarianz adäquater gewesen, konnte aber aus Zumutbarkeitsgründen nicht durchgeführt werden. Zur zusätzlichen Überprüfung der Auswirkungen und für differenziertere Analysen wurden dem Verfahren die Skalen zur Erfassung der sozialen Erwünschtheit von Ling (1967) hinzugefügt.
In Analogie zu den anderen Fragebögen wurden auch diese in die experimentelle Variation der Instruktion Faking-good-Instruktion vs. Normalinstruktion eingebunden. Diese Skalen als zusätzliche Informationsquelle miteinzubeziehen war sinnvoll, da eine *nicht* nachweisbare Mittelwertsdifferenz der Skalenwerte zwischen der Normal- und der „Faking-good“--Instruktion kein zwingenden Beleg für die Unverfälschbarkeit eines Fragebogens darstellt, sondern das Ergebnis unterschiedlicher Vorstellungen über wünschenswertes Verhalten darstellen kann. Verfälschungstendenzen in Bezug auf Mittelwertsunterschiede würden sich hierdurch kompensieren (s. hierzu u.a. Gordon & Gross, 1978). Demgegenüber würden hohe Korrelationen zwischen den Werten einer Lügen- oder Leugnungsskala mit der zu untersuchenden Skala für eine Verzerrung des Testwertes in Richtung sozialer Erwünschtheit sprechen, auch wenn kein signifikanter Mittelwertsunterschied zwischen den Instruktionsversionen auftreten sollte.

8.1.6 Zusammenfassende Betrachtung der Itemkonstruktion und Zusammenstellung der Persönlichkeitsskalen

Bei der Umsetzung der anforderungsanalytisch bestimmten Dimensionen in geeignete Testverfahren sollte versucht werden, eine Balance zwischen inhaltlicher Breite einerseits und objektiver, reliabler, konstruktvalider und ökonomischer psychometrischer Umsetzung andererseits zu finden. Um den letztgenannten Ansprüchen gerecht werden zu können, musste auf die anforderungsanalytisch ermittelte Dimensionen soziale Kompetenz, Selbstständigkeit und Kooperation sowie instrumentelle Intelligenz verzichtet werden. Umsetzungsprobleme bereitete zudem die inhaltliche Breite der Dimensionen Kreativität und Problemsensitivität, weshalb hierbei lediglich eine Fokussierung auf einzelne Konstruktfacetten möglich war. Dass dies womöglich zulasten der Reliabilität und Kriteriumsvalidität gehen könnte, musste allerdings auch auf dem Hintergrund der Zumutbarkeit des Verfahrens hingenommen werden. Neben der Orientierung an den Ergebnissen der Anforderungsanalyse spielten weitere Erwägungen bei der Skalenauswahl ebenso eine Rolle. Dies betraf die Auswahl zu testender Intelligenzdimensionen, im Weiteren besonders die Aufgabenstellung „Erfahrungswissenschaftliches, empiriebezogenes Denken“ sowie die Auswahl der Skala flexible Zielanpassung, die nach theoretischer Maßgabe des Zwei-Prozess-Modells der Zielverfolgung von Brandstädter (2002) in den Katalog der Verfahren aufgenommen wurde.
Für die Intelligenzsubtests wurden jeweils 20 Items vollkommen oder weitestgehend regelgeleitet konstruiert (s. Subtests Zahlenreihen, Zahlenmatrizen, Matrizen, Odd-One-Out), oder aus Standardverfahren (s. verbale Analogieaufgaben) zusammengestellt. Die Angemessenheit der ad hoc bemessenen Testzeitbegrenzungen wurde an einer kleinen studentischen Stichprobe überprüft und gegebenenfalls korrigiert. Subtests spezifischerer Leistungsanforderungen stellten zum einen die Skala zur Erfassung der passiven argumentativ-rhetorischen Kompetenz von Flender et al. (1996) wie auch eine frei zu beantwortende Aufgabenstellung zum empiriebezogenen Denken dar.
Insgesamt ergibt sich für die Skalenkonstruktion und Fragebogenzusammenstellung eine leichte Abweichung vom streng anforderungsanalytisch orientierten Vorgehen bei der Umsetzung der abgeleiteten Personenmerkmale durch Einbezug rationaler Skalenkonstruktionsaspekte. Durch diese Ausweitung sollte allerdings der bereits in Kapitel 7.3 angedeuteten, möglicherweise mangelnden fachspezifischen Trennschärfe der anforderungsanalytischen Ergebnisse etwas entgegengewirkt und der Prädiktorenraum um theoretisch relevante Konstrukte erweitert werden.

Den methodischen Teil dieser Arbeit abschließend, soll im Folgenden eine Begründung für die Testkonstruktion nach Maßgabe des Rasch-Modells gegeben werden. Dies scheint sinn- und zweckvoll, da ohne ein grundlegendes Verständnis der Hintergründe des Rasch-Modells die Ergebnisinterpretation der Modelltestung schwer nachvollziehbar wäre.

9. Zu den Konstruktionen der Skalen: Prinzipien und Begründung der Skalenanalysen nach dem Rasch-Modell

„Eine wichtige Voraussetzung (*conditio sine qua non*) jeglicher Modellierung besteht darin, die Abneigung gegen systematische Modellentwürfe, die vor allem empirisch arbeitende Wissenschaftler hegen, schrittweise zu verringern. Die Autoren hoffen aufrichtig, dass diese Kritik das Vertrauen in die Modellbildung festigen wird“ (Düchting, Ulmer & Ginsburg, 1996).

Wenn auch obiges Zitat nicht im Kontext von Antwortmodellen in der Psychometrie geäußert wurde, so beschreibt es dennoch sehr genau die derzeitige (meist ablehnende) Position vieler Testkonstrukteure innerhalb der psychologischen Diagnostik gegenüber dem Rasch-Modell. Fragt man nach den Gründen für diese ablehnende Haltung, so erhält man im Wesentlichen zwei Antworten. Zum einen wird argumentiert, das Rasch-Modell mit seiner Annahme konstanter Itemtrennschärfen sei zu restriktiv für den Messbegriff in der psychologischen Diagnostik, zum anderen werde beim Rasch-Modell wie auch in der Klassischen Testtheorie (KTT) der Summenscore als Maßzahl der Personenfähigkeit herangezogen, demzufolge auch beim Rasch-Modell nichts anderes „herauskomme“. Im Folgenden sollen beide Argumente jeweils getrennt hinsichtlich ihrer Implikationen und Konsequenzen genauer analysiert werden. Hierbei ist es nötig, besonders die Forderung nach konstanten Itemtrennschärfen messtheoretisch zu begründen, da sie das wesentlichste Bestimmungsstück des Rasch-Modells darstellt, gerade in Bezug auf die Definition der Messung an sich. Hierzu ist es wesentlich, diese Forderung auf die historische Entwicklung des Messbegriffs innerhalb der Psychologie zu beziehen bzw. sie daraus abzuleiten.

Für die eingehende Betrachtung des Standpunktes, das Rasch-Modell sei für die psychologische Diagnostik zu restriktiv, muss zunächst geklärt werden, auf welchen grundsätzlichen Annahmen Kritiker wie auch Befürworter des Rasch-Modells ihre jeweiligen Argumente

bauen. Kritiker des Rasch-Modells sehen in der Annahme der Konstanz der Itemtrennschärfen den Hauptgrund, weshalb das Rasch-Modell unrealistisch sei und deshalb oft genug nicht die Daten erklären könne (s. etwa Goldstein, 1980). Reduziert man nun dieses Argument auf sein eigentliches Ziel, nämlich die Beschreibung der Datenstruktur durch ein Modell, so wird gerade daran deutlich, dass es sich bei der Kontroverse nicht darum dreht, welche Methode die geeignetere sei, die Daten auszuwerten, sondern es sich vielmehr um zwei grundsätzlich verschiedene Forschungs-Paradigmen im Sinne Kuhns (1970) handelt. Nach dem traditionellen Paradigma wird ein Modell einem anderen vorgezogen, wenn es den Daten besser entsprechen kann, etwa, indem es mehr Varianz aufklärt oder nach Maßgabe statistischer Fit-Indizes besser passt. Diese Sichtweise gilt für Vertreter der mehrparametrischen Antwortmodelle wie dem Birnbaummodell (1968), bei welchem die Itemtrennschärfen variieren dürfen und gleichermaßen auch für solche der explorativen Faktorenanalyse als dem Modell kongenerischer Messungen in der Klassischen Testtheorie, bei welchem die Trennschärfen in Form von Faktorladungen unrestringiert sind. Bock und Jones (1968) drücken diesen Standpunkt klar aus, wenn sie schreiben: “First, there is the difficulty of finding a model that fits the available data and estimating model parameters” (S. 73) .... [W]hen the criterion indicates nonrandomness, an examination of residuals may suggest how the model should be modified to improve fit”(S. 5). Einen gänzlich entgegengesetzte Position verfolgen Fürsprecher des Rasch-Modells, denn wie Fisher (2004) ausführt:

> In contrast with mainstream contemporary practice in psychological measurement, the procedure does not begin with the attitude that a wide variety of models differing in the rigor of their qualitative and quantitative requirements are to be fit to data, with the model that best describes the data and all of its empirical and accidental vagaries chosen as the one with which to proceed. In this context, models are devised with the goal of describing the particular interactions and sample-dependencies characteristic of the data in hand (….) In order to mount a test of the immanent, internal coherence of all that is intrinsic to an eidos, [das abstrakte Gesetz eines Merkmals, *Anmerkung des Verfassers*] we must first specify a model that is not mathematical in a merely quantitative sense but which is fully mathematical in requiring *that the eidos be*

> *distinguished and separable from the particular questions and answers participating in it* [Hervorhebung vom Verfasser]. (S. 14)

Im Rasch-Modell ist es also das im wörtlichen Sinne maß-gebliche Ziel, bestimmte Invarianzeigenschaften beim Vergleich von Personen und Items herzustellen, sodass Aussagen über das zu untersuchende Merkmal an sich möglich sind, indem diese nicht mehr vom spezifischen Untersuchungsinstrument und der Personenstichprobe abhängen. Aus dieser Forderung ergibt sich bereits rein logisch die Konsequenz, dass das Rasch-Modell kein Modell zur *Beschreibung* der Daten sein kann, sondern, dass es vielmehr die *Anforderungen an die Daten* bzw. das Erhebungsinstrument stellt, sollen verallgemeinerbare, also invariante Aussagen möglich sein. Vereinfacht gesprochen muss also nicht das Modell den Daten, sondern die Daten müssen dem Modell genügen. Dies deutet bereits an, dass es sich beim Rasch-Modell tatsächlich um ein neues Paradigma handelt, indem es die Etablierung einer echten Messstruktur in der Psychologie darstellt. Mit einer *echten* Messstruktur ist hierbei nicht weniger gemeint, als dass Messwerte als solche erst bezeichnet werden können, wenn sie die oben genannte Personen- und Iteminvarianzeigenschaften aufweisen und die Relation zwischen beiden additiv ist. Insbesondere die Additivitäts-Bedingung eines (echten) Messwerts ist hierbei von zentraler Bedeutung. Sie zeigt sich darin, dass die Itemantworten nur durch eine Linearkombination der beiden Messmengen Personenfähigkeiten und Itemschwierigkeiten bedingt werden. Konkreter gesprochen beantwortet eine additive Messstruktur die Frage: „If a person A has more ability than person B, then how much ‚ability' must be added to B to make the performance of B appear the same as the performance of A" (Wright, 1985, S. 103). Eine Frage also, deren Antwort in der Klassischen Testtheorie durch Differenzbildung von Summenscores bereits als beantwortet vorausgesetzt wird! Im Folgenden soll nun durch einen Rekurs auf die historische Entwicklung des Messbegriffs in der Psychologie zum einen erläutert werden, wie die formale Struktur einer additiven Messung in der Psychologie aussehen muss, weshalb die oben genannten Invarianzeigenschaften für psychologische Messbereiche von Bedeutung sind und weshalb gerade für diese Anforderungen das Rasch-Modell eine notwendige bis hinreichende Bedingung darstellt.

9.1 Messen in der Psychologie: Ein gelöstes Problem oder eine problematische „Lösung“?

Die Forderung nach Invarianzeigenschaften psychologischer Messungen ist nicht etwa erst eine von Georg Rasch (1960) gewesen, sondern bereits von Fechner (1851) für die Psychologie grundlegend verlangt worden und sie findet sich ebenso in Spearmans Konzept einer „General mental ability“ (Spearman, 1904) wieder, hat also eine weit zurückreichende Tradition innerhalb der Psychologie. Insbesondere Thurstones Forderungen nach Invarianz-eigenschaften psychologischer Testdaten und seine daraus abgeleiteten Skalierungsmethoden können als Antizipationen des Rasch-Modells gesehen werden. So stellt das folgende Zitat bereits eine explizite Forderung der Stichprobenunabhängigkeit der Itemparameter dar (Thurstone, 1928):

> Within the range of objects for which the measuring instrument is intended, its function must be independent of the object of measurement…. A measuring instrument must not be seriously affected in its measuring function by the object of measurement. If a yardstick measured differently because of the fact it was a rug, a picture, or a piece of paper that was being measured, then to that extent the trustworthiness of that yardstick as a measuring device would be impaired. Within the range of objects intended, its function must be independent of the object of measurement. (S. 547)

Bereits in einem zuvor erschienenem Aufsatz (Thurstone, 1926, S. 446) erkannte er allerdings auch die Notwendigkeit von Invarianzeigenschaften nicht nur auf Item-, sondern ebenso auf Personenmesswertebene, wenn er schreibt: „...it should be possible to omit several test questions at different levels of the scale without affecting the individual score“.
Rasch (1961) greift eben diese Vorstellung einer Symmetrie von Invarianzeigenschaften zwischen Personen und Items wieder auf, wenn er fordert: “The comparison between 2 stimuli should be independent of which particular individuals were instrumental for comparison ... Symmetrically, a comparison between 2 individuals should be independent of which particular stimuli within class considered were instrumental for comparison…”
Allerdings beklagte Thurstone (1952) 24 Jahre nach seinen Ausführungen zu Invarianzeigenschaften von Messungen:

> The excuse is often made that social phenomena are so complex that the relatively simple methods of the older sciences do not apply. This argument is probably false. The analytical study of social phenomena is probably not so difficult as is commonly believed. The principal difficulty is that the experts in social studies are frequently hostile to science. They try to describe the totality of a situation and their orientation is often to the market place or the election next week. *They do not understand the thrill of discovering an invariance of some kind which never covers the totality of any situation. Social studies will not become science until students of social phenomena learn to appreciate this essential aspect of science* [Hervorhebungen vom Verfasser]. (S. 297)

Weshalb aber sah Thurstone den Status der sozialwissenschaftlicher Studien als wissenschaftlich so spezifisch von Invarianzeigenschaften abhängig? Womöglich kommt es hierbei nicht von ungefähr, dass dieser Artikel ein Jahr nach dem Erscheinen von Stevens *Handbook of Experimental Psychology* (1951) erschien. Stevens legte mit diesem Werk den Grundstein dessen, was bis heute in der Psychologie als Messung angesehen und gehandhabt wird und im Satz „kondensiert“: „...measurement is the assignment of numerals to objects or events according to a rule“. Wesentlich hierbei ist, dass Stevens keine Regel meinte, die zu numerischen Repräsentationen führt, welche die vom Messvorgang ganz *unabhängig* existierenden empirischen Relationen widerspiegeln, wie dies im vorangehenden klassischen Konzept der Messung des Repräsentationismus im Konzept einer homomorphen Abbildung verankert ist (s. u.a. N. R. Campbell, 1940). Vielmehr verstand Stevens hierunter jegliche Form numerischer Modellierung empirischer Strukturen als Regel, welche die Messung *definiert*, wenn er schreibt: „...provided a consistent rule is followed, some form of measurement is achieved“ (Stevens, 1959, S. 19).

Diese Messdefinition schafft jedoch weitaus mehr Probleme, als sie Lösungen bietet, was im Folgenden näher begründet werden soll. Das wohl grundlegendste Problem hierbei ist zunächst, dass vorausgesetzt wird, dass das zu untersuchende Merkmal quantitativ ist. Doch was bürgt in der Stevensschen Messkonzeption dafür, dass das numerische Relativ ein in seiner Struktur *quantitatives* Abbild des empirischen ist? Michell (1999) bemerkt hierzu kritisch:

> However, in so far as it is held that these procedures measure psychological attributes, there has been little serious scientific research undertaken to show that the relevant attributes are really quantitative and, therefore, that the relevant attributes are measurable. If they are not, then their character and their relationship to the standard number-generating procedures that psychologists use needs to be investigated. (S. 187)
>
> Against this hypothesis, it is known that distinct causal systems ... can systematically produce the same effect (the same score on a test) under similar conditions because of performances on mental tests, this alternative is *prima facie* plausible because exactly the same test score can result from quite different patterns of right and wrong answers, indicating the possibility of causal processes involving different attributes resulting in the same score. (S. 167)

Ebenso denkbar wäre es daher, dass qualitative anstelle von quantitativen Unterschieden hinter den Testscores stehen. Dies soll nun nicht bedeuten, dass sich nicht-quantitative Merkmale einer wissenschaftlichen Untersuchung entziehen. Beispielhaft sei hier etwa die mathematisch-linguistische Analyse grammatikalischer Strukturen anhand formaler Sprachtheorien genannt. Allerdings sind deren Schlussfolgerungen rein formal-logisch und geben daher nicht vor, dass die innere, natürliche Struktur der Sprache quantitativ ist. Vielmehr handelt es sich somit nicht um eine Messung, sondern um die *Erfassung* eines Merkmals. Was quantitative Merkmale von qualitativen also unterscheidet, sind ihre verschiedenen *internen* Strukturen: Quantitative Strukturen werden über Relationen definiert, welche die zu untersuchenden Objekte *natürlicherweise* aufweisen und welche durch die Messungen widergespiegelt werden bzw. *repräsentiert* werden müssen. „Measurement in its widest sense may be defined as the assignment of numerals to things *so as to represent facts or conventions about them*" [Hervorhebung vom Verfasser] (N. R. Campbell, 1940, S. 340). Diese quantitativen Messstrukturen müssen daher bestimmte Axiome erfüllen, wie das Beispiel der Verknüpfungsrelation (concatenation operation) des Repräsentationismus (s. u.a. N. R. Campbell, 1920, 1921/1952, 1940) zeigt: Zwei Gewichte lassen sich auf einer Waagschale zusammenlegen („verknüpfen") und deren Gesamtgewicht anhand eines dritten Gewichtes auf

der anderen Waagschale vergleichen, weil die Gewichte eine natürliche *additive* Struktur aufweisen und sich somit eine sinnvolle numerische Addition der Skalenwerte ergibt. Es besteht daher eine Eins-zu-eins-Relation zwischen der Struktur der Addition und Subtraktion von *Zahlen* und der Struktur der *Eigenschaften* der zu messenden Objekte. Diese repräsentationale Konzeption der Messung hat daher zum Ziel, die strukturellen Gesetzmäßigkeiten der Messobjekte aufzudecken: „...measurement [is] only a means to an end: we want to express the properties of systems by numerals *only because we are thereby enabled to state laws about them*" [Hervorhebung vom Verfasser] (N. R. Campbell, 1921/1952, S. 328). „Messen" im eigentlichen Sinne bedeutet daher das Vorhandensein einer additiven Merkmalsstruktur und deren Abbildung in eine ebenso geartete Messstruktur. Für die Messungen in der Psychologie ist daran wesentlich, dass dadurch eine feste Maßeinheit definiert werden kann, welche ihre Bedeutung über das gesamte Kontinuum der Merkmalsvariable behält und folglich invariant ist. Bildet man nämlich in der Psychologie Differenzen von Summenwerten, dann wird diese additive Struktur des zu untersuchenden Merkmals unterstellt. Erst wenn diese Annahme zutrifft, würde die Differenzbildung eine sinnvolle Abbildungsrelation darstellen und daher das Merkmal als quantitativ anzusehen sein. Die quantitative Struktur verschiedener psychologischer Eigenschaften ist aber eine *empirisch* zu überprüfende Hypothese, die, wenn lediglich vorausgesetzt, gegebenenfalls falsch ist und demzufolge zu fehlerhaften Schlussfolgerungen über psychische Attribute führen muss. Das Problem einer empirischen Überprüfung der Additivitätsannahme übergehen allerdings z. B. Lord und Novick (1968) mit einem rein statistischen Argument, wenn sie schreiben:

> If we construct a test score by counting up correct responses (zero-one scoring) and treating the resulting scale as having interval scale properties, the procedure may or may not produce a good predictor of some criterion. To the extent that this scaling produces a good empirical predictor the stipulated interval scale is justified. (S. 22)

Schon rein logisch betrachte kann die Güte von Testscores als Prädiktor nicht die Frage beantworten, ob sie Intervallskalen und Indikatoren quantitativer psychologischer Attribute sind. Ob sich Testscores eines Tests als nützliche Prädiktoren erweisen ist vielmehr ein praktisches Problem. Erweist sich ein Testscore als brauchbar in einer Vorhersage, so wirft dies erst die Frage auf, warum er dies ist. Es ist aber keine Antwort auf die Frage, ob überhaupt etwas im eigentlichen Sinne gemessen wurde, d.h. eine *homomorphe* Abbildung überhaupt

vorliegt. Man mag einwenden, dass sich über Testscores Ordnungsrelationen und Differenzen zwischen Personenleistungen bilden lassen. Jedoch beziehen sich diese Relationen von Testscores zunächst nur auf das *numerische* Relativ, wohingegen aber die *empirischen* Relationen von eigentlichem Interesse sind. Es muss also erst gezeigt werden, dass das numerische Relativ eine bedeutsame Abbildung des empirischen ist. Solange man also nicht weiß, wann zwei verschiedene Antwortmuster als gleichwertige Leistungen anzusehen sind oder welches eine bessere Lösung darstellt, sind nicht einmal Ordnungs- (Ordinalskala) oder sogar Äquivalenzrelationen (Nominalskala) gegeben (Fischer, 1974).

Das grundlegende Problem der Stevensschen Konzeption der Messung besteht also darin, dass sie das klassische Messkonzept des Repräsentationismus auf den Kopf stellte. Der Kontrast kann kaum deutlicher beschrieben werden, wenn Michell (1999, S. 160) bemerkt: "If measurement involves making numerical assignments to things (as the representational view has it) according to definite operations then, in accordance with operationism, measurement itself is operationally defined by the general features of this process". Wo der Repräsentationismus forderte, dass sich eine Messung aus der impliziten Struktur des Messobjekts ergeben müsse, dort kehrte Stevens dies um, indem sich laut seiner Definition messbare Eigenschaften durch Zuweisung von Zahlen ergeben. Die simple Annahme aber, dass eine noch so konsistente Regel für die Zuweisung von Zahlen zu Objekten eine empirische Relation reflektiert, ist alles andere als entdeckt zu haben, dass sie dies auch tut. Sie stellt ein Zählen im weitesten Sinne dar, nicht jedoch eine Messung im eigentlichen Sinne eines Nachweises der additiven Struktur einer Variablen.

Stevens war sich allerdings dieser Schwäche seiner Messdefinition durchaus bewusst und versuchte daher, die operationale Messdefinition methodisch zu untermauern. Er nahm hierzu an, dass die Basis jeder wissenschaftlichen Entwicklung von Begriffen das Konzept der Klassifikation sei und, dass eine Begriffsklasse anhand operationaler Kriterien definiert werden kann: „Classification can proceed only when we have criteria defining the conditions of class-inclusion, and these criteria are essentially operational tests" und, dass „the concept of that class *is defined by* the operations which determine inclusion within the class" (Stevens, zit. nach Michell, 1999, S. 171).
Die somit zwangsläufig entstehende Fülle möglicher Definitionen eines Konstruktes wird dabei als „Tugend" dieses Paradigmas angesehen, wenn Kerlinger schreibt (1979):

> Achievement may be defined by citing a standardized achievement test, a teacher-made achievement test, or grades assigned by teachers. We here have three distinctly different ways of operationally defining the same construct. The reader should not let this multiplicity of operational definitions bother him; it is part of their flexibility and strength. After all, a construct like achievement has many facets, and researchers can be interested in different facets at different times. (S. 41)

Man könnte in Kerlingers Sinne weiterführen, dass es über den Facetten operationaler Definitionen abstraktere Operationen gebe, auf welche diese wiederum rückführbar seien und sich hierüber *komplette* Definitionen von Konstrukten ergeben. Jedoch tut sich bereits auf Ebene der Facetten-Definitionen ein grundlegendes Problem auf: welches Kriterium entscheidet über die Angemessenheit der Auswahl von Operationen für die jeweilige Konstruktdefinition? Und wie kann dann entschieden werden, ob eine Facette auf eine abstrakter definierte Operation reduzierbar ist und vor allem, wann eine Definition komplett ist? Bell, Staines und Michell (2001) führen daher auch aus:

> Ironically, the very goal of operational definitions which were proposed to allow objective and reliable terms cannot be achieved simply by spelling out one or a few operational aspects of an otherwise complex concept ... the choice of ‘operations’ may be just as arbitrary or restricted as the verbal definitions which they supposedly replace. (S. 113)

Es wird bei diesem Paradigma allerdings meist weitgehend übersehen, welcher Preis für diese Flexibilität gezahlt werden muss. Besonders häufig wird dabei eine folgenreiche, rein logische Implikation des Operationalismus übergangen: Eine Operation stellt das Verfahren dar, mit dem ein Forscher ein Merkmal identifiziert. Dies impliziert jedoch nicht, dass dieses Merkmal nicht auch ohne oder anders als durch die jeweilige Operation existieren kann, da diese laut Kerlingers Zitat lediglich eine Facette des Konstrukts darstellt. Wenn nun aber das Merkmal auch ganz unabhängig von dieser bestimmten Operation bestehen kann, dann kann diese und das Merkmal nicht dasselbe bedeuten, sind also logisch distinkt. Jede Änderung in der

Operation würde streng genommen nach der Logik des Operationalismus zu einer Redefinition des Phänomens führen. Michell (1999, S. 170) bringt diese häufig übersehene Implikation des Operationalismus auf den Punkt, wenn er schreibt: „...[they] confused the knowing of something, with the thing known". Aber wie lassen sich dann überhaupt verallgemeinerbare Aussagen treffen? Die Konsequenz für die psychologische Forschung ist daher die Einschränkung auf die Operationen zur Erfassung eines Merkmals: „Science simply reduces to the study of our operations and cannot be construed as the study of an independently existing world whose secrets we penetrate via these operations" (Michell, 1997, S. 376). Es wäre demgemäß stringenter, operationale Definitionen als *Kriterien* für den wissenschaftlichen *Gebrauch* von Begriffen anzusehen, als eine sinnvolle Möglichkeit, für eindeutige, vollständige und vor allen Dingen verallgemeinerbare Begriffs-*Definitionen*: „...operations can best employed to determine if and when a term is applicable" (Bell, Staines & Mitchell, 2001, S. 114). Es ist allerdings aufschlussreich zu sehen, dass die psychologischen Forschungsaussagen, die über operationale Definitionen getätigt werden, über diejenigen des Operationalismus eigentlich weit hinausgehen (Michell, 1999):

> Yet no psychologist really means by intelligence, scores on an intelligence test, and in believing intelligence to be measurable, psychologists typically theorise about it as a quantitative attribute, on continuously related to other attributes. The ideology of operationalising, therefore, completely obscures what is really going on: psychologists are caused to ignore the distinction in meaning between theoretical concepts, like intelligence, and their observable effects, like test scores…(S. 188)

Betrachtet man bspw. allein die Aussagen der Erblichkeitsstudien zur Intelligenz (s. u.a. Petrill & Deater-Deckard, 2004; Plomin & Spinath, 2004), Metaanalysen zur allgemeinen Intelligenz im Berufskontext (Schmidt & Hunter, 1998, 2004; Schmidt, Hunter & Outerbridge, 1986) oder diejenigen zu Unterschieden zwischen Weißen und Schwarzen bezüglich Allgemeiner Intelligenz (s. z. B. Jensen, 1985), so sind dies alles Aussagen, die über den zulässigen Interpretationsrahmen operationaler Definitionen hinausgehen, da sie intendieren, verallgemeinerbare, invariante Aussagen über Merkmale an sich zu treffen und nicht über Operationen zu deren Erfassung! Es ist in diesem Zusammenhang umso erstaunlicher, dass bereits Spearman (1927, S. 15) an dieser Art des Versuches, zu wissenschaftlichen Aussagen

zu gelangen, kritisierte: „Test results and numerical tables are further accumulated; consequent action affecting the welfare of persons are proposed, and even taken, on the grounds of – nobody knows what“. Eben diese Praxis, die *Bedeutung* eines Konstruktes post hoc aus den statistischen *Fakten* oder *Effekten* zu „lesen“, führt schließlich zu einer rein logischen Absurdität wie Maraun (1998) bemerkt:

> But phenomenons are conceptualized, and to conceptualize one must understand the meaning of concepts. Their error here is akin to claiming that ‘the mountain is a place where skiers go’ and then forgetting that to know this *fact* about mountains presupposed a criterion for *mountain* (i.e. the concept's meaning). (S. 497-498)

Darüber hinaus führt der Operationalismus und die mit ihm verbundene Messdefinition von Stevens als Forschungsparadigma in die Sackgasse einer tautologischen und zirkulären Definition, betrachtet man alleine die Definition der Intelligenz. Intelligenz ist nach dem Operationalismus kein Begriff, der bereits *vor* einer Messung und *unabhängig* von ihr definiert worden ist. Im Gegenteil: Intelligenz wird als Leistungsfähigkeit beim Lösen neuer Probleme verstanden. Aber diese Leistungsfähigkeit wird wiederum durch ein oder mehrere (operational definierte!) Testergebnisse erfasst. Offensichtlich fehlt hier bereits eine Theorie, welche eine Definition der Intelligenz *vor* ihrer Messung gibt. Zwar existieren in Fortführung von Spearman faktorenanalytisch begründetet Intelligenzdefinitionen. Aber worauf basieren diese? - Auf Testergebnissen, was folglich die Zirkularität abschließt.
In inhaltlicher Fortführung zu Spearman (1927, S. 15) merkt Roskam (1989) daher an:

> ...when a researcher reports that such-and-such behavior is correlated with intelligence, you better not consult a textbook, but ask him which intelligence test he used. That, of course, brings us inevitably to the most stupid kind of definition which a scientist can give, namely: intelligence is what this test measures. Although this is formally correct as an operational definition, it is virtually meaningless and unproductive since it provokes the next question: “what is it that is measured by this test?” – where, of course, the answer “it measures intelligence” is not very enlightening and profoundly

> tautological .... this network is circular: we believe that scholastic achievement depends on intelligence (according to a common connotation), and because certain tests look like problem solving tasks and correlate with scholastic achievement, we believe that these tests measure intelligence. (S. 238-239)

Es dürfte neben den wissenschaftstheoretischen Problemen, die sich aus der Stevensschen Messkonzeption ergeben, nun auch klar werden, was sie von der eingangs zitierten von Thurstone (1928) trennt: eine operationale definierte Messung kann per se nicht zu invarianten Ergebnissen führen, weil sie je nach Operationalisierungsvorschrift scheinquantifiziert sind und anders ausfallen müssen.

Um einem an dieser Stelle wahrscheinlichen korrelationsstatistischen Einwand entgegenzutreten: Zu zeigen, dass zwei oder mehr Tests desselben Konstruktbereiches im Sinne der konvergenten Validität hoch bis sehr hoch miteinander korrelieren und daraus abzuleiten, dass sie dasselbe messen und die Ergebnisse auch invariant seien, ist eine häufige, aber ganz eigentlich falsche Schlussfolgerung. Und zwar wesentlich aus zwei Gründen. Erstens kann, wie bereits weiter oben ausgeführt, aus einem quantitativen Effekt wie Testscores nicht geschlussfolgert werden, dass a) die zugrunde liegenden Variablen ebenfalls quantitativ und vor allem b) dieselben sind. Zwei Personen können z. B. über zwei gänzlich unterschiedliche (qualitative!) Lösungsalgorithmen zum selben Testscore gelangt sein (s. z. B. Köller, Rost & Köller, 1994) wie auch Michel betont (2001, S. 213).: „...just because performances on such tests possess quantitative features (e.g. test scores), it does not follow that these features reflect the workings of exclusively quantitative causes" Daher kann also auch die Rangreihe individueller Testleistungen in zwei Tests desselben Messgegenstandes zwar hoch korreliert sein, auch wenn die Tests Unterschiedliches erfassen. Eine noch so hohe Korrelation stellt hier im besten Fall eine notwendige, aber keine hinreichende Bedingung für Konstruktidentität dar. Zweitens, und noch grundlegender, besteht der Irrtum hierbei, dass eine sehr hohe konvergente Kriteriumskorrelation häufig implizit mit Konstruktvalidität gleichgesetzt wird, wie dies z. B. Schmidt & Hunter (1999, S. 190) tun: „...they are really the same construct under two different labels". Borsboom, Mellenbergh & van Herden (2004) beschreiben diesen Fehlschluss sehr treffend, wenn sie schreiben:

> Correlations are not enough, no matter what their size. Height and weight correlate about .80 in the general population, but this does not mean that the process of letting

> people stand on a scale and reading off their weight gives one valid measurements of their height. To state otherwise is to abuse both the concept of measurement and of validity. The very fact that a correlational view of measurement allows for this kind of language abuse must be considered a fundamental weakness; any theory of validity that sustains such absurdities should immediately be dropped from consideration. Therefore, not just criterion validity but any correlational conception of validity is hopeless. (S. 1067)

Nun könnte man einwenden, bivariate Korrelationen zwischen Testergebnissen seien kein methodisch adäquates Mittel, um zu übereinstimmenden und somit generalisierbaren Aussagen zu gelangen, vielmehr müsse man Strukturen explorierende Verfahren wie die Faktorenanalyse oder Hauptkomponentenanalyse heranziehen, um zugrunde liegende Dimensionen zu identifizieren, welche möglicherweise die Ursache für alle Testinterkorrelationen sind. Ganz in diesem Sinne argumentiert Jensen (1998), wenn er die Ähnlichkeit von Faktorladungsmatrizen zwischen verschiedenen Stichproben über den Kongruenzkoeffizienten bestimmt. Hierbei gelangt er zu dem Schluss, dass die jeweils extrahierten ersten Hauptkomponenten identisch sind und ein g-Faktor allen Testleistungen zugrunde liegt: „Congruence coefficients (a measure of factor similarity) are typically above 0.95, indicating virtually identical factors, usually with the highest congruence for the g factor" (Jensen, 1998, S. 363). Es gibt zwei Gründe, welche Jensens Interpretation widerlegen. Zunächst ist der von Jensen definierte g-Faktor nicht identisch mit demjenigen von Spearman (Spearman, 1904, 1927). Für Jensen ist dieser übereinstimmend mit der ersten Hauptkomponente, wenn er schreibt: „The first principal factor is the largest common factor, and the tests that are most highly loaded on it come closest to representing the phenomenon of most central interest in the study of human abilities, namely, the common factor, which Spearman labeled g (for 'general'), that accounts for the positive manifold in the domain of ability tests" (Jensen, 1983, S. 314). Allerdings wird in der Haupkomponentenanalyse die erste Komponente aus einer Korrelationsmatrix nach dem Least-Squares-Ansatz *stets* so extrahiert, dass sie maximale Varianz aufklärt (und somit die Ladungen der Tests auf dieser Komponente am höchsten sind). Nebenbei erwähnt ist dies eine Notwendigkeit positiver symmetrischer Matrizen, die bereits von Perron 1907 entdeckt worden war (Perron, 1907). Nach der Definition von Spearman (1904, 1927) allerdings wäre der g-Faktor erst existent, wenn die Matrix der

Partialkorrelationen nach statistischer Entfernung von *g* alle null wären. Dies sollte sich zudem für alle möglichen Intelligenztestbatterien zeigen lassen. Allerdings würde dies als striktes Kriterium erfordern, dass der Rang der Korrelationsmatrix als maximale Anzahl voneinander unabhängiger Spalten gleich Eins sein müsste. Der Rang der Korrelationsmatrix bestimmt somit die Dimension der Fasern eines Testvektorbündels. Wie man unschwer einsieht, ist dies aber ein ganz anderes Kriterium als eine maximale Varianzaufklärung durch die erste Hauptkomponente, welche sich zwangsläufig und somit *nicht mehr falsifizierbar* aus der Methode der Hauptkomponentenanalyse ergibt. Was Jensen mit seiner Definition schuf, ist daher bereits definitorisch von Spearmans g-Faktor zu unterscheiden.

Allerdings könnte noch der außerordentlich hohe Kongruenzkoeffizient zwischen den ersten Hauptkomponenten als Beleg einer allen Testleistungen zugrunde liegenden Dimension angesehen werden. Dessen Höhe erweist sich tatsächlich als typisch, wenn man die Verteilung des Kongruenzkoeffizienten positiver Zufallsmatrizen betrachtet, allerdings in einem ganz anderen Sinn als Jensen (1998, S. 363) dies vermutete. Hierzu zeigen die Simulationsstudien von Schönemann (1997a), dass die Kongruenzkoeffizienten zwischen 0.995 und 0.999 variieren, wenn eine Gleichverteilung als Mutterverteilung für sie unterlegt wurde und zwischen 0.949 und 0.998 bei einer Chi-Quadrat-Verteilung als Mutterverteilung. „This means that Jensen's *g* ersatz is simply the average test score of whatever tests he analyses. It is not *g* but a travesty of Spearman's *g* (Schönemann, 2005, S. 200). Spearmans Kritik: „In truth, 'intelligence' has so many meanings that finally it has none" (Spearman, 1927, S. 15) scheint also auch heute noch aktuell zu sein, weil selbst durch ausgefeilte korrelationsstatistische Methoden nicht mehr an Definitionsklarheit zu gewinnen war, um zu invarianten und dadurch generalisierbaren Maßen zu gelangen. Es ist daher auch völlig sinnfrei z. B. zu behaupten, „Intelligence, like electricity, is easier to measure than to define it" (Jensen, 1969, S. 5) und darauf zu verweisen, dass auch die Physik eine erfolgreiche quantitative Wissenschaft wurde, bevor irgendjemand Einsicht in die hierzu notwendigen Messstrukturen hatte (s. in diesem Sinne auch Jensen, 1998). Erstens konnte sich die Physik stets auf fundamentale, d.h. nicht aus anderen Messungen ableitbare Messungen beziehen und *daraus* später entdecken, warum die damit verbundenen numerischen Operationen überhaupt so erfolgreich waren und zu exakten Konzepten führten. Gleiches kann aber für die Psychologie nicht behauptet werden, da keine fundamentalen Messungen verfügbar sind, aus denen sich auch abgeleitete begründen ließen. Zweitens kann man schlicht fragen, warum man nun, wo die Bedingungen für quantitative Messungen bekannt sind, diese nicht nutzt, um die als quantitativ angenommenen Konzepte diesbezüglich zu überprüfen? „The history of science teaches us many things, but I do not

think that one of them is that we can expect to make progress by ignoring pertinent matters" (Michell, 1999, S. 217).

Und gerade Stevens Messkonzeption stellte keine Lösung dieses Problems dar. Im Gegenteil: durch den direkten Bezug zum Operationalismus, der, grob gesagt, „*etwas* zu wissen" mit „dieses *Etwas* zu wissen" verwechselte, ermöglichte er es, den Inhalt von Konzepten post hoc in statistischen Kennwerten zu suchen, anstatt a priori falsifizierbare Hypothesen über die Struktur, die „Anatomie" eines Konstruktes zu bilden und nicht zu versuchen, dessen „Gestalt" aus Korrelationsmatrizen zu erahnen. Es ist also eben jene Misskonzeption einer Messung, die durch ihre lose Messdefinition in der Folge dazu führte, dass man in der Psychologie ignorierte, „...*that concepts must express our understanding of observations and that observations cannot serve to understand ill-defined concepts*" (Roskam, 1989, S. 240-241).

Es ist erstaunlich und vor dem Hintergrund heute noch gängiger Stevensscher Messkonzeption in der Psychologie zugleich ernüchternd zu entdecken, dass bereits vor Stevens grundlegende Abhandlungen (1951, 1959) darauf verwiesen wurde, „...dass der Messvorgang selbst Bestandteil derjenigen Theorie ist, welche man aufgrund der Messungen erst zu finden oder erhärten hoffte" (Fischer, 1974, S. 128). Sie beziehen sich auf die bereits auf S. 95 angedeuteten Messaxiome der Verknüpfungsrelation am Beispiel des Vergleichs von Gewichten auf einer Balkenwaage. Campbell (1920, 1921/1952, 1940) stellte für diese Repräsentierbarkeit physikalischer Quantität durch numerische drei empirisch überprüfbare Gesetze zur Ordnungsrelation, Additivität und zur Festlegung einer Maßeinheit auf, welche allerdings hier ihrer Länge halber nicht beschrieben werden können (für eine sehr gute Einführung hierzu s. van der Linden, 1994). Die wesentliche Aussage allerdings ist, dass eine Variable erst dann als quantitativ angesehen werden kann, wenn diese Gesetze empirisch bestätigt werden können. Variablen dieser Art sind fundamental, d.h. direkt messbar wie z. B. Masse, Volumen oder Länge. Diese Größen werden ohne Zuhilfenahme anderer Messungen direkt gemessen. Andere Variablen können nach Campbell (1920) aus diesen Messungen *abgeleitet* werden (abgeleitete Messungen). So ist z. B. Dichte = Masse/Volumen. Die abgeleitete Größe wird also als Funktion fundamentaler Größen ausgedrückt. Daher setzt jede abgeleitete Messung bereits die Gültigkeit der Gesetze fundamentaler Messung voraus! Beispielsweise stellt die Messung der Intelligenz mit einem Test das prominenteste Beispiel einer abgeleiteten Messung in der Psychologie dar. Weil nun aber eine abgeleitete Messung immer auf einer fundamentalen basiert und keine Gesetze der Relationen zwischen

fundamental messbaren Variablen der Intelligenz vorhanden sind, wäre Intelligenz nach der Definition von Campbell nicht messbar: die additive Struktur der abgeleiteten Messung lässt sich nicht aus einer fundamentalen begründen. Intelligenztests stellen hiernach vielmehr standardisierte Experimente zum Sammeln qualitativer Daten von Antworten auf Problemaufgaben dar. Das sog. Ferguson-Komitee (Ferguson et al., 1940) zur Klärung der Frage nach der Überprüfbarkeit von „quantitative estimates of sensory events" kam daher (wenn auch zerstritten) zu dem Schluss, dass eine Messung im klassischen Sinne in der Psychologie nicht möglich sei.

Genau aus diesem Grund stellte Stevens Messkonzeption den Versuch eines Auswegs aus dem Dilemma dar, allerdings mit den bekannten Implikationen und der Folge, dass „Mainstream quantitative psychology is now in the anomalous position of being unable to consider the measurability thesis in the crucial manner charateristic of normal science" (Michell, 1999, S. 191).
Dabei verfügt die moderne Messtheorie seit Luce & Tukeys grundlegendem Beitrag zum sog. *simultaneous conjoint measurement* (Luce & Tukey, 1964) längst über die Grundlage für die Überprüfbarkeit der Additivitätsannahme in der Psychologie. Die basale Idee hierbei besteht darin, dass eine fundamentale Messung über die Relation der Additivität *zwischen* Messobjekten erzielt und geprüft werden kann im Gegensatz zum oben beschriebenen klassischen Konzept der Additivität *innerhalb* von Messobjekten, bei dem sich die Verknüpfungsrelation nur auf ein einziges Attribut bezieht. Aus diesem Grund wird in der Literatur auch häufig die Bezeichnung *Additive conjoint measurement* verwendet (Perline, Wright & Wainer, 1979). Luce & Tukey (1964) zeigten nun, dass quantitative Messungen auch dann möglich sind, wenn mehrere Variablen gleichzeitig gemessen werden. Wenn man dies nun auf die Testpsychologie bezieht, handelt es sich um die gleichzeitige Messung von zwei unabhängigen Variablen, die auf eine dritte abhängige bezogen werden: Personen (A) und Itemschwierigkeit (B) bezüglich der abhängigen Variablen Itemantwort (P). (Um einem Missverständnis vorzubeugen: zwar wird bei *Anwendung* eines Tests lediglich die Personenfähigkeit gemessen, die hier interessierende simultane Messung findet jedoch zuvor als Teil des Konstruktionsprozesses des Messinstrumentes statt). Am anschaulichsten lässt sich das Zusammenwirken der Variablen in einer bivariaten Kontingenztabelle darstellen, in welcher die Zeilen die Stufen von A, die Spalten diejenigen von B darstellen und die Zelleinträge die Itemantworten (0 oder 1) darstellen, wobei die gesamte Datenmatrix nach aufsteigenden Stufen von A und B geordnet ist. Additiv verbundene Messungen würden sich nun ergeben,

wenn man Funktionen zu den tabellierten Daten finden kann, so dass das folgende additive Modell gilt (Gleichung (1)) :

$$f_1(P) = f_2(A) + f_3(B) \quad . \qquad (1)$$

Dies hieße nichts anderes, als dass sich die Leistung zusammensetzt aus Zeileneffekt + Spalteneffekt. In der Testpsychologie stellt sich die Frage nach der Möglichkeit einer solchen Abbildung der Leistung immer dann, wenn unterstellt wird, dass
Leistung = Fähigkeit - Itemschwierigkeit. Also dort, wo ein Summenwert als Anzahl gelöster Items gebildet wird und als Maß einer einzigen Leistungsfähigkeitsdimension eines Probanden herangezogen wird. Diese übliche Verrechnungspraxis der Klassischen Testtheorie impliziert also bereits das Vorhandensein dieser Form der Additivität und somit der Zulässigkeit der Summenwertbildung!
Luce und Tukey stellten nun empirisch überprüfbare Axiome auf, welche die Grundfrage des Additive conjoint measurement ansprechen. Hierbei handelt es sich um die Frage, ob die gegebenen Beobachtungen unter Beibehaltung ihrer Rangordnung in der Datenmatrix so transformiert werden können, dass sich eine additive Struktur ergibt. Bei Erfüllung dieser Axiome kann gezeigt werden, dass

a) monotone Funktionen $f_1(.)$, $f_2(.)$ und f_3 existieren, welche dem Modell unter (1) genügen und somit eine additive Repräsentation darstellen und
b) $f_1(P)$, $f_2(A)$ und $f_3(B)$ quantitative Variablen darstellen, also dieselbe Maßeinheit aufweisen.

Bei den Axiomen handelt es sich zum einen um die Transitivität von Differenzen der Beobachtungen, welche die Notwendigkeit einer additiven Struktur darstellt, zum anderen um die Lösbarkeit, welche verlangt, dass sich beide Faktoren A und B stetig verändern lassen, sodass sich eine Variation in einem Faktor durch eine entsprechende Variation im anderen kompensieren lässt. Dies stellt einen notwendigen Beweis für die Intervallskaleneigenschaft von A und B dar.
Aus Gründen der Stringenz muss für eine Herleitung der Axiome auf die Originalquelle (Luce & Tukey, 1964) und Sekundärliteratur verwiesen werden (z. B. Fischer, 1974, S. 122f.).
Aber was ist nun an diesem Ansatz das Besondere gegenüber der üblichen Messkonzeption von Stevens (1951, 1959)? Das Problem der Stevensschen Messung besteht wie ausgeführt in

der Abkehr vom klassischen Konzept der Messung hin zu einer operationalen Definition mit der Folge, dass generalisierbare Aussagen nicht mehr in dem Ausmaß möglich sind, wie es für wissenschaftliche Zwecke nötig wäre. Doch worauf fußen überhaupt generalisierbare Aussagen von so genannten objektiven Messungen? Die Grundlage hierzu stellt die Betrachtung von *Vergleichen* zwischen verschiedenen Objekten bezüglich bestimmter Eigenschaften dar. Und zwar ganz gleich, ob es sich um solche physikalischer Objekte hinsichtlich Maße wie Länge, Masse, Dichte oder eben psychologischer wie Leistungsfähigkeit oder Einstellungen von Personen handelt. Diese Vergleiche müssen in dem Sinne *objektiv* sein, als dass ihre Invarianz gegenüber einem Wechsel der Messobjekte und Messinstrumente gewährleistet sein muss, um zu kontextunabhängigen Aussagen über Eigenschaften zu gelangen. An dieser Stelle wird also klar, dass sich der Kreis zu den eingangs dargestellten Zitaten von Thurstone zu schließen beginnt: das Konzept fundamentaler Messungen der Physik kann mit dem noch grundlegenderem invarianter Vergleiche erklärt werden und überhaupt die Grundlage für den Messbegriff darstellen. Wo bei fundamentalen Messungen die Verknüpfungsrelation den Kernbegriff der Messung darstellte, ist dies beim simultaneous conjoint measurement derjenige einer festen, invarianten Maßeinheit über den gesamten Bereich der Variable. Aufgrund einer additiven Beziehung der Variablen zueinander können dadurch Vergleiche angestellt werden, indem die mit den Messobjekten verbundenen Zahlen subtrahiert werden und diese Differenz die gleiche Bedeutung über das Kontinuum behält. Ein weiteres Kernmerkmal der additiven Struktur ist, dass Vergleiche von Messobjekten innerhalb einer Klasse unabhängig von den Vergleichen in der anderen Klasse durchgeführt werden können. Unabhängigkeit bedeutet somit, dass die Ordnung der Stufen von Faktor A (Personen), welche ja durch die Itemantworten *P* bedingt wird, unabhängig davon sind, welche spezifischen Werte *B* (Itemschwierigkeiten) annimmt. Umgekehrt muss dies natürlich auch für die Ordnung der Stufen von B (Items) gelten. Würde das Additive conjoint measurement nun bspw. für einen psychologischen Leistungstest wie in Gleichung (1) gelten, so wäre es möglich, die Fähigkeit zweier beliebiger Personen (A_n und A_m) mithilfe der Itemschwierigkeit (B_t) zu vergleichen, sodass sich diese aufhebt. Dadurch wird der Vergleich zwischen den Personen unabhängig von den verwendeten Items möglich, wie folgendes Rechenbeispiel zeigt:

$$\begin{array}{rl} & P_{nt} = A_n - B_t \\ - & P_{mt} = A_m - B_t \\ \hline & P_{nt} - P_{mt} = A_n - A_m \end{array}$$

Tätigt man Aussagen wie etwa „Person A_n ist besser in abstrakter Intelligenz als Person A_m“, dann impliziert dies bereits die Unabhängigkeit des Ergebnisses von der Itemauswahl, weil man *jegliches* Problem abstrakter Intelligenz meint, mit dem Person A_n und A_m konfrontiert werden könnten. Überdies würde das Additive conjoint measurement symmetrisch zum obigen Rechenbeispiel auch zu Vergleichen der Schwierigkeit zweier Items (B_t und B_i) führen, die unabhängig von der Fähigkeitsausprägung der Personen feststellbar sind:

$$\begin{array}{rl} & B_t = A_n - P_{nt} \\ - & B_i = A_n - P_{ni} \\ \hline & B_t - B_i = P_{nt} - P_{ni} \end{array}$$

Um hierzu eine Analogie zu geben: die Schwierigkeitsdifferenz zweier unterschiedlich hoher Sprunglatten beim Hochsprung sollte stets die gleich bleiben, ganz unabhängig davon, welche Stichprobe von Hochspringern die Höhen bewältigen muss. Sollte sich allerdings die Schwierigkeitsdifferenz etwa für leistungsfähigere Springer im Vergleich zu weniger Leistungsfähigen umkehren, wäre die Messung als solche korrumpiert. Es ergäbe sich hieraus nämlich eine Wechselwirkung von Leistungsfähigkeitsausprägung und Itemschwierigkeit, welche auch *inhaltlich* schwer erklärbar wäre. Letzteres hervorzuheben ist deshalb wichtig, um klarzustellen, dass das Additive conjoint measurement keine Messtheorie darstellt, welche die „Mechanik“ der Messung über den Inhalt stellt, sondern vielmehr darauf verweist, dass man formale Messtheorie und empirische Interpretation *simultan* miteinander verknüpfen muss, wenn sich hieraus eine bedeutsame inhaltliche Interpretation bezüglich eines psychologischen Merkmals ergeben soll. Man kann deshalb nicht darauf vertrauen, dass sich ein Sinn von im Vorhinein schlecht definierten Konzepten aus den Beobachtungen ergibt. Vielmehr würde Messen also zunächst das Erstellen von eindimensionalen Strukturen voraussetzen. Messen also nicht als Selbstzweck, sondern ganz im Campbell’schen Sinne als Mittel zum Zweck zu betrachten, um inhaltliche Gesetzmäßigkeiten aufstellen zu können.
Die implizierte Symmetrie also von Invarianzeigenschaften seitens der Messobjekte (Personen) und Agentien (Items) im Additive conjoint measurement entspricht exakt den eingangs in diesem Kapitel zitierten Forderungen Thurstones, um zu wissenschaftlichen Aussagen zu gelangen: „...it should be possible to omit several test questions at different levels of the scale without affecting the individual score“ (Thurstone, 1926, S. 446), bzw.: “Within the range of objects for which the measuring instrument is intended, its function must be independent of the object of measurement” (Thurstone, 1928, S. 446). Man kann daher

zusammenfassen: Die Invarianzeigenschaften von Vergleichen, die additive Struktur und die Existenz einer festen Maßeinheit über das Kontinuum sind Eigenschaften, welche Messungen auch in der Psychologie aufweisen müssen, um überhaupt als solche bezeichnet werden zu können. Genügen Sie diesen Bedingungen nicht, dann kann das zugrunde liegende Merkmal auch nicht als quantitativ bezeichnet werden. „Die Erkenntnis eines allgemeinen Gesetzes muss der Messung vorausgehen. Wird diese Forderung nicht erfüllt, dann handelt es sich nicht um Messung, sondern um Schein-Quantifizierung, um Benennung von Ereignissen mit Zahlen anstelle beliebiger anderer Namen“ (Fischer, 1974, S. 23).

9.2 Das Rasch-Modell als probabilistische Formulierung des Additive Conjoint Measurement

Man kann nun zu recht kritisieren, ob es unter diesen Voraussetzungen überhaupt möglich ist, ein axiomatisches Messsystem auf messfehlerbehaftete Daten der Sozialwissenschaften übertragen zu können. Ein deterministisch formuliertes Axiomensystem wäre ja bereits durch eine einzige abweichende Beobachtung widerlegt. Allerdings stellt diese deterministische Datenstruktur vielmehr das Fundament für die Konstruktion einer (realistischeren) probabilistischen Messstruktur des Additive conjoint measurement dar. Ersetzt man nach diesem Prinzip die Beobachtungen der auf S. 105f. dargestellten Datenmatrix durch Wahrscheinlichkeiten, geht das Additive conjoint measurement in ein probabilistisches Messmodell über. Mit diesem Problem beschäftigte sich Rasch (1960) bei der Ableitung eines Antwortmodells für psychologische Tests, allerdings ohne direkten Bezug zur „Vorläuferliteratur“ des Additive conjoint measurement (Adams & Fagot, 1959), sondern über die Frage nach Möglichkeiten invarianter Vergleiche in der Testdiagnostik. In diesem Zusammenhang formulierte er die notwendige Bedingung, dass nur zwei Parameter benötigt werden, um die Lösungswahrscheinlichkeit bezüglich eines Items zu erklären: ein Personenfähigkeitsparameter θ und ein Itemschwierigkeitsparameter σ. Diese Überlegungen führten schließlich zum wohlbekannten Antwortmodell nach Gleichung (2):

$$P\left(X_{vi}=1\right)=\frac{\exp(\theta_v-\sigma_i)}{1+\exp(\theta_v-\sigma_i)} \qquad (2)$$

Die Gleichung besagt, dass die Lösungswahrscheinlichkeit eines Items alleine durch die Differenz der Personenfähigkeit mit der Itemschwierigkeit bestimmbar ist. Die Item-

schwierigkeit ist hierbei definiert als der Abszissenwert der 50%igen Lösungswahrscheinlichkeit einer logistischen Antwortfunktion.
Schreibt man die Modellgleichung in die logistische Form um, wird der Bezug zur Grundgleichung des Additive conjoint measurement (s. Gleichung (1)) noch offensichtlicher (Gleichung (3)):

$$\log\mathrm{it}\,P\left(X_{vi}=1\right)=\theta_v-\sigma_i \qquad (3)$$

Die zwei Messmengen Personenfähigkeit und Itemschwierigkeit sind also additiv miteinander verknüpft.
Aber welche Eigenschaft des Rasch-Modells verbindet dieses nun genau mit dem Additive conjoint measurement? Das Verbindungsglied hierbei stellt das Prinzip spezifischer Objektivität der Vergleiche dar: Vergleiche zwischen Personenparametern müssen unabhängig von den Werten der zum Vergleich herangezogenen Itemparameter sein und vice versa (sog. Separierbarkeit der Parameter). Der Leser möge sich in diesem Zusammenhang an die auf S. 107 gegebenen Rechenbeispiele im Zusammenhang mit dem additive conjoint measurement erinnern, welche dieses Prinzip verdeutlichten. Rasch (1977) konnte nun zeigen, dass die Überschneidungsfreiheit der durch Gleichung (3) gegebenen logistischen Antwortfunktionen (Itemcharacteristic Curves, ICCs) eine notwendige und hinreichende Bedingung für spezifisch objektive Vergleiche darstellt und zur Intervallskalenqualität der Logit-Skala führt (für letzteres s. Rost, 2004). Die Überschneidungsfreiheit bedeutet nichts anderes, als dass die Steigung der ICCs als Maß der Trennschärfe für alle Items gleich Eins gesetzt wird. Inhaltlich gesprochen trägt daher jedes Item gleich viel zur Differenzierung bezüglich der zu untersuchenden latenten Variable bei bzw. stellt einen gleich guten Indikator dar. Es ist in diesem Zusammenhang unerlässlich, darauf hinzuweisen, dass Items mit unterschiedlichen Trennschärfen keine spezifisch objektiven Vergleiche leisten. Das probabilistische Modell mit einander überschneidenden ICCs ist, wie eingangs beschrieben, das Birnbaum-Modell (Birnbaum, 1968), welches neben einem Schwierigkeitsparameter noch einen itemspezifischen Trennschärfeparameter (β_i) beinhaltet, wie Gleichung (4) zeigt:

$$P\left(X_{vi}\right)=\frac{\exp\left(\beta_i(\theta_v-\sigma_i)\right)}{1+\exp\left(\beta_i(\theta_v-\sigma_i)\right)} \qquad (4)$$

Als Folge können die Vergleiche zwischen Personen und Items aber nicht mehr spezifisch objektiv sein, da das Ergebnis des Vergleichs z. B. zwischen Personen nunmehr von der Itemauswahl abhängt, wie folgendes Rechenbeispiel zeigt:

$$\begin{array}{rl} & P_{nt} = \beta_i(A_n - B_t) \\ - & P_{mt} = \beta_i(A_m - B_t) \\ \hline P_{nt} - P_{mt} = & \beta_i(A_n - A_m) \end{array}$$

Es ist unmöglich, den Einfluss eines Items aus dem Vergleich zwischen zwei Personen zu eliminieren, sodass die Differenz zwischen den Fähigkeitsparametern der Personen von der Itemauswahl abhängig bleibt. Dem Leser dürfte in diesem Zusammenhang nun die konzeptuelle Nähe des Birnbaum-Modells zur Stevensschen Messkonzeption auffallen: Vergleichende Aussagen über Personen oder Items sind vollkommen abhängig von den operationalen Definitionen!
Aber welche inhaltlichen Implikationen ergeben sich aus konstanten bzw. variierenden Trennschärfen bezüglich der Interpretation der latenten Variable? Dies soll am folgenden Beispiel eines Wortkenntnistests für Kleinkinder veranschaulicht werden (nach Wright, 1997). Abbildung 8 zeigt das Ergebnis einer Rasch-konformen Kalibrierung der beispielhaft ausgewählten Wörter „away“ (i = 1) und „equestrian“ (i = 2).

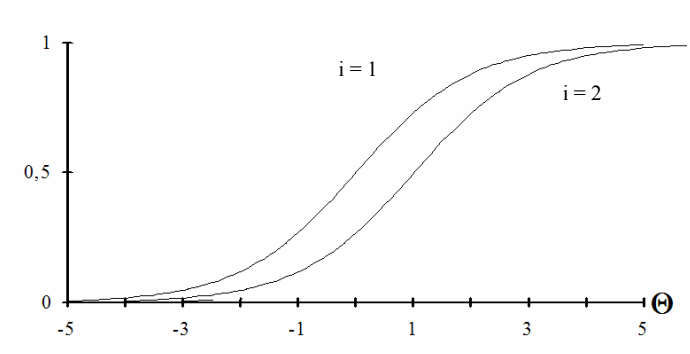


Abbildung 8: Itemcharakteristische Funktionen zweier Rasch-homogener Items

Wenn das Rasch-Modell gilt, macht es bezüglich der Lösungsrangfolge der Wörter keinen Unterschied, von welcher Stufe der latenten Variablen man diese betrachtet, stets gilt i = 2 < i = 1. Die Schwierigkeitsrangfolge bleibt somit über den gesamten Bereich des latenten Kontinuums gleich. Und dadurch stellt auch der Summenscore eine erschöpfende Statistik der Personenfähigkeit dar, da das Ergebnis des Vergleichs unabhängig von der spezifischen

Itemauswahl ist: Fähigere Personen lösen immer mehr Aufgaben als weniger fähige. Es ist egal, welche Person welche Aufgaben gelöst hat, alleine die Summe der Lösungen reicht aus, um die Information über den Leistungsgrad der Person auszuschöpfen (für einen mathematischen Beweis s. Fischer, 1974, S. 193f.; Rost, 2004, S. 122f.). Die Annahme, dass die Anzahl gelöster Aufgaben eine erschöpfende Statistik für die Personenfähigkeit sein soll, führt also *notwendigerweise* zum Rasch-Modell!
Abbildung 9 veranschaulicht nun, was passieren würde, wenn sich unterschiedliche Trennschärfen ergeben hätten.

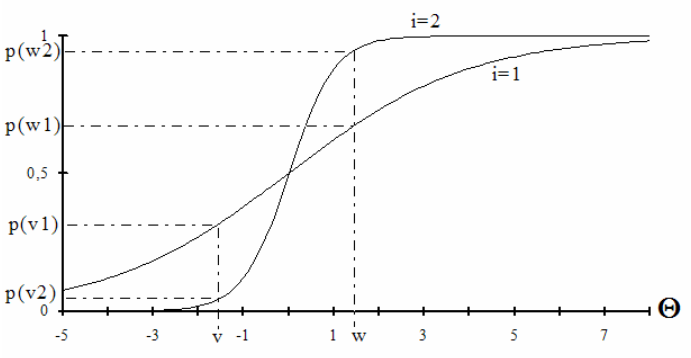


Abbildung 9: Itemcharakteristische Funktionen zweier Items mit unterschiedlichen Trennschärfen

Zunächst ist die Summe der Lösungen keine erschöpfende Statistik der Personenfähigkeit, da nicht mehr alle Items gleich viel zur Information über die Personenfähigkeit beitragen. Hier müssen also die einzelnen Itemantworten jwls. mit den Trennschärfen der Items gewichtet werden, um zu einem Maß der Personenfähigkeit aggregiert werden zu können.
Es ist weiterhin offensichtlich, dass durch sich überschneidende ICCs die Stabilität und damit die *sinnvolle* Interpretierbarkeit des Konstruktes leidet: Für Person v ist nun „equestrian“ leichter als „away“, für Person w hingegen „away“ leichter als „equestrian“. Man kann nun kritisch fragen: Was ist denn nun die Kriteriumsdefinition dieser Variable? Welches Konstrukt ist überhaupt definiert bzw. ist damit überhaupt eines bestimmt? Die Definition der zugrunde liegenden Variable verändert sich, weil sich die Bedeutung der Items verändert, je nachdem wo Personen auf der zugrunde liegenden Variablen lokalisiert sind. Warum, so kann man Fragen, ist für die fähigere Person w die Schwierigkeitsrangfolge genau umgekehrt zu derjenigen der weniger fähigen Person v? Wilson (2003) schreibt in diesem Zusammenhang treffend:

> ...the relationship between the items is no longer invariant for people at different locations. For a geographical map, this would be like saying that when you move from, say Rome to Naples, the latitudes of other places appear to change - Corsica moves south of Sardinia, say. This would make it impossible to construct a geographical map as we understand it - effectively, 'locations' no longer are consistently meaningful. In the same way, the idea of a construct map would no longer be tenable - each possible person location would possibly give a different ordering of the items, and hence the interpretation outlined above would not be useful. (S. 14).

Man kann auch sagen, dass genau an der Stelle, wo Datenkonstellationen durch zusätzliche Parameter wie im Birnbaum-Modell zugelassen werden, sich Inhalt und formales Messmodell trennen: „Because no effort is put into testing the observations in terms of the eidos [das abstrakte Gesetz eines Merkmals, *Anmerkung des Verf.*] and determining the extent to which a generalizable metrological concept emerges, the words and numbers describing the results remain tied to those results, and confusion is the inevitable consequence" (Fisher, 2004, S. 14). Wenn Campbell (1921/1952, S. 328) bemerkte, dass Messungen angestellt werden, „...only because we are thereby enabled to state laws about them", dann kann man schlicht fragen, welche Gesetzmäßigkeit überhaupt mit Modellen aufgestellt werden können, bei denen sich die Bedeutung der Indikatoren über die interessierende Variable verändert. Im Gegensatz dazu stellen beim Rasch-Modell solche Datenkonstellationen Anomalien dar, welche inhaltlich erklärt werden müssen und ggf. zu einer Revision des Konstruktes oder gar seiner Falsifikation führen, auf jeden Fall aber neue Einsichten erbringen können. Wie es Kuhn ausdrückt (1977):

> To the extent that measurement and quantitative technique play an especially significant role in scientific discovery, they do so precisely because, by displaying significant anomaly, they tell scientists when and where to look for a new qualitative phenomenon .... When measurement departs from theory, it is likely to yield mere numbers, and their very neutrality make them particularly sterile as a source of remedial suggestions. (S. 205)

Im Birnbaum-Modell und anderen mehrparametrischen Item-Response-Modellen hingegen werden solche Anomalien durch zusätzliche Parameter „maskiert“; zwar zugunsten einer besseren Modellpassung, doch zulasten einer sinnvollen Repräsentation des Inhalts durch das formale Messmodell: Inhalt und Zahlen korrespondieren nicht mehr in einer sinnvollen Weise miteinander.

Man mag gegen die Forderung der Konstanz der Itemtrennschärfen zwar einwenden, es gäbe in der Realität nun eben Items, welche ein Konstrukt unterschiedlich gut repräsentieren und demnach in ihrer Trennschärfe variieren müssen. Dem kann aber Folgendes entgegnet werden: zeigen Items bspw. eine geringere Trennschärfe als andere, so kann dies zum einen daran liegen, dass sie a) schlecht konstruiert sind (weshalb sie neu formuliert werden sollten) oder b) etwas anderes erfassen oder miterfassen (weshalb sie eliminiert werden sollten). Beide Fälle stellen aber stärker technische als Aspekte dar, denen auch im Rahmen des Birnbaum-Modells über untere Grenzen „zulässiger“ Trennschärfe begegnet werden könnte. Es ist aber durchaus auch denkbar, dass c) hierbei eine oder mehrere *Facetten* der interessierenden Variable durch verschiedene Items erfasst werden. Dann ist es aber schlichtweg falsch, diese auf einer einzigen Skala abzubilden und damit ein *eindimensionales* Konstrukt zu unterstellen, welches so nicht vorhanden ist. Ein Beispiel hierzu: Internale Kontrollüberzeugung wird stets als eine eindimensionale Variable skaliert (s. z. B. Krampen, 2000). Allerdings hat sich gezeigt, dass diese sich in die Facetten internale Kontrollüberzeugung aufgrund von Anstrengung vs. hoher Fähigkeit untergliedert, welche lediglich in mittlerer Höhe korreliert sind (Heene, 2003; Krampen, 2000, S. 112). Da dadurch zwei Personen mit jeweils ganz unterschiedlichen Ausprägungen auf beiden Facetten zum selben Rohwert gelangen können, ist es schlichtweg ein Fehlschluss, ihnen das gleiche Ausmaß internaler Kontrollüberzeugung beizumessen. Man müsste daher erstens neue inhaltliche Überlegungen über die „Konstruktanatomie“ anstellen und zweitens den Fragebogen mit einem zweidimensionalen Itemkomponenten-Rasch-Modell auswerten (s. hierzu z. B. Rost & Carstensen, 2002), um zu einer validen Aussage zu gelangen. Eine Auswertung nach dem Birnbaum-Modell würde den Facettenreichtum des Konstruktes allerdings verdecken und zu einem Fehlschluss über die Personenausprägung führen. Zudem wären spezifisch objektive Vergleiche, wie sie für eine sinnvolle inhaltliche Interpretation der latenten Variable nötig ist, unmöglich.

Die Aussagen über das Birnbaum-Modell treffen allerdings analog auf die Skalenkonstruktion nach der Klassischen Testtheorie zu. Wenn moniert wird, dass beim Rasch-Modell ja wie in der Klassischen Testtheorie auch der Summenscore als Schätzer der Personenfähigkeit verwendet

wird, so ist dem entgegenzuhalten, dass das Rasch-Modell erst prüft, ob dieser tatsächlich die gesamte Information bezüglich der Personenvariable enthält. Denn, wie bereits weiter oben ausgeführt, ist das Rasch-Modell das einzige, welches den Rohwert als suffizienten Schätzer enthält! Der gesunde wissenschaftliche Verstand würde es nun gebieten, diese und weitere Annahmen mit dem Rasch-Modell falsifizierbar zu testen, was dann aber meist mit dem Verweis darauf, dass das dieses zu restriktiv sei, unterlassen wird. Die „Logik“ hierbei ist also, das Rasch-Modell explizit als zu restriktiv abzutun, es aber dann implizit anzuwenden. Wright (1977) merkt daher an:

> ...for anyone who claims skepticism about ‘the assumption’ of the Rasch model, those who use unweighted scores are, however unwittingly, counting on the Rasch model to see them through. Whether this is useful in practice is a question not for more theorizing, but for empirical study. (S. 114)

Nun könnte man einwenden, dass es ja immerhin auf persönlichkeitspsychologischer Seite faktorenanalytisch konstruierte Fragebögen gibt, wodurch sich ein Personenmesswert aus einer Linearkombination der Faktorladungen mit den Itemantworten ergibt. Doch in fast allen Fragebögen werden für eine Auswertung die Itemantworten (ungewichtet!) aufsummiert. Sucht man sodann nach der Begründung für diese Inkonsequenz, folgt diese dem folgenden Schema (s. etwa Russell, 2002): Da die Faktorladungen stark stichprobenabhängig sind und daher schwer zu generalisieren, werden diese durch die einfache Summation gleich eins gesetzt. Folglich kann man sagen, dass hierdurch wiederum ein Rasch-Modell und zwar ein verallgemeinertes für Ratingskalen angewendet wird.
Auch der Einwand, Itemselektion nach dem Rasch-Modell und den Rationalen der KTT führten zur identischen Itemauswahl, es mache also kaum einen Unterschied, wonach man auswähle, ist schon von der theoretischen Grundlage nicht korrekt. Wie bereits oben ausgeführt, entspringt das Rasch-Modell den Überlegungen zur Invarianz von Messungen, was sich durch die Möglichkeit spezifisch objektiver Vergleiche zeigt. Wenn nun aber nach dieser Logik das Ergebnis des Personenvergleichs unabhängig von der Itemauswahl sein soll, dann ist das Rasch-Modell bereits dort verworfen, wo Items aufgrund von Item-Fit-Indizes aus dem Test entfernt werden müssten, weil nun das Ergebnis doch von einer spezifischen Itemauswahl abhängt. Dieses Vorgehen stellt im Rahmen der Klassischen Testtheorie aber gar kein Problem dar, da hier ganz im Sinne der operationalen Definition von Stevens der wahre Wert einer

Person gleichgesetzt wird mit der „wahren Intelligenztestleistungsfähigkeit“, welche sie bei beliebig häufiger Testvorgabe unter identischen Bedingungen *dieses* Tests erzielen würde. Dies soll nun nicht bedeuten, dass Item-Fit-Indizes des Rasch-Modells, wie sie im Kapitel 10.1.2 eingehend dargestellt werden, aus der Testkonstruktion zu verbannen sind; sie können wertvolle Hinweise auf Anomalien in den Daten aufdecken und darüber zu neuen inhaltlichen Überlegungen führen, ob die jeweilige Konstruktdefinition sinnvoll ist. Ein solches Ergebnis würde wiederum darauf verweisen, dass Überlegungen der Konstruktdefinition und -bedeutung *vor* der Erhebung stattfinden müssen. Der Satz, dass eine Messung Sinn *ergibt*, ist daher also ganz eigentlich falsch, denn eine Messung kann nur dann Sinn ergeben, wenn vorab ein das Konstrukt hypothetisch definierende Itemuniversum spezifiziert wurde.

Überdies gibt es aber auch bei der Beurteilung der Trennschärfe einen bedeutsamen Unterschied. Zwar führen sowohl die klassische Trennschärfeberechnung als auch diejenigen nach Item-Fit-Indizes des Rasch-Modells zu ähnlichen Ergebnissen: nach der Klassischen Testtheorie zu trennschwache Items haben meistens nach den Fit-Indizes des Rasch-Modells zu flache ICCs wie auch nach der KTT trennscharfe Items steilere ICCs aufweisen (Fan, 1998; MacDonald & Paunonen, 2002). Allerdings besteht ein gravierender Unterschied dort, wo Items nach den Annahmen des Rasch-Modells *zu* trennscharf sind, d.h. zu einem deterministischen Antwortverhalten der Personen führen, wohingegen solche Items nach der KTT als besonders gut geeignet angesehen werden, da sie ein eine hohe Unterscheidungsfähigkeit zwischen Personen aufweisen. Dieser Punkt wird auch häufig in der Testkonstruktion nach dem Rasch-Modell übersehen und derartige Items, obgleich vom Rasch-Modell als unpassend identifiziert, im Test belassen, meist mit dem Hinweis, dass dies für diagnostische Zwecke wünschenswert sei (s. in diesem Sinne etwa Rost, Carstensen & von Davier, 1999, S. 121). Dabei wird allerdings unterschlagen, dass ein systematischer „Item-Bias“ zu einer *zu hohen* Trennschärfe nach den Modellannahmen geführt haben kann. Masters (1988, S. 17) führt hierzu beispielhaft aus: „If the content of an item has been emphasized in one instructional program but either not taught or treated only superficially in another program, then that item is likely to be differentially difficult for students in those instructional groups”. Wird das Item nun in der gesamten Stichprobe kalibriert, ergibt sich eine ungewöhnlich hohe Diskriminationsleistung, weil die Teilnahme an einem bestimmten Lehrprogramm positiv mit der in diesem Item angesprochenen Fähigkeit korreliert. Das Item misst in diesem Fall daher nicht mehr alleine die eigentliche Fähigkeit, sondern ebenso die Teilnahme an einem bestimmten Lernprogramm und ist dadurch mehrdimensional.

Darüber hinaus ist es wichtig zu ergänzen, dass zu hohe Trennschärfen auch ein Effekt von Itemredundanzen sein können (E. V. Smith, Jr., 2005). Werden Items zu ähnlich konstruiert, etwa, wenn sich in Leistungstests ein Lösungsalgorithmus wiederholt, und somit der Inhalt eines Items relevante Lösungsinformationen für eines der folgenden Items aufweist, oder in Persönlichkeitsfragebögen der Iteminhalt lediglich von Item zu Item synonymisiert wird („dasselbe in grün“ darstellt), kommt es zu Abhängigkeiten zwischen den Items, die nicht mehr alleine auf die zu messende latente Variable rückführbar sind. Das Interessante an dem Problem der Itemredundanzen ist, dass hierdurch das Verdünnungsparadoxon der KTT mit erklärt und durch das Rasch-Modell aufgelöst werden kann: werden Items durch Itemredundanzen zu ähnlich, führt dies zu einer höheren bzw. *überhöhten* Reliabilität der Skala, da die Item-Interkorrelationen ansteigen. Jedoch sinkt zugleich die Validität der Skala, weil das zugrunde liegende Konstrukt nur noch partiell, also nicht ausreichend durch die Items abgebildet wird. Man weiß also immer genauer über immer weniger des Konstruktes.
Im Rasch-Modell hingegen würden sich derartige *konstruktirrelevante* Itemabhängigkeiten in einer modellinkonformen zu steilen Steigung der itemcharakeristischen Funktion manifestieren und dadurch auf das Vorhandensein einer weiteren Dimension bei der Beantwortung der Items verweisen. Hier würden daher genau solche Items eliminiert werden, welche in der KTT zu einem Abfall der Validität bei steigender Reliabilität führten. Salopp gesprochen ist hier also ein Weniger an Trennschärfe mehr.

Abschließend lässt sich somit zusammenfassen, dass die Reduktion auf einen Itemparameter zu einem Mehr an Falsifizierbarkeit und somit Einsichtsmöglichkeiten in Strukturen psychologischer Variablen führt. Dies könnte mit dazu führen, dass das Falsifikationsprinzip in der Psychologie wieder als zielführender angesehen wird, als es mittlerweile leider der Fall ist, betrachtet man beispielsweise die Möglichkeiten zu Modellanpassungen im Rahmen von Linearen Strukturgleichungsmodellen über Modifikationsindizes. „Surprisingly, vulnerability to falsification is commonly deemed by psychologists to be a fault rather than a virtue” (Michell, 1990, S. 130). Nach Meinung des Autors scheint paradoxerweise in der Falsifizierbarkeit des Rasch-Modells seine bislang geringe Verbreitung zu liegen, weil sie die Bereitschaft erfordern würde, auch empirische Fakten anzunehmen, die entgegen den traditionellen Vorstellungen stünden.

Somit beginnt sich hier der Kreis zu den eingangs dieses Kapitels gegebenen Äußerungen zu schließen: Das Rasch-Modell stellt ein gänzlich anderes Forschungsparadigma dar. Es dient

nicht der besseren Beschreibung der Daten, sondern „...instead of evaluating models representing an eidos in terms of their capacity to describe particular observations, they evaluate particular observations in terms of their capacity to participate in the eidos represented by a model" (Fisher, 2004, S. 19). Es ist die formale Umsetzung einer Messtheorie in der Psychologie, während zwei- und mehrparametrische Antwortmodelle Methoden zur Datenauswertung sind. Eben dadurch steht es auch in deutlichem Kontrast zur operationalen Definition der Messung, indem es sich durch die Invarianzforderungen auf das repräsentationale Messkonzept bezieht. Wenn das Ergebnis des Vergleichs zwischen Personen bspw. bezüglich der allgemeinen Intelligenz unabhängig von der Itemauswahl sein soll, so müsste dieses es auch unabhängig vom verwendeten Test der allgemeinen Intelligenz sein, wenn alle Tests tatsächlich dasselbe messen! Ansonsten ließen sich keine echt generalisierbaren Aussagen über allgemeine Intelligenz ableiten. Sollten sich aber die Itemparameter von Stichprobe zu Stichprobe, etwa transkulturell oder über die Geschlechter signifikant unterscheiden, so wäre die Invarianz des Konstruktes nicht mehr gegeben, der Test letztlich mehrdimensional, da es nicht anders erklärbar wäre, dass sich die Lösungswahrscheinlichkeiten von Personen mit der gleichen Fähigkeit aber in verschiedenen Stichproben verändern. Beispielsweise ließen sich Mittelwertsunterschiede bei Invarianz eines Konstruktes über die Stichproben dann nicht mehr als Unterschied in der Ausprägung, sondern nur als Unterschied in der Art interpretieren. Nur wenn sich eine Variable in invarianter Weise identifizieren ließe, so wäre sie tatsächlich allgemein zu nennen.

Abschließend soll an dieser Stelle noch betont werden, dass die Geltung des Rasch-Modells keine hinreichende, sondern bestenfalls notwendige Bedingung für die quantitative Struktur eines Merkmals ist. Ein Theoretisieren - im positiven Sinne - über die quantitative oder ggf. qualitative Struktur des psychologischem Merkmals muss daher schon *vor* der Anwendung des Rasch-Modells stattfinden. Mit dem Rasch-Modell ist allenfalls ein Punkt markiert, von dem aus begonnen werden könnte, die Quantifizierungshypothese zu überprüfen. Denn gerade die Invarianzannahmen des Rasch-Modells erfordern starke analytische Konstruktdefinitionen. Wird das Rasch-Modell als einziges Antwortmodell, welches eine Lösung des Problems verallgemeinerbarer Aussagen in der Psychologie darstellt, falsifiziert, so wäre dies nicht weniger als „Discovering the problem after it is solved" (Andrich, 2004, S. 9). Die Frage wäre dann auch ganz grundsätzlicher Natur: Wenn die Überlegungen, zu generalisierbaren Aussagen in der Psychologie zu gelangen, *notwendigerweise* zum Rasch-Modell führen, dann müsste man ebenso die Frage stellen, ob man kognitive Modelle entwickeln kann, die notwendigerweise zu diesem führen. Dieses Problem bestünde sodann in einer Reformulierung

inhaltlicher Theorien, womöglich auch in einer von Latent-Trait-Modellen überhaupt. Vielleicht, so kann man spekulieren, würde sich das Rasch-Modell auch nur als der Traum herausstellen, den die Psychometrie von der Verallgemeinerbarkeit ihrer Aussagen träumt (zu einer grundsätzlichen Auseinandersetzung mit dieser Problematik s. z. B. Schönemann, 1994). Doch auch in jedem Falle wäre die grundlegende Falsifizierbarkeit des Rasch-Modells ein Rahmen für die Überprüfbarkeit generalisierbarer Aussagen innerhalb der Psychologie generell, denn wie Cronbach (1982) es ausdrückt:

> ...the sooner all social scientists are aware that data never speak for themselves, that without a carefully framed statement of boundary conditions generalizations are misleading or trivially vague, and that forecasts depend on substantive conjectures, the sooner will social science be consistently a source of enlightenment. (S. 71)

Aus diesen Ausführungen sollte nun klar geworden sein, weshalb in dieser Arbeit die Skalenanalyse nach dem Rasch-Modell durchgeführt wird. Die messtheoretisch bessere Fundierung dürfte keiner weiteren Erläuterung bedürfen. Aber gerade hieraus ergeben sich praktisch relevante Implikationen. Gerade wenn im Rahmen von High-Stakes-Testungen Aussagen über Personenfähigkeiten getätigt werden sollen, so müssen diese strengeren Objektivitätskriterien genügen. Die strikten Invarianzforderungen des Rasch-Modells liefern hierzu einen adäquaten Bezugsrahmen: die Unabhängigkeit des Personenvergleiches von der Itemauswahl ist sowohl eine Erfordernis eindimensionaler Skalen und objektiver Vergleiche als auch ein Gebot der Testfairness. Wäre die Zulassung zu einem Studienplatz nämlich davon abhängig, welche Person welche Items vorgelegt bekommen hat, so wäre der Test als Entscheidungsgrundlage sehr zweifelhaft. Ebenso wäre die Forderung der Invarianz der Itemparameter von der Stichprobenauswahl gerade auch im Hinblick auf Testvorbereitungseffekte von enormer Bedeutung. Wenn sich etwa spezifische Items eines Tests als deutlich trainierbar erwiesen, oder gegenüber Substichprobenaufteilungen (etwa nach Geschlecht) nicht invariant zeigten, dann wären auch hier die objektiven Messeigenschaften des Testverfahrens zu hinterfragen.

10. Rationale der Testanalyse nach Rasch-Modellen

Ziel dieses Kapitels soll es sein, die Rationale und Hintergründe der Testanalyse nach Rasch-Modellen in solchen Details darzulegen, die für das Verständnis der nachfolgenden Ergebnisse bedeutsam sind. Zentral werden daher die Rationale der in dieser Arbeit verwendeten verschiedenen Modellgeltungstests (s. z. B. Rost, 2004) behandelt.

10.1 Modellgeltung im Rahmen der Rasch-Modelle

Da die Parameter eines Testmodells auch unabhängig von der Modellgeltung geschätzt werden können, ihre Interpretation, bzw. ihre Bedeutsamkeit jedoch nur bei Vorliegen einer Modellpassung gegeben ist, müssen geeignete Kennwerte für einzelne Modellannahmen herangezogen werden, die eine Beurteilung der Modellgüte ermöglichen. Allerdings kann die Modellgeltung nach Schätzung der Parameter nicht im Sinne einer Wahr-Oder-Falsch-Aussage ausfallen. Zum einen deshalb nicht, weil jedes probabilistische Testmodell von Natur aus nur mit einer bestimmten Wahrscheinlichkeit passen kann, im Gegensatz etwa zu einem deterministischen Testmodell wie der Guttman-Skala (Guttman, 1950). Zum anderen aus erkenntnistheoretischer Perspektive nicht, weil es das Wesen von Modellen ist, Vereinfachungen wesentlicher, aber verallgemeinerbarer Zusammenhänge der Realität darzustellen, und sich daher nie eine perfekte Passung zwischen beiden ergeben kann. Dies hat zur Konsequenz, dass es einer willkürlichen Grenzziehung bedarf, ab wann die Geltung eines Modells angenommen werden kann. Die Art dieser Grenzziehung ist durch zwei Wege gekennzeichnet, die zwar beide die Modellgültigkeit zum Thema haben, jedoch in unterschiedlichen Zielen resultieren können: statistische und psychologische Signifikanz. Statistische Signifikanztests machen die Beantwortung der Modellgültigkeit von jeder Subjektivität (abgesehen von der Wahl des Signifikanzniveaus) unabhängig und liefern einen Wahrscheinlichkeitswert über die Modellgültigkeit. Hierbei wird allerdings die Testlogik umgekehrt und die Modellpassung in der Nullhypothese formuliert, weshalb Bortz & Lienert (2003) von der Nullhypothese als der „Wunschhypothese“ sprechen. Nachteile dieser Vorgehensweise ergeben sich erstens aus dem monotonen Zusammenhang der Teststatistik mit der Anzahl der Beobachtungen. Numerisch gleich große Modellabweichungen können somit bei geringer Stichprobengröße zu einer Annahme des Modells führen, wohingegen sie bei größerer zu einer Ablehnung führen würden. Zweitens testet ein Signifikanztest meist nur bestimmte Annahmen, also nicht die Gesamtpassung der Daten mit dem Modell und ist somit insensitiv gegenüber anderen Modellabweichungen als den spezifizierten. Drittens ergibt sich aus dem bereits oben beschriebenen Umstand der Formulierung der Modellgeltung in Form der

Nullhypothese, dass hier nicht der Alpha-Fehler die kritische Größe darstellt, sondern vielmehr der Beta-Fehler als die Wahrscheinlichkeit, ein falsches Modell anzunehmen. Die hieraus resultierende Notwendigkeit der Kontrolle des Beta-Risikos setzt allerdings voraus, dass eine maximal tolerierbare Abweichung in Form eines Effektstärkemaßes zuvor festgelegt wurde, was wiederum das Ergebnis des Signifikanztests von subjektiven Kriterien abhängig machen würde. Alternativ ließe sich die empirisch gefundene Stärke der Modellabweichung als Alternativhypothese aufstellen und der Beta-Fehler unter dieser Alternativhypothese berechnen. Derartige Post hoc Poweranalysen sind aber vom wissenschaftstheoretischen Standpunkt nicht ausreichend. Die häufig anzutreffende Praxis der Modellannahme, sobald der Wert der Teststatistik über dem üblichen Signifikanzniveau von 5% liegt, kann ebenfalls wegen des unbekannten Beta-Fehlers nicht befriedigen. Schönemann (1981) bemerkt hierzu: „To worry about Type I Error when testing models, just because this error happens to be easier to control, is like searching for a lost dime under a lamppsot, just because the light is better there“. Viertens macht der statistische Modelltest, gleich wie oder ob die vorher beschriebenen Probleme gelöst wurden, nur Aussagen über die Passung des Datensatzes mit dem spezifizierten Modell, nicht jedoch über seine Übereinstimmung mit Alternativmodellen, welche die Daten gleich gut oder besser hätten erklären können. MacCallum et al. (1993) beschreiben das hieraus resultierende Problem in Bezug auf Strukturgleichungsmodelle (wobei dies natürlich für jede Modellbildung gilt) treffend:

> Applied research journals now routinely publish articles in which researchers present covariance structure models and argue for their validity on the basis of their meaningfulness and their fit to the empirical data. However, the possibility that for any given such model there may be alternative models that fit the data as well as the original model constitutes a clear threat to the validity of convulsions drawn by researchers using these methods (S.186).

10.1.1 Psychologische Signifikanz

Der letztgenannte Kritikpunkt an Signifikanztests als objektive und wissenschaftliche Erkenntnisgrößen weist zugleich den Weg zur Konzeption einer psychologischen Signifikanz. Im Rahmen von Rasch-Modellen spiegelt sich diese Art der Signifikanz in Maßen der *Größe der Modellabweichungen* und in *Vergleichen* von einander konkurrierenden Modellen wider.

Das Methodenspektrum ist hierbei sehr umfangreich und soll an dieser Stelle nur überblicksartig dargestellt werden, jedoch detaillierter in denjenigen Bereichen beleuchtet werden, die für diese Arbeit maßgeblich sind. Es beinhaltet verschiedene Maße und Zugänge zur Reliabilitätsbestimmung (E. V. Smith, Jr., 2001; Wright & Masters, 1982; Wright & Stone, 1979), grafischer Modellgeltungskontrollen (Embretson & Reise, 2000; Moosbrugger & Hartig, 2002; Rost, 2004), Itemfit-Maße (Rost, 2004; Rost & von Davier, 1994; Wright & Masters, 1982; Wright & Stone, 1979), Indizes zur Identifizierung von den Modellannahmen abweichender Antwortmuster (von Davier & Molenaar, 2003; von Davier & Strauss, 2003) und informationstheoretischer Maße zu deskriptiven Modellvergleichen (Bozdogan, 1987; Read & Cressie, 1988; Rost, 2004). Die Übergänge von insbesondere deskriptiven Item- und Personenfit-Maßen zu inferenzstatistischen Statistiken ist allerdings in vielen Fällen durch entsprechende Transformationen möglich, sodass sich die Grenze zwischen Größe der Modellabweichungen und ihrer zufallskritischen Absicherung auflösen lässt. Generell sind diese Kennwerte nicht als echtes Alternativkonzept zur inferenzstatistischen Absicherung zu sehen, sondern vielmehr ergänzend dazu.
Zu betonen ist daher, dass in der vorliegenden Arbeit die psychologische Signifikanz nicht zuungunsten der statistischen behandelt werden soll, sondern beide verwendet werden. Die folgenden Darstellungen der in dieser Arbeit verwendeten Modellgeltungstests und -indizes sollen dies zeigen. Hierbei folgt die Darstellung der Einteilung nach Rost (1988, 2004), der Rasch-Modelltests zum einen in solche der Überprüfung der Item-, zum anderen in solche der Personenhomogenität unterteilt.

10.1.2 Überprüfung der Itemhomogenität

Unter Itemhomogenität wird allgemein die Eigenschaft eines Tests bezeichnet, dass alle Items dieselbe latente Personeneigenschaft messen. Im dichotomen Rasch-Modell wird diese Aussage dahingehend spezifiziert, dass alle Items gleich gute Indikatoren dieser latenten Variable sind, was sich in der Restriktion gleicher Trennschärfen aller Items niederschlägt und für die vorteilhaften Eigenschaften der Eindimensionalität, spezifisch objektiver Vergleiche und der Suffizienz des Summenscores bürgt (s. hierzu Kap. 9). Es stellt sich die Frage, wie das Konzept konstanter Itemtrennschärfen als Anstieg der Itemcharakteristischen Funktion in ein zur Überprüfung dieser Annahme geeignetes Maß übersetzen lässt. Den frühesten Ansatz hierzu stellt die Test-Statistik von Martin-Löf (1973, zit. nach Rost, 1988, S. 307) dar, welche die Annahme gleicher Personenparameter für disjunkte Itemgruppen über einen bedingten Likelihood-Quotiententest prüft. Diese in der Literatur unter dem Namen „Martin-Löf-Test“

bekannt gewordene Teststatistik setzt allerdings a priori Hypothesen über maximal unterschiedliche Itemgruppen voraus. Wie van den Wollenberg (1988) allerdings treffend bemerkt, sind Hypothesen über heterogene Itemgruppierungen „...as a rule not available“. Dieses Problem aufgreifend wurden in der Folgezeit weitere Tests zur Überprüfung der Itemhomogenität entwickelt, welche meist über das Konzept der lokalen stochastischen Unabhängigkeit der Itemantworten, was im Rasch-Modell übereinstimmend mit demjenigen der Eindimensionalität ist, diese Annahme überprüfen. Für einen Überblick über die mittlerweile sehr umfangreiche Literatur zu diesem Thema s. z. B. Glas und Verhelst (1995). Trotz dieser Fülle lässt sich hierbei eine grobe Unterteilung in Residuen-basierte und Likelihood-basierte Itemfit-Maße vornehmen. In der vorliegenden Arbeit wird zur Überprüfung der Itemhomogenität und Itemauswahl der Residuen-basierten Ansatz eingesetzt, welcher von der Summe der Differenzen beobachteter und durch die Modellannahmen erwarteten Itemantworten ausgeht. Verschiedene Autoren kritisieren diesen weit verbreiteten Ansatz. So zweifeln Rogers & Hattie (1987) auf den Ergebnissen einer Monte-Carlo-Simulationsstudie die Gültigkeit der von Wright und Masters (1982) getroffenen Verteilungsannahmen der Item-Fit-Statistiken an. Als Konsequenz lassen sich keine verbindlichen zufallskritischen Grenzen angeben (Ludlow & Haley, 1992; R. M. Smith, Schumacker & Bush, 1998).

Likelihood-basierte Fit-Maße hingegen gehen nicht von der Summe erwarteter und beobachteter Itemantworten aus. Denn hierbei wird die bedingte Wahrscheinlichkeit aller Antworten auf einem Item, also der Itemvektor, unter der Bedingung des Summenscores derjenigen Personen, die das betreffende Item gelöst hatten, betrachtet. Mithilfe einer Transformation in z-Werte kann sodann geprüft werden, ob ein Itemvektor signifikant vom nach Modellannahmen zu erwartenden Lösungsmuster abweicht. Da in diesem Ansatz das gesamte Lösungsmuster in Form eines Itemvektors betrachtet wird, im Gegensatz zur Summenstatistik der Residualmaße, scheint dieser Ansatz viel versprechend hinsichtlich der Identifizierung von Modellabweichungen. Wie Ostini & Nering (2005, S. 88) allerdings bemerken: „Although the response function-based approach is promising, very little research has been conducted ... Concerns remain regarding the asymptotic properties of the statistics, particulary with estimated model parameters and in the face of model violations”. Eigene Beobachtungen mit der likelihood-basierten Q_i-Statistik (Rost & von Davier, 1994) legen die Vermutung nahe, dass oftmals die statistische Power dieser Statistik unter der residual--basierten von Wright & Stone (1982, 1979) liegt. Rost (2004, S.372) verweist bezüglich der Residuen-basierten Itemfit-Maße allerdings darauf, dass die tatsächliche Itemantwort stets

ganzzahlig ist, während ihr Erwartungswert nach den Itemfit-Maßen von Wright & Stone (1982, 1979) meist zwischen zwei Antwortalternativen liegt und daher die Antwort einer Person in den seltensten Fällen mit diesem genau übereinstimmt. Als Konsequenz ergäbe sich hieraus, dass die Residuen ein modellkonformes Item zu oft als modellinkonform identifizieren würden, dieser Ansatz also zu einem inakzeptablen Alpha-Fehler tendiert. In dieser Arbeit soll diesem Problem dadurch entgegengewirkt werden, dass neben der inferenzstatistischen Betrachtung des Itemfits auch die Größe der Abweichung als Effektstärkemaß bei der Modellgeltung und Itemauswahl mitberücksichtigt werden soll.

10.1.2.1 Darstellung der Itemfit-Statistiken Infit und Outfit

Die Ableitung der hier verwendeten Residual-Maße nimmt ihren Anfang bei der allgemeinen Definition des Erwartungswertes für diskrete Zufallsvariablen nach Gleichung (5)

$$\mu = \sum_{i=1}^{N} X_i * p_i \qquad (5)$$

Der erwartete Rohwert *E* einer Person *n* bei Item *i* ergibt sich somit aus (5) durch

$$E_{ni} = \sum_{k=0}^{M_i} k * p_{nik}, \qquad (6)$$

wobei M_i: Gesamtrohwert auf Item *i*
p_{nik}: Lösungswahrscheinlichkeit von Person n mit Rohwert *k* auf Item *i*

Die Ableitung von Residualmaßen zur Bestimmung bedeutsamer Abweichungen von den vom Modell erwarteten Itemantworten nimmt ihren Ausgangspunkt beim beobachteten Rohwert einer Person und leitet sich wie folgt ab.

Sei X_{ni} der beobachtete Rohwert von Person n bei Item i, dann ist das Rohwert-Residuum Y_{ni} als Abweichung des beobachteten vom erwarteten Rohwert bei Item i gegeben durch

$$Y_{ni} = X_{ni} - E_{ni}. \qquad (7)$$

Das standardisierte Residuum Z_{ni} ist definiert durch

$$Z_{ni} = Y_{ni} / \sqrt{W_{ni}}, \qquad (8)$$

mit

$$W_{ni} = \sum_{k=0}^{M_i} (k - E_{ni})^2 p_{nik} \qquad (9)$$

als der Varianz von X_{ni}. Durch Mittelung der quadrierten standardisierten Residuen Z_{ni} über N Personen ergibt sich der (ungewichtete) mittlere Quadratfehler (Meansquare, MNSQ) von Item i durch:

$$u_i = \sum_{n=1}^{N} Z_{ni}^2 / N, \qquad (10)$$

mit einem Erwartungswert von eins und einer Varianz von

$$s_i^2 = \sum_{n=1}^{N} (C_{ni} / W_{ni}^2) / N^2 - 1/N, \qquad (11)$$

wobei C_{ni} die Kurtosis von X_{ni} bezeichnet, welche definiert ist als

$$C_{ni} = \sum_{k=0}^{M_i} (k - E_{ni})^4 p_{nik} \qquad (12)$$

Da u_i nach Wright & Stone (1982, 1979) annähernd χ^2-verteilt ist, diese Verteilung jedoch bekanntermaßen asymmetrisch ist, kann der ungewichtete MNSQ in eine approximativ standardnormalverteilte *t*-Statistik wie folgt umgewandelt werden

$$t_i = (u_i^{1/3} - 1)(3/s_i) + s_i/3 \qquad (13)$$

Die ungewichteten Itemfit-Statistiken u_i und t_i, die sogenannten *Outfits*, sind sensitiv bereits gegenüber wenigen Ausreißer-Antworten (sog. „Off-target-responses“), etwa durch Rate-strategien von Personen, für die das betreffende Item nach Modellannahmen zu schwer ist oder durch „Achtlosigkeit“ von Personen, die das Item aufgrund ihrer Fähigkeit nach Modell-

annahmen eigentlich mit hoher Wahrscheinlichkeit hätten lösen müssen. Um diesen Effekt zu reduzieren, leiteten Wright und Stone (1982, 1979) den folgenden mit der individuellen Varianz gewichteten MNSQ ab:

$$v_i = \sum_{n=1}^{N} W_{ni} Z_{ni}^2 / \sum_{n=1}^{N} W_{ni}, \qquad (14)$$

mit einem Erwartungswert von Eins und einer Varianz von

$$q_i^2 = \sum_{n=1}^{N} (C_{ni} - W_{ni}^2) / \sum_{n=1}^{N} W_{ni}^2. \qquad (15)$$

Der gewichtete MNSQ lässt sich wie folgt in eine approximativ standardnormalverteilte Prüfgröße umwandeln:

$$t_i' = (v_i^{1/3} - 1)(3/q_i) + q_i/3. \qquad (16)$$

Die gewichteten Itemfit-Statistiken v_i und t_i', die sog. *Infits*, sind im Gegensatz zu den ungewichteten Versionen sensitiv gegenüber Verletzungen von Antworten von Personen, in deren Fähigkeitsbereich das Item eigentlich hätte maximal diskriminieren müssen (sog. „In--target-responses"). Sie sind daher geeignet, systematische Abweichungen wie Verletzungen der Annahme lokaler stochastischer Unabhängigkeit und der Eindimensionalitätsannahme zu identifizieren. Ein Beispiel soll die Logik beider Means-Square-Fit-Statistiken verdeutlichen (nach Uekawa, 2006).
Gegeben seien die folgenden Statistiken von sechs Probanden bezüglich eines Items.

Tabelle 10: Statistiken zur Beschreibung der MNSQ-Fit-Statistiken anhand von sechs Probanden (nach Uekawa, 2006).

Pbn	X_{ni}	E_{ni}	Y_{ni}	Y^2_{ni}	W_{ni}	Y^2_{ni}/W_{ni}	Z_{ni}	Z^2_{ni}	$W_{ni}*Z^2_{ni}$
1	1	.812	0.19	0.04	.15	0.23	0.48	0.23	0.04
2	1	.995	0.01	0.00	.01	0.00	0.07	0.01	0.00
3	0	.898	-0.90	0.81	.09	8.80	-2.97	8.77	0.81
4	1	.898	0.10	0.01	.09	0.11	0.34	0.11	0.01
5	1	.984	0.02	0.00	.02	0.02	0.13	0.02	0.00
6	0	.522	-0.52	0.27	.25	1.09	-1.05	1.09	0.27
Σ					0.60				
$Outfit_i$									1.70
$Infit_i$									1.85

Anmerkung.

Pbn: Probandennummer; X_{ni}: Beobachtete Antwort von Person n auf Item i; E_{ni}: Erwartete Antwort von Person n auf Item i; Y_{ni}: Residuum von Person n auf Item i; Y^2_{ni} : quadriertes Residuum von Person n auf Item i; W_{ni}: Varianz von X_{ni}; Y^2_{ni}/W_{ni}: Quotient aus z-standardisiertem quadrierten Residdum der Person n auf Item i und der Varianz von Person n auf Item i; Z_{ni}: z-standardisiertes Residuum von Person n auf Item i; Z^2_{ni}: quadriertes z-standardisiertes Residuum von Person n auf Item i; $W_{ni}*Z^2_{ni}$: Mit der Varianz gewichtetes z-standardisiertes Residuum von Person n auf Item i.

Anhand von Proband Nr. 1 lässt sich die Berechnung der einzelnen Werte wie folgt nachvollziehen. Der Erwartungswert *E* von Proband Nr. 1 ergibt sich aus der Modellgleichung des dichotomen Rasch-Modells. In diesem Fall betrug der Logit-Wert der Personenfähigkeit gleich -2.94; die Itemschwierigkeit war -4.40. Nach der Modellgleichung (2) ergibt sich für die Differenz aus Personenfähigkeit und Itemschwierigkeit von (-2.94 - [-4.40]) = 1.46 die Lösungswahrscheinlichkeit .812. Die Varianz *W* einer Person steht für die sog. Modellvarianz. Sie lässt sich direkt aus dem Erwartungswert *E* einer Person nach $W = E*(1-E)$ berechnen und bezeichnet den Anteil an Fehlervarianz, der mit dem Erwartungswert verbunden ist. Im konkreten Fall besteht eine erwartete Lösungswahrscheinlichkeit von rund 81%, allerdings mit einer Fehlervarianz für diese Modellvorhersage von .153. Bei einer rein zufälligen Lösungswahrscheinlichkeit, also bei einem Wettquotienten von 50:50, wäre die Fehlervarianz der Vorhersage verständlicherweise also am größten.

Der *z*-Wert gibt einen Eindruck von der Größe der Modellabweichung, da er skalenunabhängig ist und somit auch die Grundlage für die inferenzstatistische Beurteilung der Residuen darstellt. Anhand der Formeln (7) – (16) lässt sich nun für jeden einzelnen Probanden ein

MNSQ-Outfit- und -Infit-Wert berechnen. Der MNSQ-Outfit-Wert von 1.67 resultiert schließlich aus dem Mittelwert der quadrierten z-standardisierten Residuen Z^2_{ni}, der MNSQ-Infit Wert von 1.85 aus der Summe der mit der Modellvarianz W_i gewichteten z-standardisierten Residuen Z^2_{ni}, geteilt durch die Summe der W_{ni}. Beide Statistiken verweisen mit ihren positiven Abweichungen von ihren Erwartungswerten auf *zu viele* unerwartete Antworten auf diesem Item. Im vorliegendem Fall liegt dies an den Probanden Nr. 3 und Nr. 6, welche erwartungswidrig das Item nicht lösen konnten. Die Infit-Statistik kommt hierbei durch die Gewichtung mit W_i zu einer noch stärkeren Abweichung, weil Proband Nr. 6 einen Logit-Wert von -4.32 (Wert nicht in der Tabelle enthalten) aufweist, der nahe am Itemparameter von -4.40 liegt, also einem Bereich, in welchem die Itemcharakteristische Funktion wegen ihres dort steilsten Anstieges ihr Maximum an Informationsgehaltes bezüglich der Personenfähigkeit aufweist.

Da die z-standardisierten Infit- und Outfit-Maße einen Erwartungswert von Null haben, lässt sich in einem weiteren Schritt bestimmen, ob eine hiervon signifikante Abweichung vorliegt. Mit den MNSQ-Fit-Statistiken hingegen lässt sich die Größe der Abweichung als Effektstärkemaß beschreiben. Hierbei indizieren die MNSQ-Fit-Statistiken das Ausmaß des Zufalls innerhalb der Itemantworten: ein zu hoher Wert (größer als Eins) ergibt sich bei zu vielen unerwarteten Antworten (zu niedriger Trennschärfe), ein zu niedriger hingegen bei zu wenig unerwarteten Antworten (zu hoher Trennschärfe), etwa bei Vorliegen eines (deterministischen) Guttman-Patterns. Im Rahmen der Klassischen Testtheorie würden Items mit sehr hoher Trennschärfe im Test belassen werden. Wie bereits im Kapitel 9.2 kurz dargestellt, ist dies jedoch immer dann problematisch, wenn zwischen Items hohe Abhängigkeiten bestehen, die nicht zulasten der zu messenden latenten Personenvariable gehen, aber zu einer hohen Trennschärfe führen. Ursachen hierfür können dort gefunden werden, wo in Leistungstest ein Item denselben Lösungsweg wie ein vorhergehendes „kopiert“, daher ein Item ein anderes bereits „beantwortet“ oder in Persönlichkeitsfragebögen die Items lediglich Synonymisierungen desselben Gegenstandes darstellen (Itemredundanz), logische Abhängigkeiten zwischen den Items enthalten sind oder aber eine weitere Variable die Kontingenzen der Items mitbedingt. Letztes ist immer dann der Fall, wenn ein Itemmerkmal mit einem Subgruppenmerkmal der Untersuchungsstichprobe korrespondiert, wie etwa Testwiseness bei solchen Personen, welche Zugang zu Testtrainingskursen hatten (s. hierzu Masters, 1988). In jedem der geschilderten Fälle kommt es zu einem Verstoß gegen die Forderung lokaler stochastischer Unabhängigkeit

der Itemantworten als Grundvoraussetzung einer Trait-Messung. Smith (2005) bemerkt hinsichtlich der Auswirkungen von Itemredundanz auf die Messung der Personenfähigkeit:

> …if using a fixed cutscore, then the selection of a redundant (or locally dependent) item could increase the person estimate if the current person estimate is greater than the difficulty of the redundant item or decrease the person estimate if the current person estimate is less than the difficulty of the redundant item. (S. 16).

Allerdings sind nach Simulationsstudien von Smith (2005) und Linacre (2000) die Auswirkungen hinsichtlich der Über- oder Unterschätzung der Personenfähigkeit erst bei ca. 5% redundanter Items einer Skala gravierend. Dem ist u.a. vom Standpunkt der Konstruktvalidität entgegenzuhalten, dass dadurch zwar die Messgenauigkeit der Personenparameter nicht allzu stark betroffen sein mag, die zugrunde gelegte Personenvariable aber durch die Redundanzen bzw. lokalen stochastischen Abhängigkeiten undifferenziert abgebildet wird mit möglicherweise weit reichenden Konsequenzen für die anschließenden inhaltliche Interpretation, wie Wang, Cheng & Wilson (2005) anmerken:

> If the assumption of local item independence is violated, any statistical analysis based on it would be misleading. Specifically, estimates of the latent variables and item parameters will generally be biased because of model misspecification, which in turn leads to incorrect decisions on subsequent statistical analysis, such as testing group differences and correlations between latent variables. In addition, it is not clear what constructs the item responses reflect, and consequently, it is not clear how to combine those responses into a single test score, whether IRT is being used or not…. (S. 6)

Gerade im Falle von Studierendenauswahlverfahren mit ihrem High-Stakes-Charakter muss daher die von Smith (2005) genannt zulässige Grenze redundanter und somit lokal abhängiger Items strenger gesetzt werden.

Zur Beantwortung der Frage nach Grenzwerten zur Selektion Rasch-homogener Items nach den oben beschriebenen unstandardisierten und standardisierten Fit-Statistiken geben Wright und Stone (1979) für die unstandardisierten Werte den Bereich 1 ± 0.5, für die standardisierten 1 ± 0.9 an. Bei der Entscheidung, welcher der beiden Fit-Werte überhaupt für die Itemselektion verwendet werden soll, weist Linacre (2005c) auf die Unterscheidung von „Tests of Perfect Fit“ und „Tests of Useful Fit“ hin. „Tests of Perfect Fit“ überprüfen die Hypothese einer vollkommenen Übereinstimmung der Daten mit den Modellannahmen. Die standardisierten Werte erfüllen genau diesen Zweck. Da aber jedes Modell von Natur aus gewisse Abweichungen von den empirischen Daten zeigt, zumal bei steigendem Stichprobenumfang, empfiehlt Wright (zit. nach Linacre, 2005c) die Verwendung der unstandardisierten Werte mit dem oben angegebenen Toleranzbereich zur Beurteilung eines „Useful Fit“: „ZSTD [standardisierte MNSQ-Statistik] is only useful to salvage non-signficant MNSQ > 1.5, when sample size is small or test length is short“. Weiterhin bemerkt Linacre (2005c): „In general, mean-squares near 1.0 indicate little distortion of the measurement system, regardless of the ZSTD value”. Zur Begründung der Fokussierung bei der Überprüfung der Modellgültigkeit auf die MNSQ-Fit-Statistiken gegenüber den standardisierten Werten führt Linacre (persönliche Mitteilung vom 4.6.2006) aus: „Your fit tests are too sensitive due to the large sample size. You have turned the fit ‘magnification’ up so high that you can see every scratch on the pain of glass. The fit statistics tell you: ‘these data aren't perfect’. The mean-squares tell you ‘but they are good enough’”.

Gleichwohl weisen verschiedene Autoren korrekt darauf hin, dass s_i^2 in Formel (17) von der Varianz und Kurtosis von X_{ni} abhängt, daher von Item zu Item variiert, und es schwierig ist, einen von der Stichprobengröße unabhängigen Trennwert für die MNSQ-Fit-Statistiken anzugeben (R.M. Smith, 1994, 1996; R. M. Smith, Schumacker & Bush, 1998; Wang & Chen, 2005). Es lässt sich also feststellen, dass die Frage nach endgültigen und von der Stichprobe unabhängigen Grenzwerten zum jetzigen Zeitpunkt noch nicht abschließend beantwortet ist. Das wohl beste Verfahren, da unabhängig von den Problemen der parametrischen Ansätze, stellt Karabatsos (2001) unter Bezug zu einem Ansatz von Scheiblechner (1999) vor. Hierbei werden probabilistische Formulierungen der Axiome des Rasch-Modells als Variante des Additive conjoint measurement (Luce & Tukey, 1964) getestet. Allerdings stand dem Autor dieser Arbeit die entsprechende Software leider nicht zur Verfügung. Alternativ bieten sich bislang die Empfehlungen von Wright (1994) an, der für verschiedene Anwendungsbereiche Grenzwerte der MNSQ-Statistiken angibt. Für Testungen im High-Stakes-Bereich (wie

Studierendenauswahlverfahren) empfiehlt Wright (1994) für die Meansquare Infit- und Outfit-Statistik den Bereich von 1 ± 0.2 zur Identifizierung gut passender Items.

Hauptgrundlage der Modellgültigkeitsbeurteilung sollen in der vorliegenden Arbeit daher die Meansquare-Fit-Statistiken als Effektstärkemaße der Modellabweichungen in dem von Wright (1994) empfohlenen Toleranzbereich von 1 ± 0.2 sein, zumal bei Stichprobenumfängen von mehr als 300 Probanden (wie in der vorliegenden Arbeit der Fall) bereits sehr geringe Abweichungen signifikant werden (Linacre, persönliche Mitteilung vom 4.6.2006 sowie Linacre, 2005c). Die z-standardisierten MNSQ-Fit-Statistiken sollen jedoch als zusätzlich Informationsquelle der Modellgültigkeitsüberprüfung hinzugezogen werden.

10.1.3 Überprüfung der Personenhomogenität

Die bisherige Darstellung der Modellgültigkeitsüberprüfung ließ die Betrachtung der Falsifikation des Rasch-Modells durch Personen*heterogenität* noch außer acht. Die Personen-*homogenität* stellt neben derjenigen der Items eine weitere fundamentale Bedingung des Raschmodells dar, da angenommen wird, dass alle getesteten Personen die Testitems mit derselben Personeneigenschaft bearbeiten und nur diese Eigenschaft neben der Itemschwierigkeit die Lösungswahrscheinlichkeit maßgeblich bestimmt. Sollte hingegen eine oder mehrere Substichproben identifizierbar sein, für die sich im Vergleich zu anderen Substichproben unterschiedliche Itemparameter ergäben, würde dies einer eindimensionalen Messung widersprechen, da die Lösungswahrscheinlichkeit zusätzlich durch die Gruppenzugehörigkeit bedingt wäre. Das Prinzip zur Überprüfung der Konstanz der Itemparameter in verschiedenen Substichproben besteht folglich darin, die Itemparameter in verschiedenen Substichproben zu schätzen und zu überprüfen, ob diese sich zwischen den Gruppen unterscheiden. Für die Unterteilung der Stichprobe in zwei oder mehrere Substichproben bieten sich zwei Möglichkeiten an. Die erste besteht darin, der Stichprobenunterteilung eine Hypothese zugrunde zu legen, welche Personengruppen maximal heterogen sein könnten. Für diese werden sodann die Itemparameter getrennt geschätzt und miteinander verglichen. Eine inferenzstatistische Absicherung ist über den bedingten Likelihoodquotiententest von Andersen (1973) möglich. Problematisch hierbei ist allerdings, dass eine nicht signifikante Teststatistik lediglich zeigt, dass sich die Personenheterogenität nicht in der *vermuteten* Stichprobenaufteilung niederschlägt, ein solches Signifikanztestergebnis daher lediglich eine notwendige, aber keine hinreichende Bedingung für die Modellgeltung darstellt. Daher kommt es bei diesem Vorgehen

häufig zu einer erhöhten Beta-Fehlerrate, wenn sich die Personenheterogenität in einem anderen Merkmal als dem spezifizierten niederschlägt (s. hierzu z. B. Stelzl, 1979). Zudem sind oftmals die Voraussetzungen für einen inferenzstatistischen Schluss nicht gegeben: Damit die Prüfstatistik annähernd χ^2-verteilt ist, müssen die erwarteten Antwortpatternhäufigkeiten mindestens gleich Eins sein. Allerdings existieren (wie in dieser Arbeit) bei 20 dichotomen Items eines Subtests $2^{20} = 1\,048\,576$ mögliche verschiedene Antwortpattern. Beobachtet können in der vorliegenden Stichprobe von $N = 434$ allerdings höchstens 434 unterschiedliche Antwortpattern, womit die Voraussetzungen der Prüfverteilung stark verletzt sind. Eine Lösung wäre die Simulation der Prüfverteilung aus den vorhandenen Daten mit sog. Bootstrap-Simulationen (Efron & Tibshirani, 1985). Allerdings stand ein derartiges Verfahren zum Zeitpunkt dieser Arbeit nach Kenntnis des Autors noch nicht zur Verfügung.

Die zweite Möglichkeit, maximal heterogene Teilstichproben zu identifizieren, die zugleich einen Ausweg aus der zuvor beschriebenen Problematik bietet, besteht demgegenüber darin, die Gesamtstichprobe ohne ein zuvor festgelegtes manifestes Teilungskriterium mit mixed Rasch-Modellen (MRM) (Davier & Rost, 1995; Rost, 1990, 1991, 1996; Rost, Carstensen & Davier, 1997) zu separieren. Die Grundidee hierbei besteht darin, die Stichprobe hinsichtlich maximal unterschiedlichen Antwortverhaltens zu „entmischen“ und die Personen dadurch disjunkten und exhaustiven Klassen zuzuordnen. Hierbei wird die Forderung des Rasch--Modells nach der Konstanz der Itemparameter zwischen verschiedenen Stichproben aufgegeben und unterschiedliche Itemparameter zwischen den Personenklassen zugelassen. Infolgedessen gilt das Rasch-Modell daher nicht mehr für die gesamte Population, sondern mit jeweils unterschiedlichen Itemparametern in verschiedenen Personenklassen. Das Ergebnis ist ein sowohl klassifizierendes als auch quantifizierendes Testmodell: Zwischen den Personenklassen bestehen aufgrund heterogener Antwortprofile qualitative Unterschiede. Innerhalb jeder Klasse hingegen gilt das Rasch-Modell mit der Implikation der Messung einer quantitativen Personenvariable.

Im Rahmen von Leistungstests können sich diese Klassen als eine Separierung in Personen mit und ohne Ratestrategien (Köller, 1994), als Personenklassen mit unterschiedlichen Lösungsstrategien (Hosenfeld, Strauss & Köller, 1997; Köller, Rost & Köller, 1994) oder als Klassen unterschiedlicher Auswirkung von Testzeitbegrenzung (Bolt, Cohen & Wollack, 2002) niederschlagen. Durch die Erweiterung des mixed Rasch-Modells für mehr als zwei Antwortkategorien lassen sich weiterhin im Bereich der Persönlichkeitspsychologie unterschiedlich prägnante Typen identifizieren (Heene & Funke, 2006; Köller, Baumert & Rost, 1998; Rost, 2002) oder Personenklassen mit unterschiedlichen Antworttendenzen (Eid,

2000; Rost, Carstensen & Davier, 1997) separieren. Bei der Anwendung des mixed Rasch-Modells zur Überprüfung der Personenhomogenität wird daher analysiert, ob eine Aufteilung in zwei mixed Rasch-Klassen eine bessere Passung der Daten erzielt wird als durch das Rasch-Modell, welches bekanntermaßen lediglich eine Personenklasse annimmt.
Wie Rost und Davier (1995) zeigen konnten, stellt die Anwendung des mixed Rasch-Modells einen strengeren Modelltest zur Überprüfung der Personenhomogenität dar als der Likelihoodquotiententest von Andersen (1973). Aus diesem Grunde wird in dieser Arbeit zur Überprüfung der Personenhomogenität stets das MRM in zwei Klassen gegen das normale Rasch-Modell geprüft. Im Folgenden werden die grundlegenden und für das Verständnis notwendigen Gleichungen des MRM für dichotome Items dargestellt.

Die Modellgleichung des mixed Rasch-Modells für dichotome Items ist gegeben durch Gleichung (18):

$$p(X_{vi} = 1 | g) = \sum_{g=1}^{G} \pi_g \frac{\exp(\theta_{vg} - \sigma_{ig})}{1 + \exp(\theta_{vg} - \sigma_{ig})}, \qquad (18)$$

wobei g: Index der Personenklasse

π_g: Klassengrößeparameter mit $\sum_{g=1}^{G} \pi_g = 1$ (19)

und $\sum_{i=1}^{k} \sigma_{ig} = 0,$ (20)

was einer Summennormierung der Itemparameter in allen Klassen g entspricht.

Im Unterschied zum normalen Rasch-Modell werden die Parameter in (18) mit einem zusätzlichen Index g versehen, welcher die Abhängigkeit der Modell-Parameter von der Klassenzugehörigkeit spezifiziert.
Wie weiterhin aus Formel (18) ersichtlich, geht das MRM für G = 1 in das normale Rasch-Modell über, wobei die Klassenanzahl G hierbei eine präexperimentell vorzugebende Größe ist. Zur Überprüfung der Personenhomogenitätsannahme wird verständlicherweise die Klassenanzahl auf G = 2 gesetzt, da die Homogenitätshypothese bereits bei einer besseren Passung der Zweiklassenlösung verworfen wäre. Es stellt sich dabei die Frage, wie die wahrscheinlichste Zuordnung einer Person mit dem Antwortmuster $\underline{x}$ zu einer dieser Klassen erfolgt. Es liegt nahe, dieses Problem einer bedingten Wahrscheinlichkeit des Antwortmusters

$\underline{x}$ unter der Gruppenzugehörigkeit mit einem Bayes-Ansatz nach Formel (21) zu lösen (s. auch Rost, 2004 S. 160f.):

$$p(g|\underline{x}) = \frac{\pi_g p(\underline{x}|g)}{\sum_{h=1}^{G} \pi_h p(\underline{x}|h)}, \qquad (21)$$

wobei g: Index der Personenklasse

π_g: Klassengrößeparameter

$\underline{x}$: Antwortvektor

h: zweiter Laufindex für die latenten Klassen

Die bedingte Klassenwahrscheinlichkeit ist demnach eine Funktion der mit der Klassengröße π_g multiplizierten Wahrscheinlichkeit, mit der eine Person der Klasse g dieses Antwortmuster produziert, dividiert durch die Summe der Zuordnungswahrscheinlichkeiten. Allerdings ist hierdurch noch nicht bestimmt, wie man zu den zu schätzenden Klassengrößenparametern π_g gelangt, wenn die Modellparameter noch unbekannt sind. Hierzu wird eine zunächst willkürliche Aufteilung mit resultierenden Modellparametern durch einen Erwartungswert--Maximierungs-Algorithmus (EM-Algorithmus, s. z. B. Dellaert, 2002) in mehreren Iterationen so lange verändert bzw. optimiert, bis die Wahrscheinlichkeit der Daten unter den Modellannahmen ein Maximum erreicht. Formel (21) liefert sodann ein Maß für die Zuordnungssicherheit einer Person zu einer der Personenklassen. Als Gesamtmaß der Messgüte kann die mittlere Zuordnungswahrscheinlichkeit herangezogen werden.

Um nun einen validen Vergleich der Einklassen- gegenüber der Zwei-Klassenlösung des MRM vornehmen zu können, kann ein Modellvergleich mit den informationstheoretischen Maßen Akaikes Information Criterion (AIC), Baysian Information Criterion (BIC) und Consistent Akaikes Information Criterion (CAIC) vorgenommen werden (zur Ableitung der Indizes s. u.a. Lin & Dayton, 1997; Read & Cressie, 1988). Diese stellen allerdings rein deskriptive Maße dar, anhand derer sich *relative* Modellvergleiche vornehmen lassen, und gestatten daher keine eindeutigen Aussagen darüber, in welchem Ausmaß der Wert eines Modells kleiner sein soll als der eines konkurrierenden Modells. Folglich steht bei diesen Indizes vielmehr das *Brauchbarkeitskriterium* eines Modells als seine inferenzstatistische Absicherung im Vordergrund. Allerdings ist auch in all den hier beschriebenen Maßen das Maximum der Likelihoodfunktion als die Wahrscheinlichkeitsfunktion der Datenmatrix unter den

Modellannahmen enthalten, wie die folgenden Formeln der Informationsmaße, die in dieser Arbeit verwendet werden, zeigen (Gleichung (22) bis (24)):

$$AIC = -2\log L + 2 * n_p \qquad (22)$$

mit n_p: Parameteranzahl
logL: natürlicher Logarithmus der Likelihoodfunktion

$$BIC = -2\log L + (\log N) * n_p \qquad (23)$$

mit N: Stichprobengröße

$$CAIC = -2\log L + (\log N) * n_p + n_p \qquad (24)$$

Allen Indizes gemeinsam ist, dass sie sog. Straffunktionen darstellen: In die Formeln fließen sowohl die empirischen Likelihoodwerte als Maße der empirischen Gültigkeit als auch Gewichtungen dieser Werte mit der Parameteranzahl ein, um ein eine Verbesserung der Likelihoodfunktion durch zusätzliche Modellparameter im Sinne des Einfachheitsprinzips von Modellen zu bestrafen und sparsame Modelle zu bevorzugen. Das AIC nimmt hierbei eine gleichartige Gewichtung der Parameteranzahl n_p und des Likelihoodwertes vor, was zu einer Bevorzugung von Modellen mit vielen Parametern führt und somit dem Sparsamkeitsprinzip zuwiderläuft. Demgegenüber wird im BIC (Bayes Information Criterion) eine Gewichtung der Parameterzahl mit dem Faktor log*N vorgenommen. Bei großen Datensätzen kann die große Menge an Antwortmustern stets durch zusätzliche Parameter besser modelliert werden, durch die Gewichtung der Parameteranzahl mit dem Logarithmus von N wirkt dies jedoch einer Bevorzugung überparametrisierter Modelle entgegen. Da der AIC bei steigendem Stichprobenumfang einem systematischen Bias unterliegt, stellt der CAIC (Consistent AIC) schließlich eine Korrektur dar, die zu konsistenten Schätzungen führt. Als Auswahlkriterium empfiehlt Rost (2004, S. 344) den AIC bei kleinen Itemanzahlen mit großen Patternhäufigkeiten, den BIC hingegen bei großen Itemanzahlen und kleinen Patternhäufigkeiten. In dieser Arbeit werden Modellvergleiche zwischen der Einklassen- und der Zweiklassenlösung des MRM im Sinne des Einfachheitskriteriums und der Konsistenz der Schätzung anhand des BIC und CAIC vorgenommen. Für die Annahme einer Klassenlösung sollten hierbei beide Indizes zu konkordanten Ergebnissen bezüglich der relativen Passung führen. Ob ggf. die nach den

informationstheoretischen Maßen besser passende Zweiklassenlösung des mixed Rasch--Modells signifikant besser als die Einklassenlösung passt, lässt sich mit einem Likelihood--Quotiententest untersuchen (Rost, 2004, S. 331f.). Hierbei werden die Likelihoods der beiden Modelle nach Gleichung (25) in Beziehung gesetzt:

$$\mathrm{LR} = \frac{\mathrm{L}_0}{\mathrm{L}_1}, \qquad (25)$$

wobei L_0: Likelihood des restriktiveren Modells,

L_1: Likelihood des weniger restriktiven Modells

Der Likelihood-Quotient lässt sich schließlich in eine χ^2-verteilte Prüfstatistik umwandeln (Gleichung (26)):

$$-2\log\left(\mathrm{LR}\right) \rightarrow \chi^2, \qquad (26)$$

mit $\mathrm{df} = \mathrm{n_p}\left(\mathrm{L}_1\right) - \mathrm{n_p}\left(\mathrm{L}_0\right)$,

wobei n_p: jeweilige Parameteranzahl der Modelle.

Da der Logarithmus eines Likelihoodquotienten gleich der Differenz der log-Likelihood ist, so gilt für den Modellvergleich die folgende Prüfgröße

$$\log(LR) = \log(L_0) - \log(L_1). \qquad (27)$$

Zur Überprüfung der Personenhomogenitätsannahme des Rasch-Modells wird daher in dieser Arbeit zunächst die Zweiklassenlösung des mixed Rasch-Modells (s. u.a. Rost, 1990) anhand der informationstheoretischen Deskriptiv-Maße BIC und CAIC (s. u.a. Read & Cressie, 1988) vorgenommen; sollten sich hieraus Hinweise auf eine bessere Passung der Daten mit der Zweiklassen-Lösung ergeben, wird diese mit dem Likelihood-Quotiententest inferenz-statistisch abgesichert.

10.1.4 Reliabilitätsmaße des Rasch-Modells

Weil das Reliabilitätskonzept im Rahmen der Rasch-Modelle in großen Teilen von demjenigen der Klassischen Testtheorie abweicht bzw. dieses ergänzt, sollen im Folgenden zum besseren Verständnis diejenigen Reliabilitätsmaße dargestellt werden, die in der vorliegenden Arbeit zum Tragen kommen.

High-Stakes-Testungen wie Studierendenauswahlverfahren erfordern strenge Methoden und Standards, um die Reliabilität von Messungen zu beurteilen und die daraus abgeleiteten Schlüsse und Entscheidungen begründen zu können. Das Standardmaß einer solchen Beurteilung der internen Konsistenz eines Tests im Rahmen der Klassischen Testtheorie stellt die Formel 20 von Kuder und Richardson (KR-20) für dichotome Items dar, wie sie Gleichung (28) definiert:

$$KR-20 = \left(\frac{c}{c-1}\right) * \left(1 - \frac{\sum p_i q_i}{\sigma_X^2}\right) \qquad (28)$$

wobei
c: Anzahl der Items
p_i: Itemschwierigkeit
q_i: $1-p_i$
σ_x: Rohwertevarianz

Das Produkt $p_i q_i$ stellt die auf Basis der Binomialverteilung geschätzte Varianz eines Items i für eine Person dar, für die *dieser* p-Wert der Lösungswahrscheinlichkeit entspricht. Da jeder p-Wert den Stichprobenmittelwert darstellt, entspricht jeder dieser Werte der Lösungswahrscheinlichkeit einer durchschnittlichen Person der untersuchten Stichprobe für dieses Item. Jedes Produkt $p_i q_i$ stellt demnach eine Schätzung der mittleren Stichprobenvarianz der Itemantworten dar. Bemerkenswert daran ist, dass KR-20 dadurch eine Schätzung der Reliabilität als dem Verhältnis aus wahrere *Personen*varianz zur beobachteten *Personen*-gesamtvarianz für die jeweilig untersuchte Stichprobe unter der jeweiligen Itemauswahl darstellt. Sie ist daher vielmehr eine „Personenstichproben-Reliabilität“, indem sie die Reproduzierbarkeit der Personenrangordnung angibt, als eine Testreliabilität im eigentlichen Sinne. Linacre (1997) bemerkt hierzu pointiert: “Thus ‘reliability’ is not an index of quality (‘Is this a good measure of ...?’), but of relative reproducibility (‘How repeatable is this

measure?')". Smith (2001) bringt das Dilemma der Stichproben- und Itemauswahlabhängigkeit klassischer Reliabilitätskennwerte und der fehlenden Intervallskalenqualität von Rohwerten auf den Punkt, wenn er schreibt:

> Furthermore, in many research studies estimates of internal consistency are reported based on previously reported estimate of an assessment. However, a previously reported estimate of internal consistency is not informative unless the proposed sample has exactly the same score distribution as the sample used for the reported internal consistency. This fact brings into question the results of studies that fail to report estimates of internal consistency for the current data, instead relying on previously reported estimates to somehow carry over to the current study.
>
> Finally, the use of raw scores in calculating the sample variance is potentially misleading as raw scores are not linear (S. 283).

Die indirekte Methode einer Abschätzung der Varianz der wahren Werte über die durchschnittliche Stichproben-Fehlervarianz ist im Rahmen von Rasch-Modellen nicht erforderlich. Hierbei wird vielmehr eine *direkte* Schätzung der Fehlervarianz über den Erwartungswert der Standardschätzfehler jedes Personenparameters vorgenommen, da nach Testkalibrierung zu jedem Personenparameter der korrespondierende Standardfehler bestimmt werden kann.
Die Varianz des Fehleranteils der Messwerte über alle Personen lässt sich hierbei über das mittlere Abweichungsquadrat der Fehlervarianz der Personenparameter (MSE_p) wie in Gleichung (29) berechnen:

$$MSE_p = \sum_{n=1}^{N} S_n^2 / N, \qquad (29)$$

wobei S_n: Standardfehler jedes Personenparameters der beobachteten Varianz SD^2. Somit resultiert nun die wahre Varianz aus Gleichung (30):

$$SA_p^{\,2} = SD_p^{\,2} - MSE_p \qquad (30)$$

Schließlich resultiert die Personen-Separations-Reliabilität R_p als formales Äquivalent zur Personenreliabilität der klassischen Testtheorie durch Gleichung (31):

$$R_p = SA_p^2 / SD_p^2 = 1 - (MSE_p / SD_p^2) \qquad (31)$$

Das Besondere an dieser Form der Reliabilität ist, dass bei Modellgeltung die Fehlervarianz unabhängig von der Varianz der beobachteten Messwerte bestimmbar ist. Somit ist auch die Fehlervarianz jedes Personenparameters unabhängig davon, welche Personen sich in der Stichprobe befinden. Sie hängt allein von der Anzahl und Schwierigkeit der Items eines Tests ab. Damit ergibt sich bei Modellgeltung - im Unterschied zur „klassischen“ Reliabilität - die Stichprobenunabhängigkeit der Reliabilitätsschätzung. Zudem stellt R_p immer dann eine bessere Schätzung der Reliabilität dar, wenn der Test durch Boden- oder Decken-Effekte gekennzeichnet ist und die meisten Personenparameter dadurch hohe Standardschätzfehler aufweisen. Problematisch an KR-20, wie allerdings auch an Rp ist, dass sie sich auf keiner linearen Skala befinden, da beide auf das Intervall {0,1} beschränkt sind. Vergleiche von Reliabilitätskennwerten über verschiedene Studien hinweg werden somit erschwert (Bond & Fox, 2001). Einen alternativen Index, der im Intervall $\{0, +\infty\}$ variiert und somit auf einer Verhältnisskala liegt, stellt der Personen-Separationsindex (G_p) dar, der durch Gleichung (32) gegeben ist:

$$G_p = SA_p / SE_p = \sqrt{(R_p /(1 - R_p)}\,, \qquad (32)$$

wobei SE_p: Root Mean Square Measurement Error (RMSE), mit $RMSE = \sqrt{MSE_p}$.

G_p stellt das Verhältnis wahrer Standardabweichung zur Fehler-Standardabweichung dar. G_p ist daher ein Index der Spannweite von Personenparametern, ausgedrückt in Standardfehler-einheiten. Je höher dieser Index, desto stärker differenziert der Test zwischen den Personen. So würde etwa ein G_p von vier bedeuten, dass vier voneinander statistisch unterschiedliche Fähigkeitsbereiche mit dem Test identifiziert werden können.
Die bisher dargestellten Indizes bezogen sich alleine auf die Reliabilität und Separation von Personen. Jedoch lässt sich R_p und G_p jeweils ebenso auf Items beziehen, ersetzt man die Personenvarianz durch die Itemvarianz (Formel (33) und (34)):

$$R_i = SA_i^2 / SD_i^2 = 1 - (MSE_i / SD_i^2) \quad (33)$$

$$G_i = SA_i / SE_i = \sqrt{(R_i / (1 - R_i)} \quad (34)$$

Wie unschwer aus (33) zu ersehen, stellt dieser Index in sofern eine deutliche Verbesserung zum klassischen Reliabilitätskonzept dar, weil hier tatsächlich Itemvarianzanteile und keine von Personenantworten miteinander in Beziehung gesetzt werden. Eine besondere Bedeutung erhalten R_i und G_i bei der Beurteilung der Inhaltsvalidität, weil durch sie die Separationsreliabilität der Items bzw. die Spannweite der Itemparameter indiziert wird. Je höher beide ausfallen, desto stärker decken die Items das Konstruktkontinuum in seinen Schwierigkeitsabstufungen ab. Schumacker (2003) verdeutlicht den Unterschied zwischen Personenseparationsreliabilität und den Itemseparationskennwerten anhand der Konstruktvalidierung eines Fragebogens zur Lehrqualität, in welchem sich niedrige Reliabilitäten der Personen- im Vergleich zu sehr hohen der Itemseparation ergaben:

> The superintendents DID NOT consistently respond across the 10 items, as indicated by both the Cronbach alpha and Rasch person separation reliability indices, however the items did provide an excellent measurement ruler based on the item separation reliability index. The items are content valid as quality indicators in education based on prior research literature. A different or larger sample of superintendents would more clearly indicate the ordered preference for the quality indicators. *Persons can therefore be unreliable while the measurement ruler is reliably indicated* [Hervorhebung vom Verfasser] (S. 15).

Smith (2001) bezieht diesen, über das klassische Reliabilitätskonzept hinausgehenden Beitrag, direkt auf den Aspekt der Konstruktvalidierung, indem er schreibt: „If a sufficient (at least 2) number of item difficulty levels are unable to be identified, then one may have difficulty in interpreting the variable defined by the items ...“

Durch diese Ausführungen sollte dargestellt werden, warum sich diese Arbeit auch in Bezug auf die Reliabilität eines Tests überwiegend auf die entsprechenden Kennwerte der Rasch--Modelle bezieht: Die Reliabilitätskennwerte der Klassischen Testtheorie können zum einen problemlos äquivalent zu den entsprechenden Indizes der Rasch-Modelle dargestellt werden, wie im Falle der Personenseparations-Reliabilität, zum anderen können sie durch eine direkte Schätzung der Fehlervarianz über die Standardfehler der Personenparameter approximiert werden. Diese Methode führt gerade dann zu verlässlicheren, „ehrlicheren“ Schätzungen der Reliabilität, wenn Stichproben des oberen Fähigkeitsbereichs zur Testkonstruktion und -optimierung herangezogen werden, wie es gerade bei Personen mit Hochschulzugangsberechtigung der Fall ist. Die Reliabilitätskennwerte der Klassischen Testtheorie basieren überdies auf der irrigen Annahme einer Intervallskala der Rohwerte, wobei diese stets lediglich ordinale Relationen abbilden (s. hierzu bereits Thorndike, 1918). Der größte Vorteil der Rasch-basierten Reliabilitätskennwerte ergibt sich jedoch bei Modellgeltung aus der Stichprobenunabhängigkeit der Indizes, da die Reliabilität hierbei allein von der Anzahl und der Schwierigkeit der Items abhängt, und folglich unabhängig von der Varianz der beobachteten Messwerte ist. Streng genommen stellen die Reliabilitätskennwerte der Klassischen Testtheorie zudem lediglich Indizes der Reproduzierbarkeit der Personenrangordnung dar, und liefern daher nur einen indirekten (und geringen) Hinweis auf die Güte der Items als Bestandteile einer Messskala. Grundsätzlich wird dadurch die Leistung von Personen mit der Qualität des Testverfahrens als Messskala fälschlich gleichgesetzt. Insbesondere die Reliabilitätsindizes der Itemseparation überkommen diese Beschränkungen durch eine Verrechnung von Itemvarianzanteilen und liefern daneben Hinweise über die Breite der Erfassung des latenten Kontinuums durch die Items. Gerade durch diese Kennwerte, die kein Äquivalent in der Klassischen Testtheorie finden, wird die Interpretierbarkeit der hypothetisch zugrunde liegenden Variable bei Modellgeltung im Sinne einer „construct map“ (Wilson, 2003) mit ermöglicht und kann daher auch nicht nur als Basis einer normorientierten, sondern ebenso kriteriumsorientierten Messung genutzt werden.

10.2 Erläuterungen zu Skalenanalysen nach dem Multifacetten-Rasch-Modell

Da der Aufgabenstellung zum empiriebezogenen Denken und zu den Kreativitätsfacetten ein freies Antwortformat unterlag, erfolgte die Erfassung der Personenantworten, wie in Kapitel 8.1.3 dargestellt, durch zwei unabhängige Beurteiler anhand von Ratingskalen zu vorgegebenen Beurteilungsdimensionen bzw. dem Auswertungsschema aus dem BIS (Jaeger, Suess

& Beauducel, 1997). Ein zentrales Problem derartiger Ratings stellt ihre Anfälligkeit für verschiedene Formen von Urteilsfehlern dar (Bortz & Döring, 1995; Guilford, 1954). Ein Urteilsfehler („rater-bias“) ist hierbei über eine mangelnde Übereinstimmung von Beurteilern aufgrund a) unterschiedlicher Auffassungen der verwendeten Ratingskala oder b) beurteilerspezifische Wahrnehmung der zu beurteilenden Untersuchungsobjekte definiert (Hoyt, 2000). Der im Kontext von Leistungsbeurteilungen wohl am häufigste erwähnte Urteilsfehler ist die Tendenz zur Strenge oder Milde (Myford & Wolfe, 2003, 2004). Diese Fehler können als die Tendenz eines Beurteilers definiert werden, Einstufungen der Leistungsfähigkeit von Probandenantworten vorzunehmen, die im Durchschnitt niedriger bzw. höher liegen als diejenigen anderer Beurteiler, selbst wenn mögliche Unterschiede in der Leistungsfähigkeit der jeweils beurteilten Personen berücksichtigt werden. Eine Variante hiervon stellt die differenzielle Strenge oder Milde dar, bei welcher der jeweilige Effekt lediglich bei bestimmten Personengruppen (Frauen oder Männern, Altersgruppen, Nationalitäten u.a.) oder nur unter bestimmten Urteilssituationen (Wettkampfsituationen, Tageszeiten etc.) auftreten. Weitere mögliche Urteilseffekte sind die zentrale Tendenz und der Halo-Effekt (Myford & Wolfe, 2003, S. 396f.). Zentrale Tendenz bezeichnet die Neigung, die mittleren Antwortkategorien einer Ratingskala zu verwenden. Der Halo-Effekt ist als die Tendenz definiert, auf konzeptuell unterschiedlichen Merkmalen ähnliche Beurteilungen vorzunehmen. Dieser Effekt tritt insbesondere dann auf, wenn Globalurteile Einzelurteile „überstrahlen“.

Insgesamt betrachtet ist es das wesentliche Ziel einer objektiven Leistungsbeurteilung, derartige *konstruktirrelevante* Varianzanteile zu minimieren. Gerade bei Leistungsbeurteilungen durch verschiedene Beurteiler stellt sich daher umso dringlicher die Frage, wie das Ergebnis der Beurteilung möglichst unabhängig von derartigen Fehlerfaktoren gehalten werden kann. Denn gerade Unterschiede in der Beurteilerstrenge haben sich in zahlreichen Studien als kritische Einflussgrößen gezeigt (Eckes, 2005a, 2005b; G. Engelhard, 1994; G. Engelhard, Jr., 1992, 1996a, 1996b, 2002; Fitzpatrick, Ercikan, Yen & Ferrara, 1998; Myford, 2002; Myford & Wolfe, 2002; Wolfe, Moulder & Myford, 2001). So konnten Hoyt und Kerns (1999) in einer Metaanalyse über 79 Einzelstudien zeigen, dass 37% der Urteilsvarianz auf Urteilsfehler zurückgingen, bei nicht direkt beobachtbaren Merkmalen (z. B. Fähigkeiten) waren es sogar 49%. Selbst zeitlich ausgedehnte Beurteilertrainings von 25 und mehr Stunden ergaben keine schwächer ausgeprägten Urteilseffekte als kürzere (5 bis 24 Stunden).

Üblicherweise wird die Güte der Übereinstimmung von Beurteilern im Rahmen der Klassischen Testtheorie behandelt und evaluiert (Wirtz & Caspar, 2002). Hierbei wird meist über die Intraklassenkorrelation (s. z. B. Wirtz & Caspar, 2002) berechnet, in wieweit Gruppen von Beurteilern dem Ideal einer wechselseitigen Austauschbarkeit nahe kommen. Der entscheidende Nachteil dieses Vorgehens ist, dass die Effekte von Beurteilerfehlern unzulänglich oder gar nicht beachtet werden. Bei der unadjustierten Intraklassenkorrelation etwa werden Mittelwertsdifferenzen zwischen Beurteilern dem Zufallsfehler zugerechnet, obgleich es sich um einen systematischen Fehler handelt. Das Modell der adjustierten Intraklassenkorrelation fasst diese zwar als eigene und systematische Varianzquellen auf, wobei jedoch eine Schätzung der Strenge jedes einzelnen Beurteilers unterbleibt. Statistische Korrekturen von *Einzel*urteilen um Urteilsfehler sind somit ausgeschlossen. Zwar wäre bei zwei oder mehr Beurteilern mit entgegengesetzten Strengetendenzen im Durchschnitt eine „ausgewogene“, unverzerrte Beurteilung möglich, allerdings: ohne Kenntnis der individuellen Strengetendenz der Beurteiler würde dieser ausgleichende Effekt dem Zufall obliegen. Noch problematischer wäre eine Zuweisung von Beurteilern mit ähnlicher Strenge- oder Mildetendenz. Hier würde sich zwar mit hoher Wahrscheinlichkeit eine nach der Klassischen Testtheorie gute Beurteilerübereinstimmung in Form einer hohen Intraklassenkorrelation ergeben. Allerdings wäre diese kein Maß der Genauigkeit, sondern eines der Übereinstimmung gleicher Urteilsfehler!
Demgegenüber bietet das Multifacetten-Rasch-Modell (Linacre & Wright, 2002) die Möglichkeit, jedem Beurteiler einen eigenen Strengewert zuzuweisen und hierüber die einzelnen Leistungsbeurteilungen gegebenenfalls zu korrigieren, um zu einer fairen Leistungsbeurteilung zu gelangen. Auf die im Kapitel 9 bereits ausführlich beschriebenen Vorteile des Rasch-Modells soll in diesem Zusammenhang lediglich nur noch einmal hingewiesen werden, um im folgenden Abschnitt die grundlegende Konzeption des Multifacetten-Rasch-Modells und seine mathematischen Definition darzustellen.

10.2.1 Grundlagen des Multifacetten-Rasch-Modells und seine mathematische Definition

Das Multifacetten-Rasch-Modell (MFRM) erweitert das dichotome Rasch-Modell (Rasch, 1960) und diejenigen für polytome Antwortvariablen (Andrich, 1978; Masters, 1982).
Das Ziel einer Analyse mit dem MFRM ist es, Informationen über einzelne Ausprägungen von Elementen der Facetten einer Leistungsbeurteilung (Personenfähigkeit, Beurteilereffekte,

Aufgabenschwierigkeit etc.) zu erhalten und sie somit zu objektivieren und zu präzisieren. Wesentlich hierbei ist, dass die quantitativen Aussagen der jeweiligen Facetten-Elemente spezifisch objektiven Vergleichen genügen sollen, indem die Schätzungen der Personenleistungsfähigkeit unabhängig von der Verteilung der Strenge der Beurteiler und unabhängig von derjenigen der Aufgabenschwierigkeiten sind, wie auch die Schätzungen der Beurteilerstrenge und der Aufgabenschwierigkeiten jwls. unabhängig voneinander und von der Verteilung der Personenfähigkeit sein soll.

Das MFRM kann verschiedenste Facetten-Strukturen modellieren. In seiner einfachsten Formulierung umfasst es die Facetten Probanden und Beurteiler, in komplexeren Erhebungsdesigns können bspw. Probanden, Beurteiler, Aufgaben, Geschlecht und Beurteilungszeitpunkte modelliert werden (für einen Überblick über weitere Modellvarianten s. z. B. Eckes, 2005a; Myford & Wolfe, 2003, 2004). Das Basismodell der vorliegenden Arbeit umfasst die Messung der Fähigkeit von Probanden, der Strenge von Beurteilern und der Schwierigkeit der Aufgaben. Dieses Modell lautet in logarithmischer Schreibweise:

$$\ln\left[\frac{p_{vjik}}{p_{vjik-1}}\right] = \theta_v - \alpha_j - \sigma_l - \tau_k, \qquad (35)$$

wobei p_{vjik} : Wahrscheinlichkeit einer Einstufung von Person v bei Kriterium i durch Beurteiler j in Antwortkategorie k einer Ratingskala

p_{vijk-1} : Wahrscheinlichkeit einer Einstufung von Person v bei Kriterium i durch Beurteiler j in Antwortkategorie k-1

θ_v : Fähigkeitsparameter von Person v

σ_i : Schwierigkeitsparameter von Kriterium i

α_j : Strengeparameter von Beurteiler j und

τ_k : Schwierigkeitsparameter von Kategorie k.

Der Schwierigkeitsparameter τ von Kategorie *k* gibt die Wahrscheinlichkeit an, eine Einstufung in Kategorie *k* zu erhalten, relativ zu einer in Kategorie *k*-1. Je höher also der Parameterwert, desto geringer ist die Wahrscheinlichkeit einer Einstufung in Kategorie k. Der Modellparameter τ_k in Modellgleichung (35) definiert darüber hinaus die Struktur der Ratingskala, die den Beurteilungen zugrunde liegt. In der vorliegenden Arbeit wird das Ratingskalen-Modell von Andrich (1978) zugrunde gelegt. Bei diesem wird bei den Bewertungen bezüglich aller Aufgaben die gleiche Struktur der Ratingskala hypothetisiert,

d.h., dass die Abstände der Schwellen, zwischen allen Beurteilern als gleich groß angenommen werden. Dies impliziert, dass für alle Beurteiler die gleiche Struktur der Ratingskala gilt (vgl. auch Rost, 2004, S. 215f.) und somit die Ratingkategorien für sie die gleiche Bedeutung haben. Schwellen definieren hierbei diejenigen Punkte auf dem latenten Fähigkeitskontinuum, an denen der Übergang von einer Antwortkategorie zur nächsten stattfindet und die sich mathematisch aus den Abszissenwerten der Schnittpunkte benachbarter Kategorienfunktionen ergeben. Letztere definieren (wie beim dichotomen Rasch-Modell die itemcharakteristische Funktion) die Abhängigkeit der Antwortwahrscheinlichkeit der jeweiligen Kategorie von der latenten Variablen.

Aus (35) ergibt sich die Modellgleichung in Exponentialform mit der Wahrscheinlichkeit *p*, dass Person *v* durch Beurteiler *j* bei Aufgabe *l* ein Rating in Kategorie *k* (*k* = 0,..., *m*) erhält durch:

$$p_{vjlk} = \frac{\exp\left[k\left(\theta_v - \alpha_j - \sigma_l\right) - \sum_{s=0}^{k} \tau_s \right]}{\sum_{r=0}^{m} \exp\left[r\left(\theta_v - \alpha_j - \sigma_l\right) - \sum_{s=0}^{r} \tau_s \right]}, \qquad (36)$$

wobei $\tau_0 = 0$.

10.2.2 Maße der Qualitätssicherung des Multifacetten-Rasch-Modells

10.2.2.1 Residuenanalyse

Das Ausmaß der Abweichungen der beobachteten von den nach der Modellgleichung erwarteten Ratings kann wie im Falle des dichotomen Rasch-Modells auch im MFRM über standardisierte Residuen ausgedrückt werden. Formal lassen sich die standardisierten Residuen für Modellgleichung (36) wie folgt darstellen:

$$z_{vjl} = \frac{x_{vjl} - e_{vjl}}{w_{vjl}^{1/2}}, \qquad (37)$$

wobei x_{vjl} das beobachtete Rating für Person *v* durch Beurteiler *j* bei Aufgabe *l* ist und

$$e_{vjl} = \sum_{k=0}^{m} kp_{vjkl} \qquad (38)$$

das nach Modellgleichung (36) erwartete Rating für Person *v* auf Kriterium (oder Merkmal) *l* durch Beurteiler *j* definiert. Der Nenner in Gleichung (37) ist die Varianz des Residuums ($x_{vjl} - e_{vjl}$), die definiert ist durch

$$w_{vjl} = \sum_{k=0}^{m} (k - e_{vjl})^2 \, p_{vjlk} \, . \qquad (39)$$

10.2.2.2 Modellgeltung

Die Modellgeltung aller Parameter des Messmodells lässt sich überprüfen, indem die standardisierten Residuen über verschiedene Facetten und über die verschiedenen Elemente einer gegebenen Facette hinweg aufsummiert werden. In der Regel geschieht dies mittels der bereits in Kapitel 10.1.2.1 beschriebenen Mean-Square-Outfit- und der Mean-Square Infit--Statistiken.
Die Outfit-Statistik berechnet sich beispielhaft für die Beurteilerfacette nach:

$$\text{Outfit}_j = \frac{\sum_{v=1}^{N} \sum_{l=1}^{L} z_{vjl}^2}{N * L}, \qquad (40)$$

wobei N die Anzahl von Elementen der Facette der beurteilten Personen und L die Anzahl von Aufgaben ist.
Die Infit-Statistik ist definiert durch:

$$\text{Infit}_j = \frac{\sum_{v=1}^{N} \sum_{l=1}^{L} w_{vjl} z_{vjl}^2}{\sum_{v=1}^{N} \sum_{l=1}^{L} w_{vjl}}, \qquad (41)$$

Die Outfit-Statistik erfasst, inwieweit ein ansonsten konsistent einstufender Beurteiler unerwartete Urteile in den äußeren Skalenbereichen abgibt (“outlier-sensitive fit“). Die Infit-Statistik hingegen ist sensitiv gegenüber unerwarteten Ratings, die sich im mittleren Skalenbereich bewegen (“inlier-sensitive“).

Wie bereits in Kapitel 10.1.2.1 belegt, haben Infit- und Outfit-Statistik einen Erwartungswert von 1 und variieren im Intervall $\{0, +\infty\}$. Fitwerte deutlich größer 1 indizieren, dass die Ratingdaten mehr Variation besitzen, als nach dem Modellannahmen zu erwarten wäre, demgemäß der Zufallsanteil in den Urteilen zu groß ist. Umgekehrt weisen Fitwerte deutlich kleiner 1 darauf hin, dass die Ratingdaten weniger Variation aufweisen als vom Modell vorhergesagt, dementsprechend zu wenig Probabilistik in den Urteilen vorhanden. Sehr hohe Fitwerte verweisen also auf stark inkonsistentes Bewertungsverhalten; sehr niedrige Fitwerte hingegen können beispielsweise dadurch verursacht sein, dass Beurteiler die mittleren Antwortkategorien bevorzugt verwenden oder das Urteil mit einer Subgruppenzugehörigkeit von Probanden (etwa Geschlecht) interagiert.
Myford & Wolfe (2003, S. 409) geben für die Interpretation der Mean-Square-Statistiken im Rahmen von High-Stakes-Entscheidungen Richtwerte an. Danach sprechen Infit- bzw. Outfit-Werte im Intervall von 0.8 bis 1.2 für MFRM-Analysen für gute Modellpassungen der jeweiligen Facetten.

10.2.2.3 Separationsstatistiken

Eine hinreichend genaue Unterscheidungsfähigkeit zwischen den beurteilten Personen stellt eine notwendige Voraussetzung für die Beurteilung der Probandenfähigkeiten dar. Analog zur Beurteilung der Unterscheidungsfähigkeit im Rahmen des dichotomen Rasch-Modells 10.1.4 kann sie beim MFRM mit dem Separationsindex und der Separationsreliabilität beurteilt werden.

Der *Separationsquotient G* ist ein Maß für die Streubreite der Leistungsmaße relativiert an ihrer Genauigkeit (Wright & Masters, 1982) und berechnet sich daher als das Verhältnis von „wahrer“ Streuung der Leistungsmaße (d.h. der Streuung der Leistungsmaße nach Standardfehlerkorrektur; SD_t) zum mittleren Standardfehler der Leistungsmaße („Root Mean Square Error“; *RMSE*):

$$G = SD_T / RMSE \qquad (42)$$

Für die Beurteilerfacette gibt *G* an, wie reliabel sie sich anhand ihrer Strengemaße unterscheiden. Im Hinblick auf die Aufgaben indiziert *G* den Grad ihrer Unterscheidbarkeit anhand der Schwierigkeitsmaße. Je genauer die Messungen im Durchschnitt ausfallen (gemessen in

RMSE), desto mehr Klassen lassen sich innerhalb der jeweiligen Facette statistisch voneinander identifizieren.
Ein *G* von zwei würde z. B. hinsichtlich der Beurteilerfacette bedeuten, dass der Unterschied in der Strenge der Beurteiler zweimal größer als der Messfehler des Strengeparameters ist (siehe im übrigen die analogen Ausführungen hierzu in Kap. 10.1.4).

Der Index der *Separationsreliabilität R* ist definiert als Verhältnis von messfehlerkorrigierter zu beobachteter Varianz:

$$R = SD_t^2 / SD_x^2 = G / (1 + G) \qquad (43)$$

In Bezug auf die Beurteilerfacette zeigt die Separationsreliabilität das Ausmaß der Unterschiedlichkeit in den Strengemaßen an. *R* steht somit im Kontrast zur klassischen Interraterreliabilität, da diese ein Ähnlichkeitsindex des Urteilsverhaltens der Beurteiler darstellt, wohingegen die Separationsreliabilität erfasst, wie unähnlich sich diese hierin sind. Hohe Reliabilitätswerte stehen für eine ausgeprägte Unterschiedlichkeit innerhalb der Gruppe der Beurteiler. Im Falle der Probanden-Facette indiziert *R*, inwieweit das Beurteilungsverfahren Differenzierungen zwischen den Fähigkeitsmaßen gestattet, ist also wie bereits in Kapitel 10.1.4 als analoge Statistik zur internen Konsistenz nach Cronbachs Alpha zu interpretieren. Bezogen auf die Aufgaben-Facette zeigt *R* an, wie groß die Schwierigkeitsunterschiede ausfallen.

10.2.2.4 Homogenitätsstatistik

Zur inferenzstatistischen Absicherung der erlaubt ein Chi-Quadrat-Test (für feste Effekte) die Prüfung der Nullhypothese, dass die jeweiligen Stichprobenparameterschätzungen aus einer Population homogener Parameterwerte stammen. Für die Beurteilerfacette ist die entsprechende Statistik definiert über

$$Q = \sum_{j=1}^{J} \left(w_j \alpha_j^2 \right) - \left(\sum_{j=1}^{J} w_j \alpha_j \right)^2 \Bigg/ \sum_{j=1}^{J} w_j \qquad (44)$$

mit *J*-1 Freiheitsgraden, wobei *J*: Anzahl der Beurteiler.

Dabei ist $w_j = 1/SE^2_j$. SE_j Standardschätzfehler des Strengeparameters α_j.

Ein signifikanter *Q*-Wert zeigt an, dass sich mindestens zwei Beurteiler hinsichtlich ihrer Strengemaße überzufällig voneinander unterscheiden. Wie bei allen Signifikanztests, die auf der Chi-Quadrat-Verteilung basieren, besteht ein hohes Maß an Abhängigkeit der Teststatistik vom Stichprobenumfang. Bereits kleinste Abweichungen von der Homogenitätsannahme werden somit bei großen Stichproben als signifikant ausgewiesen.

Neben diesem Homogenitätsmaß lässt sich noch als rein deskriptives Maß der Übereinstimmung ein Prozentwert exakter Übereinstimmungen berechnen. Die klassische Interraterreliabilität (IRR) wird vom hier eingesetzten Programm FACETS (Linacre, 2005b) nicht direkt berechnet. FACETS gibt jedoch eine Multifacetten-Variante der Pearson-Korrelation an, die sogenannte „single rater-rest of the raters correlation“ (SR/ROR-Korrelation). Sie erfasst das Ausmaß an Übereinstimmung eines einzelnen Beurteilers mit denjenigen der übrigen Beurteiler. Da die Berechnung vergleichsweise komplex ist, kann für eine ausführliche Darstellung nur auf weiterführende Literatur verwiesen werden (s. z. B. Myford & Wolfe, 2003, S. 421f.). SR/ROR-Korrelationen kleiner als .30 können als gering angesehen werden, solche größer als .70 hingegen als hoch (Myford & Wolfe, 2003, S. 410).

10.2.2.5 Fairer Durchschnitt

Ein zentrales Ziel einer Multifacetten-Rasch-Analysen stellt eine möglichst genaue als auch möglichst faire Leistungsbeurteilungen dar. Leistungs- bzw. konstruktirrelevante Einflussfaktoren wie die Beurteilerstrenge oder –milde müssen daher quantifiziert werden, damit sie die endgültige Leistungsbeurteilung nicht beeinflussen bzw. nicht in systematischer Weise verzerren. Fairness würde sich hieraus ergeben, wenn für jeden Probanden dasjenige Rating ermittelt wird, das zustande käme, wenn der Proband von einem Beurteiler mit durchschnittlicher Strenge hinsichtlich einer durchschnittlich schweren Aufgabe und nach einem durchschnittlich schweren Kriterium beurteilt worden wäre. Eine Multifacetten-Rasch-Analyse berechnet daher für jedes Element jeder Facette eine erwartete Einstufung und zwar auf Grundlage der Durchschnitte der jeweils anderen Facette. Die hierüber berechnete erwartete mittlere Einstufung wird als *fairer Durchschnitt* bezeichnet. Seitens der Probanden gibt dieser, in der Metrik der Ratingskala diejenigen mittleren Einstufungen an, die unabhängig von der Strenge/Milde der Beurteiler, der Schwierigkeit der Aufgaben und derjenigen der Kriterien sind. Da allerdings diese, auf Rohwerten basierende Statistik im Gegensatz zu den Maßein-

heiten des Rasch-Modells, den Logitwerten, keine Intervallskalenqualität aufweist, werden üblicherweise für weitere Datenverarbeitungen die Logitwerte der Probandenfähigkeiten herangezogen.

Der Einfluss der Beurteilerstrenge kann nun kontrolliert werden, indem für jede beurteilte Person dasjenige Rating zu ermittelt wird, das resultieren würde, wäre sie sie von einem Beurteiler mit durchschnittlicher Strenge beurteilt worden. Aus Gleichung (36) lassen sich eine *adjustierte Wahrscheinlichkeit* vk (Gleichung (45)) und eine *adjustierte Beurteilung* v (Gleichung (46)) bestimmen:

$$\tilde{p}_{vk} = \frac{\exp\left[k\left(\theta - \bar{\alpha} - \bar{\sigma}\right) - \sum_{s=0}^{k} \tau_s \right]}{\sum_{r=0}^{m} \exp\left[r\left(\theta - \bar{\alpha} - \bar{\sigma}\right) - \sum_{s=0}^{k} \tau_s \right]}, \qquad (45)$$

wobei $\bar{\alpha}$ und $\bar{\sigma}$ die Mittelwerte der Schätzungen für die Strenge- bzw. den Schwierigkeitsparameter darstellen.

Die adjustierte Beurteilung *v* ergibt sich aus (45) durch

$$\tilde{x}_v = \sum_{r=0}^{m} r\tilde{p}_{vr.} \qquad (46)$$

Insbesondere die Notwendigkeit einer Objektivierung der Leistungsbeurteilung durch Kontrolle der Urteilsstrenge, wie überhaupt von Urteilsfehlern, macht das MFRM gerade für den Bereich von Segmenten in Auswahlverfahren *ohne* gebundenes Antwortformat attraktiv bzw. prädestiniert es hierfür. Seine Anwendung beschränkt sich selbstverständlicherweise nicht –wie im Falle der vorliegenden Arbeit– auf die Beurteilung schriftlich erbrachter Leistung, sondern könnte ebenso im Rahmen von teil- oder vollstrukturierten Interviews erfolgen. Darüber hinaus stellt es durch die Korrektur von *Einzel*urteilen um Beurteilerfehler eine notwendige Grundlage auch *rechtserheblicher* Auswahlentscheidungen dar. Ein Auswahlverfahren, welches erwiesenermaßen sich als anfällig gegenüber idiosynkratischen Bewertungsfehlern herausstellte, wäre nämlich bereits auf der rein rechtlichen Ebene invalidiert.

Wegen seiner Vorteile sowohl in theoretischer als auch praktischer Hinsicht stellte daher das MFRM auch in den in dieser Arbeit durchgeführten Analysen für alle Aufgaben im freien Antwortmodus die Auswertungsgrundlage dar.

Mit diesen Ausführungen schließt nun die Darstellung der Rasch-Modellgeltungsrationale dieser Arbeit. Die folgenden Abschnitte behandeln die empirischen Ergebnisse dieser Arbeit. Zunächst wird ein Überblick über die Zusammensetzung der Untersuchungsstichprobe gegeben, um darauf folgend die Ergebnisse der Modellgeltungstests zu berichten. Schließlich werden die Ergebnisse der Validitätsanalysen berichtet.

11. Stichprobe der Markmalsträger

An der Erhebung nahmen insgesamt 434 Studierende verschiedener Fachrichtungen der Universität Heidelberg im Zeitraum von Oktober 2004 bis April 2005 teil. Eine detaillierte Darstellung der Verteilung der Testteilnehmer auf die verschiedenen Studienfächer findet sich in Anhang K. Bei Teilnehmern anderer Studienfächer als Psychologie wurden, um die Vergleichbarkeit hinsichtlich der Testergebnisse zwischen den Fächern zu gewährleisten, nach Möglichkeit nur Studierende des ersten Semesters zum Wintersemester 2004 bei der Testpersonenrekrutierung angeworben. In sehr wenigen Fällen (s. Anhang L) ließ sich eine geringfügige Durchmischung mit Studierenden höherer Semester nicht vermeiden. In der Geschlechterverteilung der Gesamtstichprobe waren weibliche Studienteilnehmer mit 63.7% gegenüber 36.3% männlichen deutlich überrepräsentiert. Dies ist dadurch zu erklären, dass Studierende der Psychologie die größte Teilstichprobe darstellten und in diesem Studienfach der Frauenanteil generell größer ist. Der Mittelwert der Altersverteilung betrug 23.33 Jahre bei einer Standardabweichung von 4.11. Neben diesen demografischen Variablen wurden noch die Abiturdurchschnittsnote sowie die Abiturnoten in Mathematik, Deutsch und Englisch erfragt. Einen Überblick über deren deskriptive Statistiken gibt Tabelle 11.

Tabelle 11: Deskriptivstatistiken angegebener Abiturnoten in der Gesamtstichprobe

	N	***Minimum***	***Maximum***	***M***	***SD***
Abiturdurchschnitt	422	1.00	3.70	1.86	0.65
Mathematiknote	388	1.00	6.00	2.07	1.10
Deutschnote	379	1.00	5.00	1.93	0.87
Englischnote	351	1.00	6.00	1.87	0.93

Die unterschiedliche Anzahl vorhandener Abiturnoten, insbesondere der Fachnoten, resultierte daraus, dass diese freiwillig angegeben wurden und zudem manchen Teilnehmern nicht mehr in Erinnerung waren.

Die Studierenden der Psychologie stellten mit $n = 183$ die größte Teilstichprobe dar, weil diese Arbeit auf die Konstruktion und Evaluation eines psychologiespezifischen Auswahlinstrumentes abzielte. Da, wie schon in Kapitel 2 beschrieben, hierbei auch erste retrospektive Validitäten von Studierenden des Hauptstudiums mit zurückliegenden Vordiplomnoten gewonnen werden sollten, wurden auch aus dieser Gruppe Testteilnehmer angeworben. Die Gesamtaufteilung der Stichprobe im Diplomstudiengang Psychologie zeigt Tabelle 12

Tabelle 12: Semesterverteilung der Testteilnehmer im Fach Psychologie

Semester	**1**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**15**	**Gesamt**
Anzahl Testteilnehmer	81	27	1	24	1	15	1	11	5	7	6	2	2	183
Anzahl Testteilnehmer nach Grund-/ Hauptstudium	109			74										

Anmerkung. Semester ohne Testteilnehmer sind nicht aufgeführt

12. Studienteilnehmerrekrutierung und Testablauf

Um zum einen die individuellen Testergebnisse der Studienteilnehmer für eine spätere Evaluation über die Immatrikulationsnummer mit Studienleistungen in Beziehung setzen zu können, wurde diese unter schriftlicher Zusicherung der Verwendung der Daten nur im Rahmen einer Dissertation abgefragt. Über das Auskunftsrecht der Studienteilnehmer über die von ihnen abgespeicherten Daten über den Datenschutzbeauftragten der Universität Heidelberg informierte ein separates Informationsblatt vor Testbeginn.

Die Teilnehmerrekrutierung im Fach Psychologie erfolgte über verschiedene Wege. Sämtliche Teilnehmer des ersten Semesters in Psychologie wurden am ersten Tag der „Erstsemester--Kompaktseminar-Woche“ (EKS-Woche) mit Erklärungen über den Hintergrund der Studie sowie einer Anrechnung von zwei Versuchspersonenstunden bei vollständiger Testteilnahme

und der Möglichkeit individueller Testergebnisrückmeldung in Form von Prozenträngen zur freiwilligen Teilnahme angeworben. Bei dieser Testung musste aufgrund von Motivationsproblemen durch die Testlänge der Testteilnehmer bei der Bearbeitung des Inventars zur Messung der passiven argumentatorischen und rhetorischen Kompetenz (Flender, Christmann, Groeben & Mlynski, 1996) auf dessen zweites Szenario („Macht Fernsehen Aggressiv?“) verzichtet werden, was in Hinblick auf die nachfolgend zu bearbeitenden Teile des Gesamttests anders nicht zu lösen war. Die Rekrutierung der Studierenden höherer Semester des Grundstudiums und Studierender des Hauptstudiums fand sowohl über die Darstellung des Projektes in Lehrveranstaltungen statt als auch über Aushänge am Psychologischen Institut und Anwerbungen über E-Mail-Verteiler. Testpersonen aus dem dritten Semester wurden bei vollständiger Testteilnahme zwei Versuchspersonenstunden angerechnet und die individuelle Testrückmeldung angeboten. Die Datenerhebung der Substichprobe aus dem Hauptstudium erstreckte sich bis zum April 2005 und fand je nach Gruppengröße, die von 5 bis 15 Personen variierte, am Psychologischen Institut Heidelberg statt. Die Teilnehmer dieser Substichprobe erhielten für die Testteilnahme 15 €.

Die Anwerbung von Testpersonen der Fächer Biologie, Chemie, Ethnologie und Geographie erfolgte ebenfalls in der EKS-Woche dieser Fächer im Oktober 2004 mit Kurzdarstellung des Projektes und dessen Hintergründen durch den Autor sowie der Zusicherung von 15 € Versuchspersonengeld bei vollständiger Testteilnahme und, bei Interesse, individueller Ergebnisrückmeldung. Die Erhebung in diesen Fächern fand im November 2004 direkt in den Hörsälen der entsprechenden Institute statt, um die Schwelle einer Nicht-Teilnahme wegen zu langer Anfahrtswege möglichst niedrig zu halten.
Die Rekrutierung von Studienteilnehmern in den Fächern Rechtswissenschaften, Volkswirtschaftslehre und Soziologie erfolgte durch direkte Ansprache von Dozenten der jeweiligen Fakultäten und fand im Dezember 2004 bis Februar 2005 unter gleichen Vorgaben wie in den übrigen nicht-psychologischen Fächern unter Überlassung von Vorlesungsterminen durch Dozenten statt.

Alle Testungen wurden zur Wahrung der Durchführungsobjektivität vom Autor selbst durchgeführt. Die beiden Testheftversionen (Faking-Good-Instruktion vs. Normalinstruktion) wurden randomisiert auf die Testteilnehmer verteilt, indem beide Versionen vorab in abwechselnder Folge auf die Testbearbeitungsplätze verteilt wurden. Nach einführenden Worten über den Hintergrund der Evaluationsstudie wurde den Testteilnehmern vor jeder Durchführung der einzelnen Subtests die Instruktion vorgelesen, auf etwaige Rückfragen

eingegangen und schließlich nach Angabe der Testzeitbegrenzung der Test durchgeführt. Vor der Bearbeitung der Fragebögen wurden die Testanden noch einmal angehalten, sich die Instruktion sorgfältig durchzulesen und zu befolgen, um dann mit der Bearbeitung zu beginnen. Nach Testdurchführung erhielten die Probanden das Probandengeld in Höhe von 15 € mit dem Hinweis, bei Interesse an einer individuellen Ergebnisrückmeldung eine E-Mail an den Autor mit Angabe der Testheftnummer zu senden.

13. Datenaufbereitung und Datenverarbeitung

Die Dateneingabe der Multiple-Choice-Items sowie der Auswertung der freien Antworten der Aufgabe zum Denken in empirische Fragestellungen erfolgte durch den Autor und durch eine studentische Hilfskraft. Die Auswertung jeder der freien Antworten zum Denken in empirischen Fragestellungen erfolgte wie im Kapitel 8.1.3 beschrieben durch zwei Personen und unabhängig voneinander mithilfe des dargestellten Auswertungsschemas. Die Auswertung der Kreativitätsaufgaben aus dem Berliner Intelligenzstrukturtests (BIS) (Jäger, Süss & Beauducel, 1997) erfolgte in gleicher Weise durch den Autor und einer weiteren studentischen Hilfskraft nach Maßgabe des Auswertungsschemas aus dem Handbuch zum BIS.
Die statistische Datenanalyse wurde mit SPSS 12.0 (SPSS, 2004), die Skalenanalysen nach Maßgabe der Rasch-Modelle mit WINSTEPS 3.57 (Linacre, 2005d), WINMIRA 2001 (von Davier, 2001) und ConQuest 3.1 (Wu, Adams & Haldane, 2005) durchgeführt. Zur Analyse der Aufgaben mit freier Beantwortung (Aufgabe zum Denken in empirischen Fragestellungen und Kreativitätsaufgaben) nach dem Multifacetten-Rasch-Modell wurde das Programm Facets 3.59 (Linacre, 2005a) verwendet.

14. Ergebnisse

14.1 Skalenanalysen nach Rasch-Modellen

Im Folgenden werden die Ergebnisse der Skalenanalyse nach Rasch-Modellen beschrieben. Dieses betrifft auf der einen Seite Modellgeltungstests zum dichtotomen Rasch-Modell für alle Aufgaben, welche im Multiple-choice- Format vorlagen, zum anderen solche, im freien Antwortformat. Keiner Modellgeltung unterzogen werden hingegen alle Persönlichkeitsskalen. Wenn auch letzteres gerade auf dem Hintergrund einer Konstruktvalidierung unter einer Faking-Good-Bedingung gegenüber einer Normalinstruktion interessant gewesen wäre, so wurde eine Auswertung in dieser Richtung dennoch unterlassen, da die zentrale Fragestellung dieser Arbeit war, wie sehr die *Kriteriums*validitäten von Fragebogendaten hiervon beeinflusst werden.

14.1.1 Messfehlerbereinigte Testinterkorrelationen

Um für die Konstruktvalidierung einen ersten, rein deskriptiven Überblick der Zusammenhänge der sechs Subtests mit gebundenem Antwortformat (also ohne die Aufgaben zum empiriebezogenen Denken und zu Kreativitätsfacetten) zu erhalten, wurden die Interkorrelationen der Personenparameter nach dem Rasch-Modell über das Programm ConQuest 3.1 (Wu, Adams & Haldane, 2005) berechnet. Hierzu wurden alle Tests im Multiple-Choice-Format als sechsdimensionales Rasch-Modell modelliert, wobei jeder Subtest eine eigene Dimension definierte. Das eindimensionale Rasch-Modell wird dadurch zum Mehrdimensionalen Itemkomponenten Rasch-Modell (MULTIRA, s. Rost & Carstensen, 2002) verallgemeinert. Der Vorteil liegt hierbei in der Möglichkeit, durch eine simultane Testanalyse der sechs Subtests auch die bivariaten Verteilungen der Tests zu modellieren und so Schätzer für die Korrelationen der Testergebnisse zu erhalten. Zudem erhält man durch ConQuest über die Annahme einer bivariaten Normalverteilung der latenten Variablen *messfehlerbereinigte* Interkorrelationen der Testergebnisse (s. auch Rost, 2004, S. 263f.). Einen Überblick über die Ergebnisse dieser Analyse gibt Tabelle 13.

Tabelle 13: Messfehlerbereinigte Interkorrelationen der sechs Subtests im Multiple-Choice-Format (N = 434)

	A	O.O.O.	ZR	ZM	M
A					
O.O.O.	.64				
ZR	.61	.57			
ZM	.37	.30	.66		
M	.47	.55	.53	.37	
SPARK	.48	.31	.31	.19	.23

Anmerkung. A: verbale Analogien; O.O.O.: verbale Odd-One-Out;
ZR: Zahlenreihen; ZM: Zahlenmatrizen; M: Matrizen; SPARK: Skala zur Erfassung der passiven argumentativ-rhetorischen Kompetenz (gekürzte Form)

Alle Tests interkorrelieren positiv in mittlerer Höhe, bilden also die für Intelligenztests typische „positive manifold“. Die über Fishers *Z*-Transformation (s. Bortz, 1999, S. 210) ermittelte mittlere Korrelation fällt mit .44 moderat aus. Das vergleichsweise hohe Leistungsniveau der studentischen Stichprobe sowie der Umstand, dass es sich bei den Tests mit Ausnahme von SPARK um noch suboptimale Testvorversionen handelte, führten offenbar zu geringeren Zusammenhängen. Betrachtet man die Interkorrelationen der Subtests innerhalb eines gemeinsamen Intelligenzbereichs, so fallen diese höher als solche mit anderen Bereichen aus. Auffallend hierbei ist allerdings die vergleichsweise hohe Korrelation des Subtests Zahlenreihen mit beiden Subtests zur verbalen Intelligenz. Der Matrizentest als Maß fluider Intelligenz korreliert relativ am höchsten mit den Subtests verbale Odd-One-Out und Zahlenreihen. Erwartungsgemäß korreliert SPARK (Flender, Christmann, Groeben & Mlynski, 1996) als Test mit großen Verbalanteilen höher mit den Subtests zu verbalen als mit denjenigen zur numerischen und dem zur fluiden Intelligenz.

14.1.2 Modellgeltungstests

Im Folgenden werden die Ergebnisse der Modellgeltungstests nach dem dichotomen und dem Multifacetten Rasch-Modell berichtet. Hierbei werden stets zunächst die Ergebnisse der Itemhomogenität und anschließend diejenigen der Personenhomogenität dargestellt. Um eine

stringente und geraffte Darstellung der Ergebnisse zu ermöglichen, werden eingangs stets die Tabellen zu den Ergebnissen der Itemfit-Maße berichtet, damit sich der Leser zunächst selbst einen groben Überblick verschaffen kann. Im Anschluss daran werden detailliertere Erläuterungen hierzu gegeben.

14.1.2.1 Subtest „verbale Analogien“

Tabelle 14 gibt einen ersten Überblick über die wesentlichen deskriptiven Statistiken und Maße zur Modellgültigkeit der Items dieses Subtests. Zu detaillierten Erklärungen der Modellgültigkeitsindizes Infit und Outfit s. Kapitel 10.1.2.1.

Tabelle 14: Nach Modellpassung absteigend geordnete Itemstatistiken zum Subtest „verbale Analogien“ ($N = 434$)

Item-Position	**Item-name**	p_i	**Schwierig-keit** σ_i	**SE**	**MNSQ Infit**	**ZSTD Infit**	**MNSQ Outfit**	**ZSTD Outfit**
1	A1	.91	-3.24	0.18	1.08	0.7	1.60	**2.3**
2	A2	.30	0.49	0.11	1.17	**3.2**	1.28	**3.1**
6	A6	.50	-0.5	0.11	1.14	**3.6**	1.23	**4.0**
8	A8	.68	-1.43	0.11	1.08	1.5	1.20	**2.4**
3	A3	.50	-0.54	0.11	1.11	**3.0**	1.18	**3.2**
12	A12	.28	0.59	0.12	1.08	1.5	1.17	1.9
4	A4	.32	0.35	0.11	1.12	**2.5**	1.16	**2.0**
5	A5	.65	-1.25	0.11	1.12	**2.6**	1.15	**2.1**
7	A7	.74	-1.78	0.12	1.05	0.8	1.12	1.3
9	A9	.71	-1.56	0.11	1.03	0.6	1.09	1.0
13	A13	.24	0.87	0.12	1.00	0.0	0.94	-0.5
17	A17	.04	3.09	0.25	0.96	-0.1	0.60	-1.1
10	A10	.59	-0.98	0.11	0.94	-1.5	0.92	-1.3
14	A14	.30	0.49	0.11	0.90	**-2.0**	0.83	**-2.2**
20	A20	.07	2.45	0.20	0.89	-0.7	0.53	**-2.1**
18	A18	.09	2.17	0.18	0.87	-0.9	0.56	**-2.3**
19	A19	.11	1.92	0.16	0.84	-1.4	0.65	**-2.0**
11	A11	.77	-1.97	0.12	0.81	**-3.0**	0.69	**-3.1**
15	A15	.37	0.1	0.11	0.79	**-5.3**	0.73	**-4.5**
16	A16	.26	0.73	0.12	0.79	**-3.8**	0.68	**-3.6**
M		.42	0.00	0.13	0.99	0.0	0.97	0.0
SD			1.6	0.04	0.12	2.4	0.29	2.5

Anmerkung: p_i: Itemschwierigkeit nach der Klassischen Testtheorie; Schwierigkeit σ_i: Logit der Itemschwierigkeit; *SE*: Standardfehler des Itemparameters

Zunächst ist festzustellen, dass eine sehr gute Übereinstimmung zwischen den Rangreihen der Schwierigkeitsmaßen des Rasch-Modells und denjenigen der Klassischen Testtheorie besteht, wie die sehr hohe Rangkorrelation nach Kendals mit τ_b = -.99 belegt. Der Mittelwert der Itemschwierigkeit nach der Klassischen Testtheorie liegt mit .42 im mittleren Bereich. Im Rasch-Modell hingegen unterliegen die Itemparameter einer Summennormierung, um den frei wählbaren Nullpunkt der Differenzskala zu bestimmen, wodurch der Itemmittelwert stets gleich Null ist. Das Ausmaß der Modellabweichungen ist (quantifiziert über die MNSQ-Statistiken) bei einigen Items nicht besonders groß, gemessen an dem in dieser Arbeit für Testungen im High-Stakes-Bereich von Wright (1994) empfohlenen Toleranzbereich von 0.80 – 1.20. Allerdings werden die Items A14, A15, A16, A18, A19 und A20 sowohl nach Infit als auch Outfit bzw. deren z-standardisierten Werten als zu gut passend indiziert werden (sog. „Overfit“), da diese Werte *negativ* von ihren jeweiligen Erwartungswerten abweichen. Offenbar messen diese Items neben verbalem schlussfolgerndem Denken eine weitere Dimension, welche positiv mit dem Testgesamtergebnis korreliert ist. Betrachtet man nun die Position dieser Items, so fällt auf, dass sie im letzten Drittel des Subtests positioniert sind. Die Lösungswahrscheinlichkeit dieser Items hängt somit neben der Grundfähigkeit zusätzlich von der Bearbeitungsgeschwindigkeit ab, wodurch die Forderung nach lokaler stochastischer Unabhängigkeit der Itemantworten verletzt ist (Douglas, 2005; Van den Wollenberg, 1979, 1985). In der Folge kommt es hierdurch zu Überschätzungen der Itemparameter: Wären diese Items etwa zu Anfang des Tests positioniert gewesen, hätten die Schwierigkeitsschätzungen niedrigere Werte ergeben. Aufseiten der Personenparameter kann dies hingegen zu einer Unterschätzung führen, da Probanden, die nicht bis zur Bearbeitung der Items zum Ende des Tests gelangten, diese unter Power-Test-Bedingungen ggf. hätten lösen können. Offensichtlich wurde die Testzeit bei dieser Vorversion zu knapp angesetzt.

Der Mittelwert der Personenparameter liegt bei -0.54 bei einer Standardabweichung von 1.02. (M_{Rohwert} = 8.40, SD_{Rohwert} = 3.0). Der leicht negative Wert der Personenparameter zeigt an, dass dieser Subtests für die Probandenstichprobe insgesamt etwas zu schwer war. Die Reliabilitätskennwerte nach dem Rasch-Modell und der Klassischen Testtheorie (Kuder-Richardson-20, KR-20) zeigt Tabelle 15.

Tabelle 15: Personen- und Itemseparationsreliabilitäten Subtest „verbale Analogien“

	Separationsreliabilität *R*	**Separationsindex *G***
Personen	.66	1.40
Items	.99	11.49
KR-20	.68	

Sowohl nach der Separationsreliabililtät *R* als auch nach KR-20 ergeben sich Werte unterhalb des befriedigenden Bereiches für Leistungstests. Auch der Separationsindex *G* verweist auf eine lediglich mäßig gute Unterscheidbarkeit der Probanden anhand ihrer Fähigkeitswerte. Demgegenüber zeigt sich aufseiten der Items sowohl anhand der Itemseparationsreliabilität als auch nach dem Separationsindex eine sehr gute Verteilung der Schwierigkeitsunterschiede über das latente Kontinuum. Der Separationsindex *G* zeigt an, dass gut 11 statistisch voneinander distinkte Schwierigkeitsstufen zwischen den Items identifiziert werden können. Tabelle 16 gibt nun einen Überblick über die Modellvergleiche zur Überprüfung der Personenhomogenitätsannahme des Rasch-Modells.

Tabelle 16: Modellvergleiche anhand informationstheoretischer Maße und dem Likelihood-Quotiententests Subtest „verbale Analogien“

	BIC	**CAIC**	**Log L**	**-2(log(cL_{RM}) -log(cL_{2KL}))**
Dichotomes Rasch-Modell	8799.42	8820.42	-4335.94	444.9**
Zweiklassen mixed Rasch-Modell	8488.11	8531.11	-4113.49	

Anmerkung: Log L: Logarithmierte Likelihood der Daten unter den jwl. Modellannahmen; BIC: Bayes Information Criterion; CAIC: Consistent Akaikes Information Criterion; cL_{RM}: Bedingte Likelihood des dichotomen Rasch-Modells; cL_{2KL}: Bedingte Likelihood des mixed Rasch-Modells in zwei Klassen; -2(log(cL_{RM})-log(cL_{2KL})): Bedingter Likelihood-Quotiententest

Ein Vergleich des dichotomen Rasch-Modells mit der Zweiklassenlösung des mixed Rasch-Modells ergibt nach den informationstheoretischen Maßen BIC und CAIC für letztere eine bessere Passung, die sich zudem nach dem Likelihood-Quotiententest als signifikant erweist. Es besteht also in diesem Subtest eine bedeutsame Personenheterogenität. Die Größe

der ersten Klasse liegt hierbei bei 62%, diejenige der zweiten bei 38% der Gesamtstichprobe. Die mittlere Zuordnungswahrscheinlichkeit als Maß der Zuordnungsreliabilität liegt für die erste Klasse bei .97, für die zweite bei .94, kann also als sehr zuverlässig angesehen werden. Um einen Eindruck über die Art der Personenheterogenität zu erhalten, gibt Abbildung 10 die Schätzungen der Itemparameter in beiden der Klassen wieder.

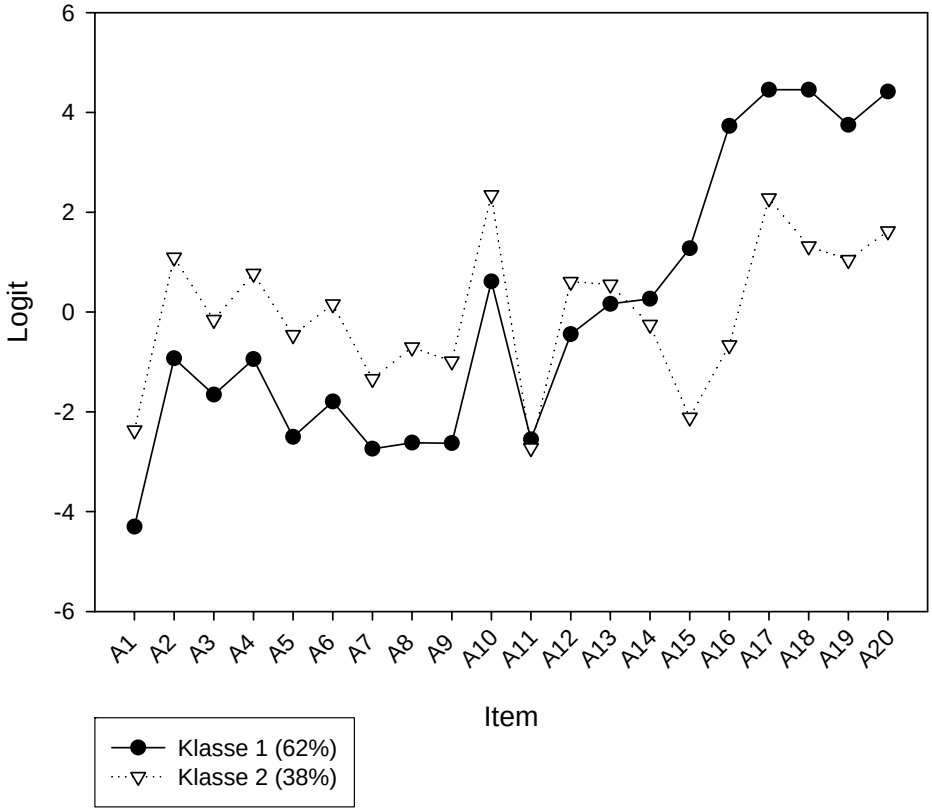


Abbildung 10: Itemparameter nach dem mixed Rasch-Modell in zwei Klassen für Subtest „verbale Analogien“ (N = 434)

Es ist bei dieser Form der Abbildung mit dem Verlauf der Itemschwierigkeiten wichtig zu betonen, dass hieraus keine Informationen über das Niveau zwischen den Personenklassen gezogen werden können! Auch im mixed Rasch-Modell wird eine Summennormierung der Itemparameter vorgenommen und zwar in jeder Klasse getrennt. Die Summe der Itemparameter ist daher stets in jeder Klasse gleich null. In ihrer mittleren Höhe können sich also die Itemparameter zwischen den Klassen nicht unterscheiden. Informationen über die mittlere Lösungswahrscheinlichkeiten der Items in jeder Klasse, also quantitative Unterschiede wie etwa „Könner“ gegenüber „Nicht-Könnern“, sind hier also nicht enthalten. Dafür lässt sich aus dieser Form der Abbildung ablesen, welche Aufgaben die Unterschiedlichkeit der beiden

Klassen *strukturell* ausmachen, etwa, wenn unterschiedliche Lösungsstrategien identifizierbar sein sollten.

Die Abbildung bestätigt die Schlussfolgerungen aus der Itemhomogenitätsanalyse. Man erkennt einen deutlich ausgeprägten Inangriffnahme-Effekt. Klasse 1 kann am besten als bearbeitungslangsamer, Klasse 2 hingegen als bearbeitungsschneller bezeichnet werden, da die Unterschiede in den Itemparametern zum Ende des Tests sehr deutlich werden. In Klasse 1 sind die Items A 16 bis A20 deutlich schwerer als in Klasse 2. Um der Frage nachzugehen, ob sich die Klassen auch in der Testgesamtleistung unterschieden, wurde für die Personenparameter ein t-Test für unabhängige Stichproben mit ungleichen Varianzen berechnet. Hier ergaben sich sehr hohe Effektstärken hinsichtlich der Überlegenheit der Klasse 2 (M_{PP} = 0.05, SD_{PP} = 0.66, M_{RW} = 10.06, SD_{RW} = 2.34) gegenüber der Klasse 1 (M_{PP} = -1.78, SD_{PP} = 0.90, M_{RW} = 6.59, SD_{RW} = 2.20) in der Subtestgesamtleistung: $t(418) = 24.24$, $p < .001$; Cohen's $d = 2.44$, $r_{ptbis} = .76$.
Vergleicht man die Itemparameter beider Klassen Abbildung 10 miteinander, so lässt sich allerdings ein leicht ausgeprägter Speed-Accuracy-Effekt (Lajoie & Shore, 1986) der Klasse 2 im Kontrast zur Klasse 1 feststellen, da hier die Items der ersten Testhälfte (A1 bis A11) *schwerer* ausfallen. Offensichtlich führte die höhere Bearbeitungsgeschwindigkeit gerade bei den Items zu Testanfang zu leichten Einbußen der Präzision.

Die aus zu geringer Testzeit resultierende Zweidimensionalität des Tests führt also, wenn die weitere Dimension Bearbeitungsgeschwindigkeit nicht zusätzlich erfasst wird, zu verfälschten Schätzungen nicht nur der Item-, sondern auch der Personenparameter. Bolt, Cohen und Wollack (2002) schlagen eine Auswertung von zeitlimitierten Tests mit Effekten einer Personenheterogenität nach dem mixed Rasch-Modell vor, da dadurch getrennte und verlässliche Schätzungen der Item- und Personenparameter möglich sind. In den hier durchgeführten Analysen hätte dies jedoch insbesondere für die Untersuchung der Kriteriumsvaliditäten getrennte Ergebnisse für beide Klassen bedeutet, was zum einen zulasten der Stringenz und der Interpretierbarkeit bzw. Verallgemeinerbarkeit der Ergebnisse geführt hätte. Zwar befinden sich die meisten von der Speed-Komponente betroffenen Items immer noch im Toleranzbereich von 0.8 – 1.2, allerdings ist die Anzahl dieser Items groß, um den Speed--Effekt unberücksichtigt zu lassen, zumal sich der Effekt in einer substanziellen Personenheterogenität niederschlägt. Daher wurde in der vorliegenden Arbeit der Weg über die Formulierung eines Rasch-Modells gegangen, welches die Speed-Komponente als potenziell relevante Einflussvariable neben der Grundfähigkeitskomponente bei der

Lösungswahrscheinlichkeit *kontrolliert*. Die Idee des Autors bestand darin, die Wahrscheinlichkeit einer korrekten Antwort als Linearkombination aus Personengrundfähigkeit, Itemschwierigkeit und einem Bearbeitungsgeschwindigkeits-Parameter zu modellieren. Um einen Parameter einer derartigen Speed-Komponente zu erhalten, wurde die Länge des Antwortvektors jeder Person, also die Anzahl in Angriff genommener Aufgaben, herangezogen. (Die Anzahl korrekter Antworten steht demnach für die Power-Komponente). Das somit spezifizierte Modell lautet in logarithmischer Schreibweise:

$$\log\left(p_{vi}\right) = \theta_v - \sigma_i - \delta_v, \qquad (47)$$

wobei θ_v: Personenfähigkeit einer Person v

σ_i: Schwierigkeit des Items i

δ_v: Speedparameter der Person v (Anzahl in Angriff genommener Items)

Die Parameter dieses Modells wurde mit dem Programm ConQuest 3.1 (Wu, Adams & Haldane, 2005) geschätzt. Tabelle 17 gibt einen Überblick über die Ergebnisse durch die Modell-Reformulierung.

Tabelle 17: Itemstatistiken im Subtest „verbale Analogien“ nach Modell-Reformulierung (*N* = 434) (Fortsetzung der Tabelle auf folgender Seite)

Itemname	**Item-schwierigkeit σ_i**	**SE**	**MNSQ Outfit**	**ZSTD Outfit**	**MNSQ Infit**	**ZSTD Infit**
A1	-2.70	0.13	1.30	1.1	1.05	0.4
A2	0.62	0.09	1.17	0.7	1.12	**2.6**
A3	-0.32	0.09	1.04	0.3	1.04	1.4
A4	0.45	0.09	1.11	0.5	1.08	1.5
A5	-0.96	0.09	1.11	0.5	1.08	1.5
A6	-0.28	0.09	1.14	0.6	1.12	**4.6**
A7	-1.44	0.09	1.07	0.3	1.03	0.5
A8	-1.18	0.09	1.12	0.5	1.07	1.6
A9	-1.24	0.09	1.04	0.2	1.02	0.4
A10	-0.74	0.09	1.01	0.1	1.00	-0.1
A11	-1.72	0.10	0.81	-0.7	0.87	-1.9

Itemname	**Item-schwierigkeit σ_i**	**SE**	**MNSQ Outfit**	**ZSTD Outfit**	**MNSQ Infit**	**ZSTD Infit**
A12	0.68	0.09	0.99	0.1	1.01	0.2
A13	0.87	0.10	0.92	-0.3	0.95	-0.8
A14	0.49	0.09	0.90	-0.3	0.93	-1.9
A15	0.16	0.09	0.79	-0.8	0.82	**-6.8**
A16	0.64	0.10	0.79	-0.9	0.83	**-4.2**
A17	2.06	0.16	0.92	-0.2	0.98	-0.1
A18	1.65	0.13	0.77	-0.9	0.89	-1.0
A19	1.29	0.13	0.82	-0.6	0.87	-1.6
A20	1.66	0.47	0.84	-0.5	0.93	-0.6
Item Separations-Reliabilität R = .99						

Anmerkung: Itemschwierigkeit σ_i: Logit der Itemschwierigkeit; *SE*: Standardfehler des Itemparameters; MNSQ Outift: Meansquare-Outfit-Statistik; MNSQ Infit: Meansquare Infit-Statistik; ZSTD Outfit: z-standardisierte MNSQ-Outfit-Statistik; ZSTD Infit: z-standardisierte MNSQ-Infit-Statistik

Man erkennt eine deutlich verbesserte Item-Modell-Passung durch die zusätzliche Spezifizierung einer Speed-Komponente. Die Item-Separationsreliabilität liegt mit .99 weiterhin sehr hoch und zeigt eine sehr gute Schwierigkeitsstreuung der Items über die latente Variable an. Die in der Ergebnisausgabe fett markierten Items zeigen weiterhin signifikante Modellabweichungen. Diese Items wurden daher für eine weitere Analyse nach Modellgleichung (47) entfernt und die verbleibenden Items rekalibriert. Die resultierenden Itemfit--Werte zeigten allesamt keine signifikanten Abweichungen. Der größte positive z-standardisierte Infit-Wert ergab sich für Item A8 mit z_{Infit} = 1.8, der größte negative für Item A19 mit z_{Infit} = -1.7. Daher wurde die Modellgültigkeit der verbleibenden 16 Items für das Modell nach Gleichung (47) angenommen. Aus der Reduktion der Skala resultierte eine interne Konsistenz nach KR-20 von α = .62.

14.1.2.2 Subtest „Odd-One-Out verbal"

Tabelle 18 gibt einen ersten Überblick über die wesentlichen deskriptiven Statistiken und Maße zur Modellgültigkeit der Items dieses Subtests.

Tabelle 18: Nach Modellpassung absteigend geordnete Itemstatistiken zum Subtest „Odd-One-Out verbal“ (N = 434)

Item-Position	Item-name	p_i	Schwierigkeit σ_i	SE	MNSQ Infit	ZSTD Infit	MNSQ Outfit	ZSTD Outfit
1	O21	.29	1.83	0.11	1.17	**3.2**	2.05	**9.5**
8	O28	.90	-1.71	0.17	1.07	0.6	1.31	1.5
4	O24	.37	1.41	0.11	1.07	1.7	1.18	**2.6**
11	O31	.54	0.61	0.10	1.10	**3.1**	1.12	**2.4**
9	O29	.52	0.71	0.10	1.05	1.7	1.07	1.4
14	O34	.82	-0.90	0.13	0.99	-0.1	1.06	0.5
13	O33	.65	0.10	0.11	1.01	0.2	1.04	0.7
7	O27	.76	-0.52	0.12	1.02	0.4	1.04	0.5
5	O25	.30	1.77	0.11	1.03	0.6	1.03	0.4
3	O23	.76	-0.48	0.12	1.03	0.5	1.01	0.1
10	O30	.55	0.58	0.10	0.98	-0.7	0.98	-0.4
18	O38	.18	2.53	0.13	0.96	-0.5	0.89	-0.8
15	O35	.53	0.68	0.10	0.95	-1.7	0.91	-1.8
19	O39	.20	2.41	0.13	0.94	-0.8	0.83	-1.4
20	O40	.32	1.66	0.11	0.90	**-2.2**	0.85	**-2.0**
17	O37	.56	0.52	0.10	0.89	**-3.2**	0.88	**-2.5**
2	O22	.99	-3.91	0.42	0.88	-0.2	0.77	-0.2
6	O26	.99	-4.1	0.46	0.86	-0.2	0.67	-0.3
16	O36	.79	-0.67	0.12	0.85	**-2.2**	0.79	**-2.2**
12	O32	.95	-2.53	0.23	0.85	-0.7	0.64	-1.3
M		.60	0.00	0.16	0.98	0.0	1.01	0.3
SD			1.85	0.10	0.09	1.6	0.29	2.5

Anmerkung. p_i: Itemschwierigkeit nach der Klassischen Testtheorie; Schwierigkeit σ_i: Logit der Itemschwierigkeit; *SE*: Standardfehler des Itemparameters

Wie bereits im Subtest „verbale Analogien“ besteht eine sehr gute Übereinstimmung der Rangreihen nach Schwierigkeitswerten der Klassischen Testtheorie und nach den Itemparametern des Rasch-Modells mit τ_b = -.99. Der mittlere *p*-Wert von .60 weist die Items insgesamt als moderat schwer für die vorliegende Stichprobe aus. Eine signifikant zu schlechte Modellpassung zeigen die Items O21, O24 und O31. Ähnlich, wenn auch im Ausmaß geringer als im Subtest verbale Analogien, zeigt sich ein Speed-Effekt, sodass die Items gegen Ende des Tests hin (ab Item O36) den Infit- und Outfit-Maßen höhere Diskriminationsleistungen zeigen. Diese Items werden daher vom Rasch-Modell als zu trennscharf indiziert, da ihre Lösungswahrscheinlichkeit zusätzlich von der Bearbeitungsgeschwindigkeit mitbedingt wird.
Der Mittelwert der Personenparameter lag bei 0.80 bei einer Standardabweichung von 0.84. ($M_{Rohwert}$ = 12.00, $SD_{Rohwert}$ = 5.34). Der leicht positive mittlere Logitwert zeigt, dass dieser

Subtest etwas zu leicht für diese Probandenstichprobe war. Dieser Eindruck wird auch durch die Personenseparationsstatistiken in Tabelle 19 bestätigt.

Tabelle 19: Personen- und Itemseparationsreliabilitäten Subtest "Odd-One-Out verbal"

	Separationsreliabilität *R*	**Separationsindex *G***
Personen	.60	1.28
Items	.99	9.98
KR-20	.61	

Sowohl die Personenseparationsreliabilität *R*, der Separationsindex *G* als auch KR-20 verweisen auf eine eher geringe Personendifferenzierungsfähigkeit des Subtests. Die Streuung der Itemschwierigkeiten über das latente Kontinuum hingegen liegt nach beiden Itemseparationsstatistiken im sehr guten Bereich. Insbesondere der Separationsindex *G* zeigt an, dass knapp 10 statistisch voneinander distinkte Schwierigkeitsstufen zwischen den Items identifiziert werden können.
Tabelle 20 stellt die Ergebnisse der Personenhomogenitätsüberprüfung dar.

Tabelle 20: Modellvergleiche anhand informationstheoretischer Maße und dem Likelihood-Quotiententests für Subtest „Odd-One-Out verbal"

	BIC	**CAIC**	**Log L**	**$-2(\log(cL_{RM})-\log(cL_{2KL}))$**
Dichotomes Rasch-Modell	8933.14	8954.14	-4402.80	23.94 n.s.
Zweiklassen mixed Rasch-Modell	9042.67	9085.67	-4390.83	

Anmerkung: Log L: Logarithmierte Likelihood der Daten unter den jwl. Modellannahmen; BIC: Bayes Information Criterion; CAIC: Consistent Akaikes Information Criterion; cL_{RM}: Bedingte Likelihood des dichotomen Rasch-Modells; cL_{2KL}: Bedingte Likelihood des mixed Rasch-Modells in zwei Klassen; $-2(\log(cL_{RM})-\log(cL_{2KL}))$: Bedingter Likelihood-Quotiententest

Hinsichtlich der Personenhomogenität nach Tabelle 20 zeigt sich nach den informations-theoretischen Maßen und dem Log-Likelihood-Test keine bessere Passung der Zweiklassen--Lösung des mixed Rasch-Modells gegenüber dem dichotomen Rasch-Modell. Offensichtlich

wirkt sich der im Kontext von Tabelle 18 beschriebene leichte Speed-Effekt nicht gravierend auf die Personenhomogenität aus.

Für die Endform des Subtests für diese Arbeit wurden daher alle sechs nach Tabelle 18 ausgewiesenen signifikant vom Modell abweichenden Items eliminiert (O21, O24, O31, O36, O37 und O40) und eine erneute Analyse nach dem Rasch-Modell vorgenommen. Nach dieser Analyse zeigte Item O35 einen signifikanten Overfit nach den z-standardisierten Infit- und Outfit-Werten (z_{Infit} = -1.98 bzw. $z_{Outfit.}$= 2.07) und wurde für eine weitere Kalibrierung eliminiert. Danach zeigte kein Item mehr eine signifikante Abweichung von den Modellerwartungen. Die größten Abweichungen hatte Item O29 mit z_{Infit} = 1.89 und z_{Outfit} = 1.60. Die Modellgültigkeit der somit verbleibenden 13 Items wird angenommen. Durch die Itemselegierung sank KR-20 von α = .61 auf α = .57 ab, da (wie bereits im Subtest „Verbale Analogien“) nach der Klassischen Testtheorie trennstärkere Items selegiert wurden, welche jedoch das Rasch-Modell als zu trennscharf identifizierte.

14.1.2.3 Subtest „Zahlenreihen“

Tabelle 21 stellt die Resultate der wesentlichen deskriptiven Statistiken und Maße zur Modellgültigkeit der Items dieses Subtests dar.

Tabelle 21: Nach Modellpassung absteigend geordnete Itemstatistiken zum Subtest „Zahlenreihen“ (N = 434) (Fortsetzung der Tabelle auf folgender Seite)

Item-Position	**Item name**	p_i	**Schwierigkeit** σ_i	**SE**	**MNSQ Infit**	**ZSTD Infit**	**MNSQ Outfit**	**ZSTD Outfit**
1	ZR41	.91	-3.24	0.18	1.08	0.7	1.6	**2.3**
2	ZR42	.30	0.49	0.11	1.17	**3.2**	1.28	**3.1**
6	ZR46	.50	-0.50	0.11	1.14	**3.6**	1.23	**4.0**
8	ZR48	.68	-1.43	0.11	1.08	1.5	1.2	**2.4**
3	ZR43	.50	-0.54	0.11	1.11	**3.00**	1.18	**3.2**
12	ZR52	.28	0.59	0.12	1.08	1.5	1.17	1.9
4	ZR44	.32	0.35	0.11	1.12	**2.5**	1.16	**2.0**
5	ZR45	.65	-1.25	0.11	1.12	**2.6**	1.15	**2.1**
7	ZR47	.74	-1.78	0.12	1.05	0.8	1.12	1.3
9	ZR49	.71	-1.56	0.11	1.03	0.6	1.09	1.0
13	ZR53	.24	0.87	0.12	1.00	0.00	0.94	-0.5
17	ZR57	.04	3.09	0.25	0.96	-0.1	0.60	-1.1
10	ZR50	.59	-0.98	0.11	0.94	-1.5	0.92	-1.3
14	ZR54	.30	0.49	0.11	0.9	**-2.0**	0.83	**-2.2**
20	ZR60	.07	2.45	0.20	0.89	-0.7	0.53	**-2.1**

Item-Position	Item name	p_i	Schwierigkeit σ_i	SE	MNSQ Infit	ZSTD Infit	MNSQ Outfit	ZSTD Outfit
18	ZR58	.09	2.17	0.18	0.87	-0.9	0.56	**-2.3**
19	ZR59	.11	1.92	0.16	0.84	-1.4	0.65	**-2.0**
11	ZR51	.77	-1.97	0.12	0.81	**-3.00**	0.69	**-3.1**
15	ZR55	.37	0.10	0.11	0.79	**-5.3**	0.73	**-4.5**
16	ZR56	.26	0.73	0.12	0.79	**-3.8**	0.68	**-3.6**
M		0.46	0.00	0.13	0.99	0.00	0.97	0.0
SD			1.6	0.04	0.12	2.4	0.29	2.5

Anmerkung: p_i: Itemschwierigkeit nach der Klassischen Testtheorie; Schwierigkeit σ_i: Logit der Itemschwierigkeit; *SE*: Standardfehler des Itemparameters

Für die Items des Subtests Zahlenreihen ergibt sich wiederum eine sehr gute Übereinstimmung hinsichtlich der Rangreihe nach den „klassischen" Schwierigkeitsindizes und den Itemparametern des Rasch-Modells mit τ_b = -.99. Mit einem durchschnittlichen Schwierigkeitsindex von .43 erwies sich der Test für diese Stichprobe als mittelschwer. Der Mittelwert der Personenparameter lag bei -0.54 bei einer Standardabweichung von 0.94. ($M_{Rohwert}$ = 8.40, $SD_{Rohwert}$ = 5.9). Somit war der Test tendenziell etwas zu schwer für diese Probandenstichprobe. Wie bereits im Subtest „Verbalen Analogien" tritt wiederum ein prägnanter Speed-Effekt der Items aus dem letzten Testdrittel hervor. Die Items ZR54, ZR55, ZR56, ZR58, ZR59 und ZR60 zeigen mit signifikant negativen Abweichungen einen Overfit und somit eine Verletzung der lokalen stochastischen Unabhängigkeit der Itemantworten an. Tabelle 22 gibt einen Überblick über die Separationsstatistiken der Personen- und Item-Ebene.

Tabelle 22: Personen- und Itemseparationsreliabilitäten Subtest „Zahlenreihen"

	Separationsreliabilität *R*	**Separationsindex *G***
Personen	.78	2.00
Items	.99	11.29
KR-20	.80	

Die Personenseparationsstatistiken nach dem Rasch-Modell und nach KR-20 liegen im befriedigenden Bereich. Die Separationsstatistiken auf Itemebene hingegen können als sehr gut bezeichnet werden, insbesondere zeigt der Separationsindex *G* eine sehr hohe Streuung der

Items über die latente Variable an. Hiernach können gut 11 statistisch voneinander verschiedene Schwierigkeitsstufen innerhalb der Itemgruppe identifiziert werden. Tabelle 23 gibt einen Überblick über die Prüfung der Personenhomogenität.

Tabelle 23: Modellvergleiche anhand informationstheoretischer Maße und dem Likelihood-Quotiententests Subtest „Zahlenreihen“

	BIC	**CAIC**	**Log L**	**-2(log(cL_{RM})-log(cL_{2KL}))**
Dichotomes Rasch-Modell	7963.19	7984.19	-3917.83	231.1**
Zweiklassen mixed Rasch-Modell	7865.70	7908.70	-3802.28	

Anmerkung: Log L: Logarithmierte Likelihood der Daten unter den jwl. Modellannahmen; BIC: Bayes Information Criterion; CAIC: Consistent Akaikes Information Criterion; cL_{RM}: Bedingte Likelihood des dichotomen Rasch-Modells; cL_{2KL}: Bedingte Likelihood des mixed Rasch-Modells in zwei Klassen; -2(log(cL_{RM})-log(cL_{2KL})): Bedingter Likelihood-Quotiententest

Beide informationstheoretischen Maße weisen die Zweiklassenlösung als besser passend als das dichotome Rasch-Modell aus. Die Differenz des Log-Likelihood-Tests ist zudem signifikant, was auf eine substanzielle Personenheterogenität verweist. Die mittlere Zuordnungswahrscheinlichkeit war zudem für Klasse 1 mit .94 und für Klasse 2 mit .91 sehr reliabel. Ein t-Test für unabhängige Stichproben mit ungleichen Varianzen für die Personenparameter ergab sehr hohe Effektstärken bezüglich der Überlegenheit der Klasse 1 (M_{PP} = 0.98,
SD_{PP} = 0.95, M_{RW} = 13.24, $SD_{Rohwert}$ = 2.49) gegenüber Klasse 2 (M_{PP} = -0.50, SD_{PP} = 1.10, M_{RW} = 9.32, SD_{RW} = 2.30) in der Subtestgesamtleistung: $t(350.83) = 14.65$, $p < .001$; Cohen’s d = 1.58, r_{ptbis} = .61. Abbildung 11 gibt einen Einblick in die Art der Personenheterogenität.

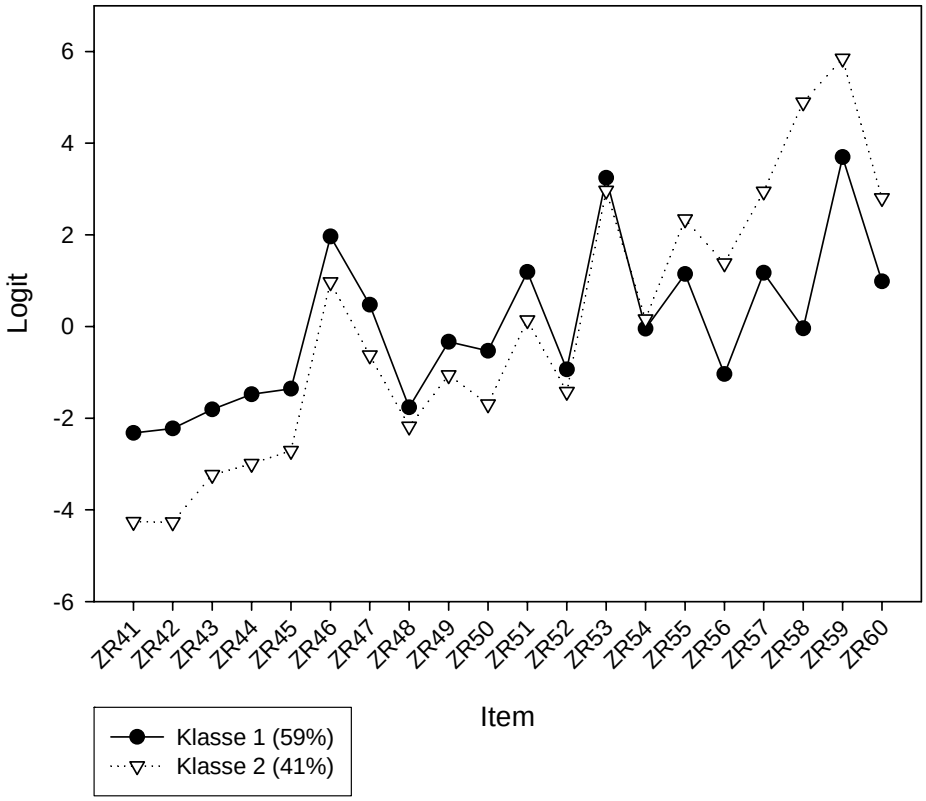


Abbildung 11: Itemparameter nach dem mixed Rasch-Modell in zwei Klassen für Subtest „Zahlenreihen“

Wiederum ergibt sich die Personenheterogenität durch zwei Personenklassen, die sich hinsichtlich ihrer Bearbeitungsgeschwindigkeit unterscheiden. Anders als beim Subtest Analogien handelt es sich bei Klasse 1, der größeren von beiden Klassen, um die bearbeitungsschnellere Gruppe, Klasse 2 hingegen stellt die bearbeitungslangsamere Gruppe dar. Auch zeigt sich bei Klasse 1 ein tendezieller Speed-accuracy-trade-off für die ersten fünf Items.

Um die Speed-Komponente bei der Beurteilung der Personenfähigkeit mit zu beachten, wurde abermals das Modell nach Gleichung (47) angewendet, weil die Menge der von der Speed--Komponente betroffenen Items zu groß war, auch wenn deren MNSQ-Fit-Statistiken überwiegend im Toleranzbereich von 0.8 - 1.2 lagen. Die Ergebnisse der Itemfitmaße zum reformulierten Modell zeigt Tabelle 24.

Tabelle 24: Itemstatistiken im Subtest „Zahlenreihen“ nach Modell-Reformulierung ($N = 434$)

Itemname	**Item-schwierigkeit σ_i**	**SE**	**MNSQ Outfit**	**ZSTD Outfit**	**MNSQ Infit**	**ZSTD Infit**
ZR1	-2.65	0.10	1.44	1.5	1.04	0.3
ZR2	-2.41	0.10	1.45	1.5	1.04	0.2
ZR3	-2.23	0.10	1.43	1.5	1.02	0.2
ZR4	-1.91	0.09	1.16	0.6	1.03	0.3
ZR5	-1.79	0.09	1.38	1.3	1.06	0.6
ZR6	1.75	0.08	1.10	0.5	1.02	0.5
ZR7	0.15	0.08	1.33	1.2	1.26	**6.1**
ZR8	-1.77	0.09	0.92	-0.2	0.95	-0.4
ZR9	-0.54	0.08	1.05	0.3	1.02	0.4
ZR10	-0.99	0.08	1.15	0.7	1.06	0.9
ZR11	0.93	0.08	1.31	1.2	1.24	**5.8**
ZR12	-1.04	0.09	1.05	0.3	0.99	-0.2
ZR13	3.26	0.10	0.98	0.0	0.96	-0.2
ZR14	0.13	0.08	0.89	-0.4	0.91	-1.8
ZR15	1.46	0.08	0.87	-0.5	0.91	-1.7
ZR16	0.18	0.08	0.71	-1.2	0.73	**-7.5**
ZR17	1.57	0.08	0.96	-0.1	0.98	-0.4
ZR18	0.82	0.08	0.70	-1.3	0.72	**-8.5**
ZR19	3.63	0.10	0.91	-0.3	0.97	-0.1
ZR20	1.46	0.38	0.89	-0.4	0.93	-1.5

Item Separationsreliabilität $R = .99$

Anmerkung: Itemschwierigkeit σ_i: Logit der Itemschwierigkeit;
SE: Standardfehler des Itemparameters; MNSQ Outift: Meansquare-Outfit-Statistik; MNSQ Infit: Meansquare Infit-Statistik; ZSTD Outfit: z-standardisierte MNSQ-Outfit-Statistik; ZSTD Infit: z-standardisierte MNSQ-Infit-Statistik

Es resultiert durch die Reformulierung eine deutlich bessere Passung der Items unter Beachtung des Speed-Parameters. Die Item-Separationsreliabilität liegt mit .99 weiterhin sehr hoch und indiziert eine sehr gute Schwierigkeitsstreuung der Items über die latente Variable. In einer weiteren Analyse wurden die in Tabelle 24 mit einem signifikanten Over- oder Underfit ausgewiesenen Items eliminiert (in der Tabelle fett markiert) und die übrigen Items erneut einer Analyse nach dem reformuliertem Rasch-Modell unterworfen. Danach zeigte Item A14 einen signifikanten Overfit nach der Infit-Statistik ($MNSQ_{Infit} = 0.90$, $z_{Infit} = -2.7$). Auch wenn die Größe der Modellabweichung nach dem MNSQ-Wert nicht groß ist, wurde dieses Item

zwecks weiterer Verbesserung der Skala in einer folgenden Analyse ausgeschlossen. Hiernach zeigte kein Item eine signifikante Modellabweichung. Die relativ größte Modellabweichung zeigte Item A15 mit $MNSQ_{Infit} = 0.92$, $z_{Infit} = 1.69$ bzw. $MNSQ_{Outfit} = 0.89$, $z_{Outfit} = 1.14$. Die Modellgültigkeit der somit verbleibenden 14 Items wurde daher angenommen. Die interne Konsistenz nach KR-20 der verkürzten Skala stieg von $\alpha = .80$ auf $\alpha = .85$ an, da in dieser Skala mehr Items mit signifikantem Underfit selegiert wurden.

14.1.2.4 Subtest „Zahlenmatrizen“

Tabelle 25 stellt die Resultate der wesentlichen deskriptiven Statistiken und Maße zur Modellgültigkeit der Items des Subtests „Zahlenmatrizen“ dar.

Tabelle 25: Nach Modellpassung absteigend geordnete Itemstatistiken zum Subtest „Zahlenmatrizen“ ($N = 434$)

Item-Position	**Item-name**	p_i	**Schwierigkeit** σ_i	**SE**	**MNSQ Infit**	**ZSTD Infit**	**MNSQ Outfit**	**ZSTD Outfit**
2	ZM62	.88	-4.12	0.18	0.93	-0.5	2.52	**2.8**
7	ZM67	.01	5.03	0.61	1.23	0.6	2.41	1.2
18	ZM78	.03	3.18	0.30	0.99	0.0	2.06	1.3
12	ZM72	.20	0.6	0.14	1.13	1.6	1.73	**3.1**
14	ZM74	.10	1.7	0.18	1.09	0.8	1.67	1.7
8	ZM68	.80	-3.23	0.14	1.15	1.8	1.53	1.8
15	ZM75	.30	-0.17	0.12	0.94	-1.1	1.52	**3.3**
3	ZM63	.86	-3.81	0.16	1.01	0.2	1.42	1.2
1	ZM61	.76	-2.93	0.13	1.03	0.5	1.42	1.7
13	ZM73	.18	0.78	0.14	0.94	-0.6	1.21	1.0
9	ZM69	.50	-1.34	0.11	1.00	0.0	1.15	1.3
10	ZM70	.11	1.5	0.17	0.94	-0.5	1.12	0.5
11	ZM71	.04	2.78	0.26	1.11	0.6	0.79	-0.2
16	ZM76	.28	-0.01	0.12	0.98	-0.3	0.94	-0.4
6	ZM66	.73	-2.7	0.13	0.88	**-2.0**	0.76	-1.2
4	ZM64	.70	-2.47	0.12	0.87	**-2.2**	0.82	-1.0
5	ZM65	.61	-1.94	0.12	0.82	**-3.8**	0.8	-1.5
17	ZM77	.03	3.09	0.29	0.73	-1.2	0.2	-1.5
20	ZM80	.02	4.04	0.41	0.71	-0.8	0.46	-0.2
M		0.38	0.00	0.28	0.97	-0.4	1.29	0.8
SD			3.14	0.38	0.13	1.3	0.6	1.5

Anmerkung: p_i: Itemschwierigkeit nach der Klassischen Testtheorie; Schwierigkeit σ_i: Logit der Itemschwierigkeit; *E*: Standardfehler des Itemparameters

In die Skalenanalyse des Subtests Zahlenmatrizen gingen 19 der ursprünglich 20 Items ein, weil Item ZM19 von keiner Person gelöst werden konnte. Dieses Item wäre in einer Testrevision folglich durch ein leichteres zu ersetzen. Für die übrigen Items zeigt sich wiederum eine sehr gute Übereinstimmung der Rangfolge „klassischer“ Schwierigkeitswerte mit derjenigen nach den Itemparametern des Rasch-Modells mit τ_b = -.99. Die durchschnittliche Itemschwierigkeit nach dem „klassischen“ Schwierigkeitsindex liegt mit .38 im mittleren Bereich. Der Mittelwert der Personenparameter zeigt mit einem Wert von -1.32 bei einer Standardabweichung von 1.60 an, dass dieser Test insgesamt für die Probandenstichprobe etwas zu schwer war (M_{Rohwert} = 7.10, SD_{Rohwert} = 5.36). Hinsichtlich der Itempassungen zeigen sich deutliche Abweichungen. Dies betrifft besonders die Items ZR64, ZR65 und ZR66, welche nach dem Infit-Maß des Rasch-Modells als signifikant zu trennscharf ausgewiesen werden. Dies kann jedoch kein Effekt der Testzeitbegrenzung sein, da diese Items alle zu Beginn des Subtests platziert sind. Eine Erklärung für diesen Verstoß gegen die lokale stochastische Unabhängigkeit der Itemantworten muss also auf anderer Ebene gesucht werden. Die folgende Betrachtung dieser Items und ihrer jeweiligen Lösungsregel bringt dahingehend Klarheit.

ZM64

?	**33**	**8**
48	**42**	**5**
22	**12**	**9**

*Regel: (33+8)**+1** = 42*

ZM65

67	**12**	**?**
78	**8**	**83**
54	**13**	**64**

*Regel: (67+12)**-3** = 76*

ZM66

77	**89**	**33**
3	**4**	**8**
81	**94**	**?**

*Regel: (33**+1**)+8 = 42*

Analysiert man die drei Lösungswege nach Gemeinsamkeiten, so fällt auf, dass zur Aufgabenlösung jeweils eine nicht in der Zahlenmatrix vorhandene Zahl ergänzt werden muss (in den Lösungregeln jwls. fett markiert). Da zudem die Items unmittelbar aufeinanderfolgen, ist es für Probanden, die einmal dieses Schema durchschaut haben ein leichtes, dieses zur nächsten Aufgabe zu übertragen. Der Lösungsweg des einen Items bietet also Hinweise auf den des folgenden. Vor einer Itemselektion muss allerdings zunächst das Ergebnis der Personenhomogenitätsanalyse betrachtet werden.
Einen Überblick über die Personen- und Itemseparationsstatistiken liefert Tabelle 26.

Tabelle 26: Personen- und Itemseparationsreliabilitäten Subtest „Zahlenmatrizen“

	Separationsreliabilität *R*	**Separationsindex *G***
Personen	.75	1.64
Items	.99	11.03
KR-20	.78	

Die Personenseparationsstatistiken nach dem Rasch-Modell und der KR-20 verweisen auf eine befriedigende Unterscheidbarkeit der Probanden anhand ihrer Fähigkeitswerte. Gleichwohl müsste diese für Selektionszwecke über eine Testrevision erhöht werden. Auf Itemebene resultiert eine sehr gute Streuung der Items und somit ihrer Funktion als Skala über einen weiten Bereich der zu messenden Variable. Die Itemseparationsreliabilität liegt sehr hoch und nach dem Separationsindex *G* können hier gut 11 statistisch voneinander distinkte Schwierigkeitsstufen innerhalb der Items identifiziert werden.
Tabelle 27 gibt die Ergebnisse zur Überprüfung der Personenhomogenitätsannahme wieder.

Tabelle 27: Modellvergleiche anhand informationstheoretischer Maße und dem Likelihood-Quotiententests Subtest „Zahlenmatrizen"

	BIC	**CAIC**	**Log L**	**$-2(\log(cL_{RM})-\log(cL_{2KL}))$**
Dichotomes Rasch-Modell	6283.10	6304.10	-3077.78	4.76 n.s.
Zweiklassen mixed Rasch-Modell	6271.94	6314.94	-3005.40	

Anmerkung: Log L: Logarithmierte Likelihood der Daten unter den jwl. Modellannahmen; BIC: Bayes Information Criterion; CAIC: Consistent Akaikes Information Criterion; cL_{RM}: Bedingte Likelihood des dichotomen Rasch-Modells; cL_{2KL}: Bedingte Likelihood des mixed Rasch-Modells in zwei Klassen; $-2(\log(cL_{RM})-\log(cL_{2KL}))$: Bedingter Likelihood-Quotiententest

Hier kommen BIC und CAIC zu unterschiedlichen Ergebnissen. Während BIC das Zweiklassenmodell als leicht besser passend ausweist, kommt CAIC zum umgekehrten Schluss. Da allerdings der Log-Likelihood-Test nicht signifikant wird und CAIC die verlässlichere Statistik darstellt, weil es mit steigendem Stichprobenumfang konsistent ist (Read & Cressie, 1988), wird hier die Einklassenlösung bevorzugt und die Personenhomogenität angenommen.

Für die Itemauswahl wurden zunächst alle nach Tabelle 25 als signifikant von ihren Erwartungswerten abweichende Items eliminiert und erneut eine Analyse nach dem Rasch-Modell durchgeführt. Die Items ZM67 und ZM78 zeigten hierbei nach den Meansquare-Outfit-Werten von $MNSQ_{Outfit} = 5.05$ bzw. $MNSQ_{Outfit} = 4.79$ deutliche Modellabweichungen. Allerdings ergab sich lediglich für Item ZM78 ein signifikanter z-standardisierter Wert von $ZSTD_{Outfit} = 2.8$. Wegen des vergleichsweise hohen Standardfehlers von Item ZM67 ($SE_{ZM67} = .63$) und des damit verbundenen Teststärkeverlustes fiel die deutliche Modellabweichung insignifikant aus ($ZSTD_{ZM67} = 1.7$). Da sich in diesem Fall das Effekstärkemaß der Signifikanztestung als überlegen erweist, wurde neben Item ZM78 auch Item ZM67 für die nachfolgenden Analyse eliminiert. Nach dieser Itemseparation wies kein Item eine signifikante Abweichung von den Erwartungswerten der Infit- und Outfit-Statistik auf. Die relativ größte Modellabweichung wies Item ZM68 mit $MNSQ_{Outfit} = 1.19$ auf. Die Modellgültigkeit wurde für die verbleibenden 12 Items daher angenommen. Aus der Verkürzung der Skala resultierte eine Absenkung der internen Konsistenz nach KR-20 von $\alpha = .78$ auf $\alpha = .70$.

14.1.2.5 Subtest „Matrizen“

Tabelle 28 gibt einen Überblick über die wesentlichen deskriptiven Statistiken und Maße zur Modellgültigkeit der Items dieses Subtests.

Tabelle 28: Nach Modellpassung absteigend geordnete Itemstatistiken zum Subtest „Matrizen“ (N = 434)

Item-Position	**Item name**	p_i	**Schwierigkeit** σ_i	**SE**	**MNSQ Infit**	**ZSTD Infit**	**MNSQ Outfit**	**ZSTD Outfit**
2	M82	.30	0.81	0.11	1.18	**3.5**	1.65	**6.8**
4	M84	.75	-1.41	0.12	1.09	1.6	1.20	**2.1**
12	M92	.35	0.55	0.11	1.12	**2.7**	1.18	**2.6**
5	M85	.59	-0.59	0.10	1.13	**3.3**	1.17	**2.9**
13	M93	.09	2.44	0.17	1.08	0.7	1.17	0.8
6	M86	.29	0.85	0.11	1.01	0.1	1.09	1.1
1	M81	.90	-2.69	0.17	1.01	0.1	1.06	0.4
11	M91	.47	-0.01	0.10	1.00	0.0	1.02	0.3
9	M89	.66	-0.92	0.11	0.96	-0.9	1.01	0.2
19	M99	.17	1.66	0.14	0.90	-1.2	0.98	-0.1
14	M94	.25	1.1	0.12	0.98	-0.3	0.97	-0.3
7	M87	.92	-2.91	0.18	0.98	-0.1	0.90	-0.4
15	M95	.44	0.09	0.10	0.96	-1.1	0.97	-0.6
8	M88	.76	-1.5	0.12	0.96	-0.6	0.95	-0.5
10	M90	.81	-1.84	0.13	0.94	-0.7	0.91	-0.7
16	M96	.20	1.42	0.13	0.94	-0.9	0.83	-1.5
3	M83	.91	-2.81	0.18	0.92	-0.6	0.85	-0.7
18	M98	.22	1.31	0.12	0.89	-1.7	0.85	-1.4
20	M100	.06	3.00	0.22	0.86	-0.8	0.48	**-2.2**
17	M97	.20	1.45	0.13	0.85	**-2.1**	0.8	-1.8
M		.47	0.00	0.13	0.99	0.0	1.0	0.4
SD			1.72	0.03	0.09	1.5	0.22	2

Anmerkung: p_i: Itemschwierigkeit nach der Klassischen Testtheorie; Schwierigkeit σ_i: Logit der Itemschwierigkeit; *SE*: Standardfehler des Itemparameters

Auch für den Matrizen-Subtest zeigt sich eine sehr gute Übereinstimmung der Rangfolge der Schwierigkeitskennwerte nach Klassischer Testtheorie und dem Rasch-Modell bei τ_b = -.99. Die leicht negative klassische Trennschärfe von Item M82 lässt sich aus der Uneindeutigkeit eines Distraktors durch einen Fehler in der graphischen Darstellung erklären. Der Mittelwert des „klassischen“ Schwierigkeitsindexes von .48 weist die Items als mittelschwer aus. Der Mittwelwert der Personenparameter liegt bei -.17 bei einer Standardabweichung von 0.78 (M_{Rohwert} = 9.3, SD_{Rohwert} = 5.62). Insgesamt war der Test also für die Probandenstichprobe

etwas zu schwer. Analog zum Subtest „Verbale Analogien“ und Zahlenreihen erkennt man anhand der MNSQ-Fit-Statisiken eine zunehmende Differenzierungsfähigkeit der Items gegen Ende des Tests hin, welche im Falle der Items M97 und M100 als jeweils signifikanter Overfit ausgewiesen wird, auch wenn die Größe der Abweichung nach den MNSQ-Fit-Statistiken gering ausfällt und im Toleranz-Intervall von 0.8 – 1.20 liegt. Einen jwls. signifikanten Underfit, somit eine zu geringe Trennschärfe, weisen (neben dem bereits genannten Item M82) die Items M84, M85 und M92 auf. Tabelle 29 gibt Auskunft über die Personen- und Itemseparationsindizes.

Tabelle 29: Personen- und Itemseparationsreliabilitäten Subtest „Matrizen“

	Separationsreliabilität *R*	**Separationsindex *G***
Personen	.73	1.70
Items	.99	12.31
KR-20	.74	

Die Indizes der Personenseparation fallen in Bezug auf *R*, der KR-20 und dem Personen-separationsindex *G* befriedigend aus. Die entsprechenden Indizes der Itemseparation hingegen verweisen auf eine sehr gute Separation. Nach dem Separationsindex *G* können somit gut 12 statistisch voneinander unterschiedliche Schwierigkeitsstufen innerhalb der Items identifiziert werden.
Tabelle 30 zeigt das Ergebnis der Überprüfung der Personenhomogenität.

Tabelle 30: Modellvergleiche anhand informationstheoretischer Maße und dem Likelihood-Quotiententests Subtest „Matrizen“

Modell	**BIC**	**CAIC**	**Log L**	**-2(log(cL_{RM})-log(cL_{2KL}))**
Dichotomes Rasch-Modell	8733.71	8754.71	-4303.09	196.92**
Zweiklassen mixed Rasch-Modell	8670.40	8713.40	-4204.63	

Anmerkung: Log L: Logarithmierte Likelihood der Daten unter den jwl. Modellannahmen;
BIC: Bayes Information Criterion; CAIC: Consistent Akaikes Information Criterion; cL_{RM}: Bedingte Likelihood des dichotomen Rasch-Modells; cL_{2KL}: Bedingte Likelihood des mixed Rasch-Modells in zwei Klassen; -2(log(cL_{RM})-log(cL_{2KL})): Bedingter Likelihood-Quotiententest

Sowohl nach dem BIC als auch nach CAIC ergibt sich eine bessere Modellpassung der Zwei-gegenüber der Einklassenlösung des mixed Rasch-Modells. Zudem wird der Likelihood-Quotiententest signifikant und weist auf eine substanzielle Personenheterogenität hin. Die mittlere Zuordnungswahrscheinlichkeit liegt mit .99 für Klasse 1 und .92 für Klasse 2 sehr hoch. Abbildung 12 gibt einen Eindruck von der Art der Personenheterogenität.

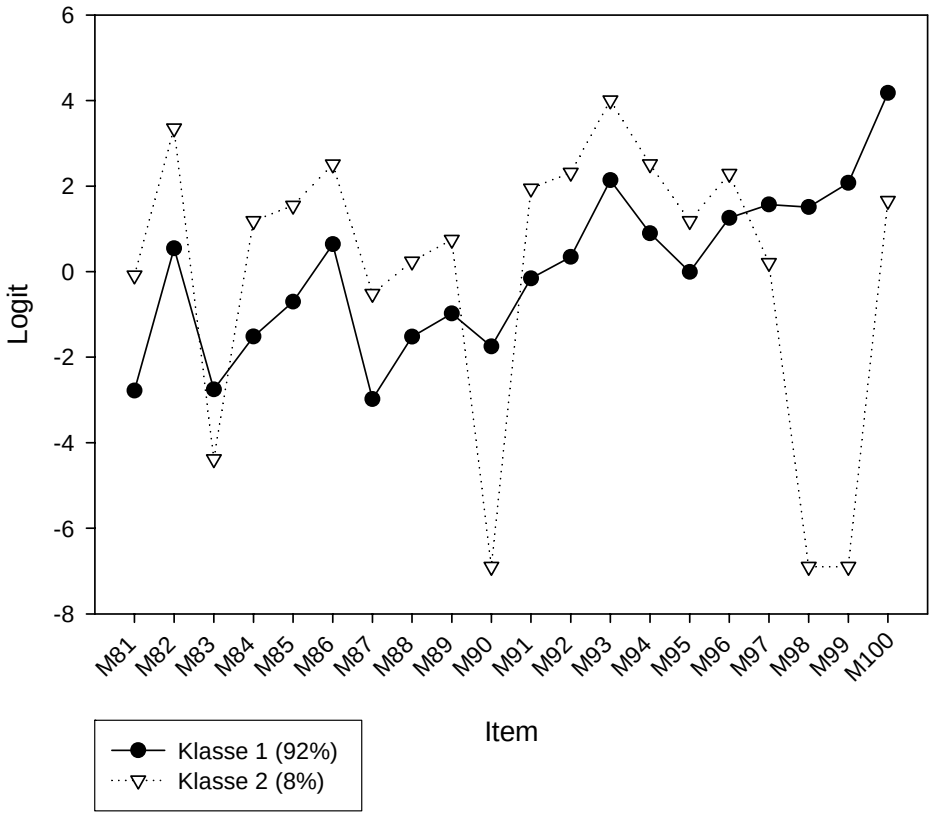


Abbildung 12: Itemparameter nach dem mixed Rasch-Modell in zwei Klassen für Subtest „Matrizen“

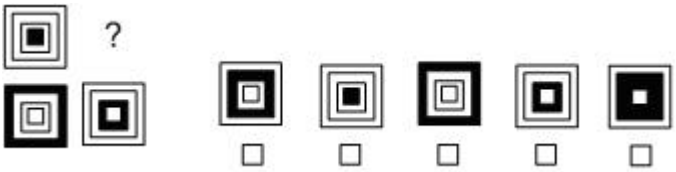
Anders als beim Subtest verbale Analogien oder Zahlenreihen ergibt sich keine einfache Beziehung der bearbeitungsschnelleren Klasse 2 gegenüber einer bearbeitungslangsameren Klasse 1. Zunächst lässt sich zwar auch zwischen ihnen ein deutlicher Unterschied in den Itemschwierigkeiten für die letzten drei Items (M97 bis M100) ausmachen, wobei sich M100 auch für Klasse 2 als schwer erwies, sodass wohl auch hier die Testzeitbegrenzung einen einflussreichen Faktor für diese Itemschwierigkeit darstellte. Ein t-Test für unabhängige Stichproben mit gleichen Varianzen für die Personenparameter ergab nach den Effektstärken

eine deutliche Überlegenheit der Klasse 2 (M_{PP} = 1.83, SD_{PP} = 0.93, M_{RW} = 12.78, SD_{RW} = 2.59) gegenüber der Klasse 1 (M_{PP} = -0.33, SD_{PP} = 0.78, M_{RW} = 9.01, SD_{RW} = 2.32) in der Subtestgesamtleistung: $t(432) = 15.45$, $p < .001$; Cohen's $d = 2.73$, $r_{ptbis} = .59$. Allerdings kann man wie im Subtest „Verbale Analogien" einen tendenziellen Speed-accuracy-trade-off der Klasse 2 gegenüber Klasse 1 ausmachen.
Jedoch erklärt der Speed-Effekt nicht die sehr deutlich geringer ausgeprägte Schwierigkeit von Item M83 und M90 in Klasse 2 (s. Abbildung 12). Die folgende Einzelanalyse dieser Items scheint daher angebracht, um die Gründe hierfür zu identifizieren.

Item M83

Lösung: Wahlalternative drei

Item M90



Lösung: Wahlalternative eins

An beiden Items ist auffallend, dass vornehmlich *visuell-räumliche* Regeln für ihre Lösungen nötig sind und keine (oder nur sehr gering) *analytisch-logische.* So bestehen die Schwierigkeitskomponenten von M83 aus dem Oberflächenmerkmal des Schwarz-Weiss-Kontrastes und der Raumlage des Dreiecks. Auch bei M90 ist die Variation der Oberflächenstruktur als Schwarz-Weiss-Kontrast entscheidend für seine Lösung. Diese Items sprechen daher eine kognitive Komponente an, welche in der Literatur als Wahrnehmungs-, Figural- oder auch Gestaltfaktor bezeichnet wird und auch in den Raven Advanced Progressive Matrices (Raven, Raven & Court, 1998) gefunden wurde (s. hierzu u.a. Gallini, 1983; Mackintosh & Bennett, 2005; van der Ven & Ellis, 2000; Vigneau & Bors, 2005).

Es stellt sich somit einmal mehr die Frage nach der Eindimensionalität von Tests, welche vorgeben, nur einen g-Faktor zu messen. In jedem Fall lassen sich (allerdings post hoc) die deutlich niedrigeren Itemparameter von M83 und M90 in Klasse 2 dadurch erklären, dass Personen dieser Klasse über eine höhere Wahrnehmungsgeschwindigkeit als diejenigen in Klasse 1 verfügen und somit gerade Items, deren Komponenten stärker visuelle-räumliche Wahrnehmungsprozesse als analytisch-logische Regelhaftigkeiten ansprechen, werden von diesen Probanden besonders leicht gelöst.

Für die Testkonstruktion bedeutet dies jedoch, auch in diesem Test den Speed-Effekt mithilfe des Modells nach Gleichung (47) zu kontrollieren. Zunächst wurde vorab Item M82 aufgrund der Uneindeutigkeit eines Distraktors entfernt und mit den verbleibenden Items eine Rekalibrierung vorgenommen. Die Ergebnisse der Item-Modellgeltung zeigt Tabelle 31.

Tabelle 31: Itemstatistiken im Subtest „Matrizen“ nach Modell-Reformulierung ($N = 434$)

Item	**Item-Schwierigkeit σ_i**	**SE**	**MNSQ-Outfit**	**ZSTD-Outfit**	**MNSQ-Infit**	**ZSTD-Infit**
M81	-2.90	0.13	1.18	0.70	1.03	0.30
M83	-2.50	0.14	0.95	-0.10	0.99	0.00
M84	-1.20	0.10	1.14	0.60	1.09	1.70
M85	-0.45	0.09	1.11	0.50	1.09	**3.10**
M86	0.86	0.10	0.96	-0.10	0.97	-0.60
M87	-2.50	0.14	1.00	0.10	1.00	0.00
M88	-1.30	0.10	1.04	0.30	1.02	0.30
M89	-0.80	0.10	1.02	0.20	1.01	0.30
M90	-1.66	0.11	1.05	0.30	1.01	0.20
M91	0.04	0.09	0.96	-0.10	0.96	-1.80
M92	0.54	0.09	1.07	0.40	1.06	1.90
M93	2.28	0.14	1.05	0.30	1.00	0.00
M94	0.96	0.10	0.96	-0.10	0.98	-0.50
M95	0.05	0.09	0.97	0.00	0.97	-1.20
M96	1.27	0.11	0.90	-0.40	0.94	-0.90
M97	1.14	0.11	0.83	-0.60	0.90	-1.70
M98	1.10	0.11	0.84	-0.60	0.89	-1.90
M99	1.63	0.11	0.84	-0.60	0.92	-1.00
M100	2.10	0.48	0.94	-0.20	0.97	-0.20
Itemseparations-reliabilität $R = .99$						

Anmerkung: Itemschwierigkeit σ_i: Logit der Itemschwierigkeit; *SE*: Standardfehler des Itemparameters; MNSQ Outift: Meansquare-Outfit-Statistik; MNSQ Infit: Meansquare Infit-Statistik; ZSTD Outfit: z-standardisierte MNSQ-Outfit-Statistik; ZSTD Infit: z-standardisierte MNSQ-Infit-Statistik

Es resultiert eine deutlich verbesserte Modellgeltung der verbleibenden Items. Lediglich Item M85 weist einen signifikanten Outfit-Wert auf und wurde für eine erneute Analyse eliminiert. Hiernach wies sich kein Item eine signifikante Modellabweichung auf. Die größte Modellabweichung zeigte Item M81 ($MNSQ_{Outfit} = 1.19$, $ZSTD_{Outfit} = 0.92$). Die Modellgeltung dieser 19 Items wurde daher angenommen. KR-20 stieg durch die Verkürzung der Skala um Item M82 mit schlechter Modellpassung von $\alpha = .74$ auf $\alpha = .76$ an.

14.1.2.6 Subtest „SPARK“

Tabelle 32 gibt zunächst die Ergebnisse der Modellgeltung auf Itemebene wieder.

Tabelle 32: Nach Modellpassung absteigend geordnete Itemstatistiken des Subtests „SPARK“ ($N = 434$)

Item-Position	**Item name**	p_i	**Schwierigkeit** σ_i	**SE**	**MNSQ Infit**	**ZSTD Infit**	**MNSQ Outfit**	**ZSTD Outfit**
1	AK1	.29	1.50	0.08	1.20	**3.1**	1.56	**7.1**
2	AK2	.52	0.07	0.08	1.03	0.5	0.96	-0.5
4	AK4	.69	-1.09	0.08	0.78	**-3.5**	0.60	**-7.0**
3	AK3	.70	-0.98	0.08	0.95	-0.7	0.86	**-2.1**
5	AK5	.45	0.53	0.17	1.06	1.1	1.05	0.7
M		.53	0.00	0.10	1.00	0.1	1.01	-0.36
SD			1.08	0.04	0.15	2.4	0.35	5.1

Anmerkung: p_i: Itemschwierigkeit nach der Klassischen Testtheorie; Schwierigkeit σ_i: Logit der Itemschwierigkeit; *SE*: Standardfehler des Itemparameters

Die Rangreihe der Itemschwierigkeiten nach der Klassische Testtheorie und derjenigen nach dem Rasch-Modell stimmen mit $\tau_b = -.80$ in diesem Subtest gut miteinander überein. Die mittlere Itemschwierigkeit nach der Klassische Testtheorie von .53 weist die Items als moderat schwer aus. Der Mittelwert der Personenparameter liegt bei 0.16 bei einer Standardabweichung von 1.70 ($M_{Rohwerte} = 2.65$, $SD_{Rohwerte} = 1.58$), wodurch der Test damit insgesamt für die Probandenstichprobe als etwas zu leicht ausgewiesen wird, wobei aber zugleich die große Varianz der Fähigkeitsverteilung auffällt.

Die Modellgeltung der Items ist insgesamt sehr problematisch. Item AK1 (Identifizierung der Zeile, in der eine argumentative Auffälligkeit steht und Benennung dieser anhand von neun Multiple-Choice-Alternativen) erweist sich durch einen besonders hohen positiven und signifikanten MNSQ-Outfit-Wert als zu trennschwach. Hierfür mag die Gestaltung der

Distraktoren verantwortlich sein. Diese sind vermutlich zu leicht als unwahrscheinliche Lösungen zu identifizieren und ermöglichen so auch eine Lösung nach dem Ausschlussprinzip unplausibler Distraktoren. Die Distraktorverteilung über die neun Antwortkategorien nach lässt dies zumindest vermuten, weil sich die Antworten überwiegend auf die Alternativen 2 und 3 verteilen, wie Tabelle 33 zeigt:

Tabelle 33: Distraktorverteilung von Item AK1

Antwortalternative	**Häufigkeit**	**Prozent**
0 (keine Auffälligkeit wahrgenommen)	38	8.8
1	17	3.9
2	179	41.2
3	**173**	**39.9**
4	6	1.4
5	3	0.7
6	5	1.2
7	5	1.2
8	8	1.8
9	0	0.0

Anmerkung: Die korrekte Antwortalternative ist fett markiert

Deutliche Modellabweichungen besonders in Form eines Overfits ergeben sich für die Items AK3 und AK4, beide Aufgaben scheinen also lokal voneinander abhängig zu sein. Bei der inhaltliche Analyse der Items AK2 bis AK4 fällt nun Folgendes auf: In allen drei Items ist die korrekte Beurteilung dreier Argumente als Pro-, Contra- oder Weder-Noch-Argument in Bezug auf die folgende These verlangt

> „Die einzelnen Bestandteile unseres Bildungssystems sind aufeinander abgestimmt. Man muss sie deswegen auch in einer bestimmten Reihenfolge durchlaufen. Nur so kann man sich die Voraussetzungen für die erfolgreiche Bewältigung der nächst höheren Stufe innerhalb des Bildungssystems verschaffen.“

Mit den zu beurteilenden Argumenten, ob Pro-, Contra-, oder Weder-noch-Argument:

AK2:

„Und sehen Sie, sie will ja arbeiten an der Universität. Immer, wenn ich mit ihr rede, sagt sie das. Und sie meint dass wirklich ernst.“ *Korrekte Antwort:* Weder-noch-Argument.

AK3:

„Aber es gab doch auch schlechte Schüler, die an der Universität gut waren. Einstein hatte doch auch schlechte Noten!“ *Korrekte Antwort:* Contra-Argument.

AK4:

„Auf der Schule werden nämlich Grundfertigkeiten vermittelt, z. B. Grundkenntnisse in Mathematik und Fremdsprachen, die für ein erfolgreiches Studium an der Universität unerlässlich sind.“ *Korrekte Antwort:* Pro-Argument.

Offensichtlich stellt das Argument in Item AK4 das inhaltlich-logisch „gespiegelte“, invertierte von AK3 dar. Die Lösung von AK4 hängt somit stark auch davon ab, ob eine Person bereits AK3 gelöst hat, weil hier bereits Information enthalten ist, die zur Lösung von AK4 *übertragen* werden kann. Wenn also die korrekte Lösung von AK3 das Contra-Argument ist, muss, wegen der logischen Abhängigkeit durch Iteminhaltsinvertierung, das Pro-Argument die Lösung für AK4 sein. Somit trägt Item AK4 für die Leistungsfähigkeitsmessung zu wenig eigene statistische Information bei und ist daher für die Fähigkeitsmessung redundant. Eine Testrevision müsste daher das Argument in AK4 gegen eines austauschen, welches nicht nur eine bloße Variation desjenigen von AK3 darstellt.

Tabelle 34: Personen- und Itemseparationsstatistiken Subtest „SPARK“

	Separationsreliabilität *R*	**Separationsindex *G***
Personen	.44	0.89
Items	.99	14.10
KR-20	.69	

Ebenfalls unterschiedlich fällt die Beurteilung der Personenseparations-Statistiken nach KR-20 und der Separationsreliabilität *R* des Rasch-Modells aus. Während KR-20 eine ausreichende Reliabilität anzeigt, ergibt die Personenseparationsreliabilität *R* des Rasch-Modells mit .44 einen deutlich schlechteren Wert, wie auch der Personenseparationsindex *G* sehr niedrig

ausfällt. Die Summenwerte weisen eine sehr große Varianz auf. Bei nur sechs unterschiedlichen Summenscores beträgt die Varianz 2.56. Unter anderem kommt dies dadurch zustande, dass 26.3% der Stichprobe die Extremwerte Null oder Fünf aufweisen. Nach KR-20 ist dies mit keinerlei Problemen verbunden, weil hier die Fehlervarianz als konstant für alle Personenfähigkeitsausprägungen angenommen wird. Die hohe Varianz führt dadurch zu einer höheren internen Konsistenz. Nach der Methode des Rasch-Modells zur Abschätzung der Fehlervarianz über den Erwartungswert von Standardschätzfehlern der einzelnen Personenparameter hingegen sind es eben jene Extremwerte, die mit einem hohen Standardschätzfehler behaftet sind. Für den Minimum-Wert von Null liegt dieser bei 1.71, für den Maximum-Wert bei 1.74. Beide Werte liegen sogar noch höher als die Standardabweichung der Personenparameter von 1.70. Es sind daher nur die 73.7% Personen mit Rohwerten von Eins bis Vier, die für eine Reliabilität größer Null sorgen. Die Rasch-Reliabilität kommt hier (ohne die Annahme gleich großer Fehlervarianz für alle Personenfähigkeiten) zu einer realistischeren Schätzung der tatsächlichen Reliabilität.
Die Ergebnisse zur Personenhomogenitätsüberprüfung fasst Tabelle 35 zusammen.

Tabelle 35: Modellvergleiche anhand informationstheoretischer Maße und dem Likelihood-Quotiententests Subtest „SPARK“

Modell	**BIC**	**CAIC**	**Log L**	**-2(log(cL_{RM})-log(cL_{2KL}))**
Dichotomes Rasch-Modell	2543.67	2549.67	-1253.62	170.30**
Zweiklassen mixed Rasch-Modell	2415.89	2428.89	-1168.47	

Anmerkung: Log L: Logarithmierte Likelihood der Daten unter den jwl. Modellannahmen; BIC: Bayes Information Criterion; CAIC: Consistent Akaikes Information Criterion; cL_{RM}: Bedingte Likelihood des dichotomen Rasch-Modells; cL_{2KL}: Bedingte Likelihood des mixed Rasch-Modells in zwei Klassen; -2(log(cL_{RM})-log(cL_{2KL})): Bedingter Likelihood-Quotiententest

Es zeigt sich, dass nach beiden informationstheoretischen Maßen die Zweiklassenlösung eine bessere Passung aufweist, die zudem nach dem Likelihood-Quotiententest signifikant ausfällt. Die mittlere Zuordnungswahrscheinlichkeit liegt mit .99 für Klasse 1 und .93 für Klasse 2 jeweils sehr hoch und ist somit sehr reliabel. Abbildung 13 gibt die Art der Personenheterogenität wieder.

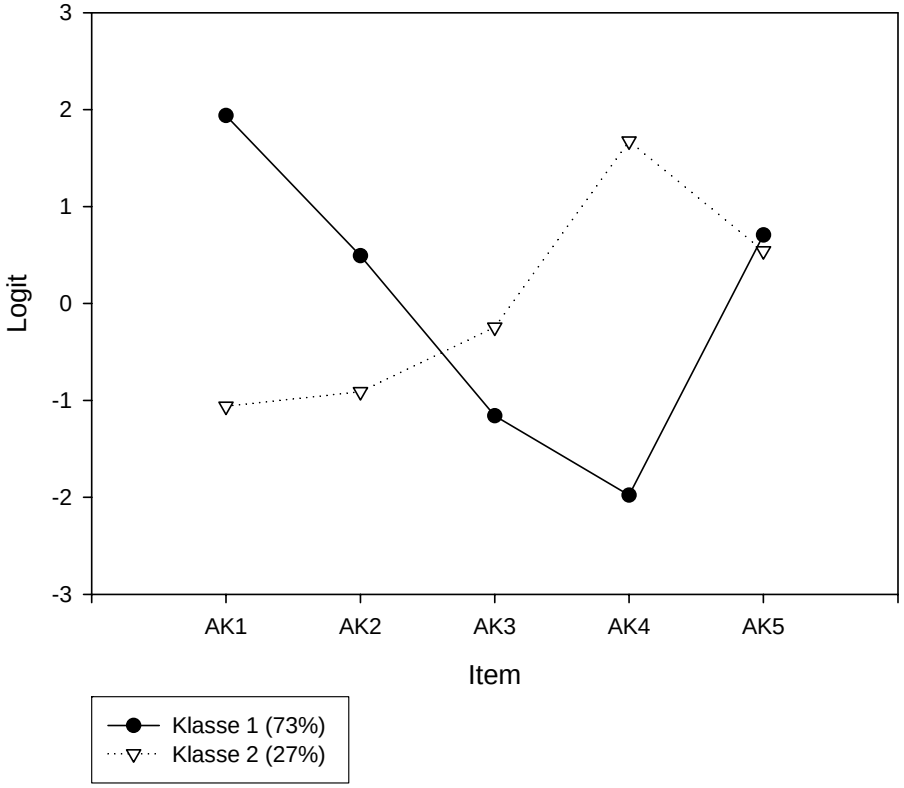


Abbildung 13: Itemparameter nach dem mixed Rasch-Modell in zwei Klassen für Subtest „SPARK“

Ein t-Test über die Mittelwerte für Stichproben mit ungleichen Varianzen ergab große Effektstärken und somit eine Überlegenheit in der Testgesamtleistung der Klasse 1 (M_{PP} = 1.14, SD_{PP} = 1.19, M_{RW} = 3.45, SD_{RW} = 1.19) gegenüber der Klasse 2 (M_{PP} = -2.17, SD_{PP} = 0.77, M_{RW} = 0.48, SD_{RW} = 0.58): $t(318.69) = 33.87$, $p < .001$; Cohen’s $d = 4.27$, $r_{ptbis} = .88$.

Wie bereits in den Ausführungen zur entsprechenden Abbildung im Subtest „verbale Analogien“ erläutert, darf die Überschneidung der Itemprofile nicht quantitativ interpretiert werden. So ist Klasse 1, obgleich die Items AK1 und AK2 hier höhere Schwierigkeiten aufweisen, quantitativ betrachtet die leistungsfähigere. Vielmehr zeigt die Abbildung, dass es gerade die nach den Item-Fit-Statistiken als „overfitting“ indizierten Items sind (AK3 und AK4), welche die *strukturellen* (nicht quantitativen!) Unterschiede zwischen den Klassen ausmachen. Klasse 1 profitiert durch ihre höhere Leistungsfähigkeit stärker von der Item-redundanz von AK4: AK4 ist leichter als AK3, weil seine Lösung neben der Leistungsfähigkeit zusätzlich von der Lösung von AK3 mitbedingt wird. Bei der weniger leistungsfähigen Klasse 2 hingegen wirkt sich die Abhängigkeit genau in umgekehrter Richtung aus. Als Ergebnis der Personenheterogenität lässt sich daher sagen, dass die Zweiklassenlösung nicht etwa durch

unterschiedliche Lösungsstrategien zustande kommt, sondern durch die logische Abhängigkeit der Items AK3 und AK4. Neben der Verbesserung aufseiten der Distraktoren von Item AK1 müsste daher eine Testrevision ebenfalls an einer Ersetzung von AK4 ansetzen.

Für die vorliegende Arbeit ergibt sich nun das Problem, dass eine Eliminierung der besonders von dem Modellerwartungen abweichenden Items AK1 und AK4 in einer Skala mit lediglich drei Items resultieren würde. Zu wenige also, um bei der ohnehin sehr niedrigen Reliabilität die Personenfähigkeit überhaupt noch verlässlich erfassen zu können. An dieser Stelle muss daher auf eine rein vorläufige Lösung ausgewichen werden, indem auf die Skalenkonstruktion nach der Klassischen Testtheorie zurückgegriffen wird. Aus diesem Grund wurde nach der Logik der Klassischen Testtheorie das trennschwächste Item (AK1) eliminiert, Item AK3 und AK4 als jwls. trennscharfe Items hingegen im Test belassen. Durch die Eliminierung von Item AK1 stieg die interne Konsistenz nach KR-20 von .69 auf .74 an.

Im Folgenden werden die Einzelergebnisse der Analysen nach dem Multifacetten--Rasch-Modell der Kreativitätsfacetten Ideenflüssigkeit und Ideenflexibilität und zur Aufgabenstellung „empiriebezogenes Denken“ berichtet. Zu den Einzelheiten der Bewertungskriterien sei an dieser Stelle für die Kreativitätsfacetten lediglich auf Kapitel 8.1.3 und für die Aufgabe zum empirischen Denken auf Anhang D verwiesen, wo jeweils detaillierte Informationen aufgeführt sind.

14.1.2.7 Kreativitätsfacetten

Für die Analyse der Kreativitätsfacette *Ideenflüssigkeit* wurden diejenigen Aufgaben herangezogen, welche von den Probanden unter Zeitlimitierung die Produktion folgender Leistungen verlangte:

1) Möglichst viele verschiedene Zeichnungen aus einer vorgegebenen Figur herstellen (in den Ergebnisdarstellungen als „Zeichnungen, Menge“ abgekürzt).
2) Möglichst viele Sätze, bestehend aus vorgegebenen drei Wörtern produzieren (in den Ergebnisdarstellungen als „Sätze, Menge“ abgekürzt).
3) Möglichst viele Verwendungsmöglichkeiten eines Gegenstandes nennen (in den Ergebnisdarstellungen als „Verwendungen, Menge“ abgekürzt).

Da die Daten der Probanden als Mengenleistungen vorlagen, wurde ein MFRM (Linacre & Wright, 2002) mit einer zwanzigstufigen Kategorienanzahl modelliert. Die Kategorienanzahl von 20 ergab sich hierbei aus dem Maximalwert der produzierten Lösungsmenge in dieser Stichprobe. In die Analyse gingen die Leistungen der Gesamtstichprobe von $N = 434$ doppelt und unabhängig voneinander ausgewerteten Testprotokollen ein.
Tabelle 36 bis Tabelle 38 geben die Ergebnisse der MFRM-Analyse, getrennt nach den Facetten Beurteiler, Aufgaben und Probanden wieder.

Tabelle 36: Statistische Fit-Analyse für die Beurteilerfacette Subtest Ideenflüssigkeit ($N = 434$)

OBSVD SCORE	OBSVD COUNT	OBSVD AVERAGE	FAIR-M AVRAGE	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTBIS	BEURTEILER
8045	1302	6.2	6.04	.47	.02	.85	-3.5	.92	-2.0	.49	1
8267	1302	6.3	6.20	.40	.02	1.03	.6	1.11	2.6	.47	2
8156.0	1302.0	6.3	6.12	.44	.02	.94	-1.5	1.01	.3	.48	MEAN (COUNT: 2)
157.0	.0	.1	.11	.04	.00	.12	2.9	.14	3.4	.01	S.D. (SAMPLE)

MODEL, SAMPLE: RMSE .02 ADJ (TRUE) S.D. .04 SEPARATION 2.45 RELIABILITY .86
MODEL, FIXED (ALL SAME) CHI-SQUARE: 7.0 D.F.: 1 SIGNIFICANCE (PROBABILITY): .01
RATER AGREEMENT OPPORTUNITIES: 1302 EXACT AGREEMENTS: 902 = 69.3%

Anmerkung. *OBSVD SCORE*: Summe der beobachteten Ratings über alle Aufgaben; *OBSVD COUNT*: Anzahl beobachteter Ratings je Beobachter; *OBSVD AVERAGE*: Durchschnittliches beobachtetes Rating des jwl. Beurteilers; *FAIR-M AVERAGE*: Fairer Durchschnitt (um Strengefehler korrigierte mittlere Beurteilung); *MEASURE*: Logit-Strengewert des jwl. Beurteilers; *MODEL S.E.:* Standardfehler des Strengeparameters; *INFIT/OUTFIT MNSQ*: Meansquare-Fit-Statistiken; *ZSTD*: z-standardisierte Meansquare-Fit-Statistiken; *PTBIS*: SR/ROR-Korrelation.

Die Nullhypothese, dass beide Beurteiler kein unterschiedliches Ausmaß an Strenge zeigen, wird durch den Chi-Quadrat-Test verworfen. Ebenso indizieren die Separationsreliabilität R und der Separationsindex G (in der FACETS-Ausgabe als „Reliability“ bzw. „Separation“ gekennzeichnet) eine gute Trennbarkeit der Beurteiler anhand ihrer Strengemaße. Auf deskriptiver Ebene zeigt sich, dass Beurteiler 1 mit einem Logit-Strengewert von 0.47 gegenüber Beurteiler 2 mit einem von 0.40 strenger war. Auf Rohwertebene entspricht dieser Strengeparameterunterschied einem mittleren Rating von 6.2 (Beurteiler 1) zu 6.3 (Beurteiler 2). Das mittlere Ausmaß der Unterschiedlichkeit in der Strenge ist daher allerdings als gering zu betrachten. Gleiches lässt sich auch für die Auswirkungen der Strenge bezüglich der

Korrektur für eine faire Beurteilung sagen. Zwar zeigen beide Auswerter einen Strengeeffekt, da die fairen mittleren Ratings jwls. niedriger als die beobachteten liegen (vgl. „Obsvd Average“mit „Fair-M Average“) und auch die positiven Strenge-Logits (in der Facets-Ausgabe als „Measure“ betitelt) verweisen auf einen Strengeeffekt beider Beurteiler. Allerdings fallen diese jeweiligen Differenzen des beobachteten Ratings zum korrigierten fairen Mittelwert mit $Ratingmittelwert_{\text{Auswerter 1}} = 6.2$ zu $Ratingmittelwert_{\text{Auswerter 2}} = 6.04$ bzw. $Fairer\ Durchschnitt_{\text{Auswerter 1}} = 6.3$ zu $Fairer\ Durchschnitt_{\text{Auswerter 2}} = 6.2$ sehr gering aus. Die SR/ROR-Korrelationen liegen mit .49 bzw. .47 im befriedigenden Bereich wie auch die prozentualen Übereinstimmungen mit 69.3%. Offenbar ist das Auswertungsschema des BIS (Jaeger, Suess & Beauducel, 1997) für diese Teilaufgabe verlässlich. Allerdings weisen die MNSQ-Fit-Statistiken Infit und Outfit insbesondere bei Beurteiler 1 auf einen Halo-Effekt hin. Sowohl Infit als auch Outfit liegen mit 0.85 und 0.92 signifikant unter ihrem Erwartungswert von 1 (s. die entsprechenden z-transformierten Werte „Zstd“). D.h., dass die Beurteilungen von Auswerter 1 über alle drei Aufgaben hinweg betrachtet zu vorhersagbar waren, zu wenig Probabilistik enthielten. Offenbar hatte Auswerter 1 Schwierigkeiten, zwischen den Einzelaufgaben zu unterscheiden und übertrug ähnliche Einschätzungen von Teilaufgabe zu Teilaufgabe. Rater 2 hingegen zeigt einen signifikanten Outfit von 1.11. Offensichtlich produzierte dieser Auswerter signifikante Ausreißer-Werte, indem er nach Modellerwartungen einigen eigentlich schwächeren Probanden bessere Leistungen und vice versa attestierte. Allerdings liegt die Größe der Modellabweichung bei beiden Beurteilern im Bereich einer guten Modellpassung von 0.80 bis 1.20 (s. Myford & Wolfe, 2003, S. 409). Insgesamt betrachtet sind demnach die Auswirkungen der unterschiedlichen Strenge und anderer Urteilseffekte auf die Objektivität der Leistungsbeurteilung als sehr gering zu betrachten, weshalb für diese Facette von einer guten Teilanpassung an das Modell gesprochen werden kann.

Einen Überblick über die Modellgeltung der Aufgabenfacette gibt Tabelle 37.

Tabelle 37: Statistische Fit-Analyse für die Aufgabenfacette Subtest Ideenflüssigkeit (N = 434)

OBSVD SCORE	OBSVD COUNT	OBSVD AVERAGE	FAIR-M AVRAGE	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTBIS	ITEMS
6356	868	7.3	7.05	-.03	.02	1.09	1.6	1.15	2.8	.42	1 ZEICHNUNGEN MENGE
2860	868	3.3	3.28	.14	.03	1.17	3.4	1.16	3.2	.35	2 SÄTZE MENGE
7096	868	8.2	8.21	-.11	.02	.74	-5.6	.74	-5.6	.53	3 VERWENDUNGEN MENGE
5437.3	868.0	6.3	6.18	.00	.02	1.00	-.2	1.01	.2	.43	MEAN (COUNT: 3)
2262.5	.0	2.6	2.58	.13	.00	.23	4.8	.24	5.0	.09	S.D. (SAMPLE)

MODEL, SAMPLE: RMSE .02 ADJ (TRUE) S.D. .13 SEPARATION 5.70 RELIABILITY .97
MODEL, FIXED (ALL SAME) CHI-SQUARE: 60.4 D.F.: 2 SIGNIFICANCE (PROBABILITY): .00

Anmerkung. *OBSVD SCORE*: Summe der beobachteten Beurteilungen über alle Aufgaben; *OBSVD COUNT*: Anzahl beobachteter Beurteilungen je Aufgabe; *OBSVD AVRAGE*: Durchschnittlich beobachtete Beurteilung für jede Aufgabe; *FAIR-M AVRAGE*: Fairer Durchschnitt (um Strengefehler korrigierte Aufgabenschwierigkeit in Rohwerteinheiten); *MEASURE*: Logit der jwl. Aufgabenschwierigeit; *Model S.E.:* Standardfehler der Aufgabenschwierigkeit; *INFIT/OUTFIT MNSQ*: Meansquare-Fit-Statistiken; *ZSTD*: z-standardisierte Meansquare-Fit-Statistiken; *PTBIS*: Trennschärfe der jwl. Aufgabe.

Die Nullhypothese, dass die drei Aufgabenschwierigkeiten nicht verschieden sind, wird durch den Chi-Quadrat-Test verworfen. Sowohl der Separationsquotient („Separation“) als auch die Separationsreliabilität („Reliability“) weisen mit sehr hohen Werten von G = 5.7 bzw. R = .97 auf eine sehr gute Unterscheidbarkeit der Aufgaben bezüglich ihrer Schwierigkeiten hin. Die schwierigste der drei Aufgaben ist die „Menge zu produzierender Drei-Wort-Sätze“ mit einem Logit von σ_2 = 0.14, die „Menge verschiedener Verwendungsmöglichkeiten für Schaumstoffpolster“ mit σ_3 = -0.11 hingegen die leichteste.

Die faire Schwierigkeitseinstufung („Fair-M Avrage“), d.h., der um die Strenge der Beurteiler und der Verteilung der Probandenfähigkeit bereinigte Wert, liegt für alle drei Aufgaben nur knapp unter dem unadjustierten („Obsvd Average“). Relativ am stärksten muss hierbei die mittlere Schwierigkeitseinschätzung der Aufgabe möglichst vieler Zeichnungen von 7.3 auf 7.05 adjustiert werden.

Betrachtet man die Itemfit-Maße, so fällt besonders auf, dass die Aufgabe zur „Menge verschiedener Verwendungsmöglichkeiten für Schaumstoffpolster“ zwar nach Kriterien der Klassischen Testtheorie mit einer Trennschärfe von .53 am besten zwischen den Probanden zu differenzieren vermag, nach dem MFRM allerdings als zu gut passend indiziert wird, betrachtet man die signifikanten z-standardisierten MNSQ-Fit-Statistiken. Die unstandardisierten Werte liegen mit jwls. 0.74 deutlich unterhalb ihres Erwartungswertes, was ein

deutlicher Hinweis auf Mehrdimensionalität dieser Aufgabe ist. Vermutlich hängt gerade die Produktion verschiedener Verwendungsmöglichkeiten eines Gegenstandes zusätzlich von spezifischen *Wissens*strukturen über Anwendungsbereiche für Schaumstoffpolster und nicht alleine von der Ideenflüssigkeit ab.

Allerdings zeigen auch die beiden anderen Aufgaben signifikante Modellabweichungen. So erzielten nach der z-standardisierten Outfit-Statistik einige Ausreißer mit geringen Ausprägungen auf der Gesamtskala nach den Modellannahmen unerwartet hohe Werte in der Aufgabe zur Produktion möglichst vieler Zeichnungen nach Teilvorlage. Die Aufgabe zur Produktion möglichst vieler Dreiwort-Sätze zeigt sich nach der Infit-Statistik anfällig für unerwartete Antworten im Fähigkeitsbereich, in welchem sie eigentlich kalibriert ist wie auch gegenüber einigen Ausreißerantworten, wie die signifikante Outfit-Statistik anzeigt. Gleichwohl liegen die MNSQ-Fit-Statistiken nahe an ihren Erwartungswerten und im akzeptablen Modellgeltungsbereich für High-Stakes-Testungen von 0.80 bis 1.20.

Die Passung der Aufgabenfacette mit dem Modell erweist sich also insbesondere wegen der Aufgabe zur „Menge von Verwendungsmöglichkeiten“ als problematisch, da durch sie die Eindimensionalität der Skala zerstört wird, vermutlich wegen zusätzlich relevanter spezifischer Wissensstrukturen beim Bearbeiten der Aufgabe. Sie wurde daher eliminiert und eine Rekalibrierung der Skala durchgeführt. Die Itemfitindizes der verbleibenden Items „Menge zu produzierender Drei-Wort-Sätze“ und „Menge zu produzierender Zeichnungen“ indizierten für High-Stakes-Testungen nach den Meansquare-Statistiken eine jeweils gute Modellpassung ($MNSQ\text{-}Infit_{Menge\ Zeichnungen} = .85$, $ZSTD\text{-}Infit_{Menge\ Zeichnungen} = 1.7$ $MNSQ\text{-}Outfit_{Menge\ Zeichnungen} = .84$ $ZSTD\text{-}Outfit_{Menge\ Zeichnungen} = 1.8$; $MNSQ\text{-}Infit_{Menge\ Sätze} = 1.12$, $ZSTD\text{-}Infit_{Menge\ Zeichnungen} = 3.7$; $MNSQ\text{-}Outfit_{Menge\ Sätze} = 1.14$, $ZSTD\text{-}Outfit_{Menge\ Sätze} = 4.1$).

Die Ergebnisse zur Modelpassung der Probanden zeigt Tabelle 38.

Tabelle 38: Statistische Fit-Analyse für die Probanden-Facette Subtest Ideenflüssigkeit ($N = 434$)

OBSVD SCORE	OBSVD COUNT	OBSVD AVERAGE	FAIR-M AVRAGE	MODEL MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	
37.6	6.0	6.3	6.17	.00	.25	.93	-.4	1.01	-.3	MEAN (COUNT: 434)
11.6	.0	1.9	1.93	.73	.02	1.13	1.5	1.19	1.7	S.D. (SAMPLE)

MODEL, SAMPLE: RMSE .25 ADJ (TRUE) S.D. .69 SEPARATION 2.74 RELIABILITY .88
MODEL, FIXED (ALL SAME) CHI-SQUARE: 4050.5 D.F.: 433 SIGNIFICANCE (PROBABILITY): .00
MODEL, RANDOM (NORMAL) CHI-SQUARE: 391.6 D.F.: 432 SIGNIFICANCE (PROBABILITY): .92

Anmerkung. *OBSVD SCORE*: Mittelwert beobachteter Gesamtbeurteilungen; *OBSVD COUNT*: Mittelwert beobachteter Beurteilungen je Aufgabe; *OBSVD AVRAGE*: Durchschnittlich beobachtete Beurteilung je Aufgabe; *FAIR-M AVRAGE*: Fairer Durchschnitt beobachtete Beurteilung je Aufgabe; *MEASURE*: Mittlerer Logit der Probandenfähigkeit; *MODEL S.E.:* Mittelwert des Standardfehlers des Logits der Probandenfähigkeit; *INFIT/OUTFIT MNSQ*: Mittelwert der Meansquare-Personenfitindizes; *ZSTD*: z-standardisierte Meansquare-Personenfitindizes.

Die Nullhypothese, wonach die beobachtete Stichprobe eine Zufallsstichprobe aus einer normalverteilten Grundgesamtheit darstellt, kann nach dem Chi-Quadrat-Test mit einer Irrtumswahrscheinlichkeit von 92% beibehalten werden. Die Nullhypothese nicht verschiedener Personenparameterwerte wird zurückgewiesen. Die deskriptiven Statistiken der Personenseparation weisen eine gute Differenzierungsfähigkeit der Personen anhand ihrer Leistungsmaße auf. Der Separationsquotient („Separation“) und die Separationsreliabilität („Reliability“) verweisen mit Werten von 2.74 bzw. .88 auf eine befriedigende Differenzierungsfähigkeit des Verfahrens bezüglich der Probandenfacette hin. Die mittlere Fähigkeitsausprägung von 0.00 weist darauf hin, dass der Test weder zu schwer noch zu leicht für diese Probandenstichprobe war. Die hier dargestellten Fit-Statistiken Infit und Outfit stellen sogenannte Personenfit-Indizes dar und beurteilen die Konsistenz jeder einzelnen Personenantwort mit den Modellannahmen. Die Mittelwerte liegen seitens der MNSQ--Statistiken mit 0.98 (Infit) und 1.01 (Outfit) zwar nahe an ihren Erwartungswerten, allerdings fällt die jwls. hohe Standardabweichung von 1.13 bzw. 1.19 auf. Auch die symmetrisch verteilten und somit leichter bezüglich ihrer Standardabweichungen interpretierbaren z-standardisierten Werte mit Mittelwerten von -0.4 (ZSTD-Infit) und -0.3 (ZSTD-Outfit) weisen mit Standardabweichungen von 1.5 bzw. 1.7 auf Personenheterogenität hin. Auch besteht ein relativ hoher mittlerer Anpassungsfehler dieser Facette von *RMSE* = .25.

Eine Inspektion der Antwortmuster über signifikant abweichende Personenfit-Indizes ergab, dass die Varianz der Personenfit-Maße durch die schlecht zu den Modellannahmen passende Aufgabe zu möglichst vielen Verwendungsmöglichkeiten von Schaumstoffpolster erhöht wird. Ein typisches Antwortmuster war z. B. 11, 4, 2 (Proband Nr. 369). Dieses Antwortmuster ist insofern mit den Modellannahmen unverträglich, da dieser Proband mit einem Personenparameter von $\theta_{369} = 1.48$ eine hohe Mengenleistung in der ersten Aufgabe erbrachte, welche mit einem Logit von $\sigma_1 = -0.03$ mittelschwer ist, jedoch in der dritten Aufgabe (viele Verwendungsmöglichkeiten von Schaumstoffpolster) mit einem niedrigeren Logit-Schwierigkeitswert von $\sigma_3 = -0.11$ deutlich weniger produzierte. Diese Form der Personenheterogenität ist also überwiegend durch die Mehrdimensionalität der Aufgabe vermittelt und kein Ausdruck etwa von zugrunde liegenden heterogenen Mischverteilungen etwa im Sinne von Kreativitätstypen. Die durchgeführte Eliminierung dieser Aufgabe für die endgültige Kalibrierung ist also auch vor diesem Hintergrund sinnvoll.

Insgesamt betrachtet ergibt sich für die Modellgültigkeit der Aufgaben zur Ideenflüssigkeit eine befriedigende Passung. Die signifikanten Unterschiede in den Strengetendenzen der Beurteiler belegen die Notwendigkeit eines Parameters zur Quantifizierung dieses Urteilsfehlers und einer Adjustierung der Urteile um diesen Effekt. Als problematischer erweist sich die Modellpassung der Aufgabenfacette. Hier zeigt sich insbesondere, dass die Aufgabe zur Produktion möglichst vieler Verwendungsmöglichkeiten von Schaumstoffpolster mehrdimensional ist. Eine zusätzlich enthaltene Antwortkomponente mag, als Post-hoc-Hypothese, das Vorliegen spezifischer Wissensstrukturen sein. Diese Aufgabe müsste daher von den Testautoren modifiziert werden, und zwar mit einem weniger spezifischen Verwendungsgegenstand als Schaumstoffpolster.
Seitens der Probandenfacette resultiert für die nahe an ihren Erwartungswerten liegenden Mittelwerten der Personenfit-Indizes eine gute Modellgeltung. Allerdings verweisen die relativ hohen Standardabweichungen dieser Indizes auf vorhandene Personenheterogenität, was jedoch, wie dargestellt, überwiegend ein Resultat der mehrdimensionalen Aufgabe zur Mengenleistung von Verwendungsmöglichkeiten für Schaumstoffpolster ist. Eine Revision dieser Aufgabenfacette würde sich günstig sowohl auf die Modellpassung der Aufgaben- als auch Probandenfacette auswirken.

Die folgenden Ergebnisdarstellungen beziehen sich auf die Analyse der Skala *Ideenflexibilität.* Die Beurteilung der Testprotokolle für die bereits oben beschriebenen Aufgaben erfolgte

hierbei jeweils nach der Vielfalt der Lösungen als der Anzahl unterschiedlicher Kategorien, in welche die Probandenlösungen fielen.
Tabelle 39 gibt eine Übersicht zur Modellgeltung der Beurteilerfacette.

Tabelle 39: Statistische Fit-Analyse für die Beurteilerfacette Subtest Ideenflexibilität ($N = 434$)

OBSVD SCORE	OBSVD COUNT	OBSVD AVERAGE	FAIR-M AVRAGE	MODEL MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTBIS	BEURTEILER
4877	868	5.6	5.49	.48	.03	.96	-.8	.97	-.6	.45	1
4779	868	5.5	5.40	.56	.03	1.01	.2	1.04	.8	.42	2
4828.0	868.0	5.6	5.44	.52	.03	.99	-.3	1.00	.1	.43	MEAN (COUNT: 2)
69.3	.0	.1	.06	.05	.00	.04	.8	.05	1.0	.02	S.D. (SAMPLE)

MODEL, SAMPLE: RMSE .03 ADJ (TRUE) S.D. .04 SEPARATION 1.62 RELIABILITY .72
MODEL, FIXED (ALL SAME) CHI-SQUARE: 3.6 D.F.: 1 SIGNIFICANCE (PROBABILITY): .06
RATER AGREEMENT OPPORTUNITIES: 868 EXACT AGREEMENTS: 430 = 49.5%

Anmerkung. *OBSVD SCORE*: Summe der beobachteten Ratings über alle Aufgaben; *OBSVD COUNT*: Anzahl beobachteter Ratings je Beobachter; *OBSVD AVERAGE*: Durchschnittliches beobachtetes Rating des jwl. Beurteilers; *FAIR-M AVRAGE*: Fairer Durchschnitt (um Strengefehler korrigierte mittlere Beurteilung); *MEASURE*: Logit-Strengewert des jwl. Beurteilers; *MODEL S.E.:* Standardfehler des Strengeparameters; *INFIT/OUTFIT MNSQ*: Meansquare-Fit-Statistiken; *ZSTD*: z-standardisierte Meansquare-Fit-Statistiken; *PTBIS*: SR/ROR-Korrelation.

Die Kennwerte zur Modellgeltung der Beurteilerfacette ergeben keinerlei Hinweis auf fehlende Anpassung. So liegt der mittlere Modellanpassungsfehler mit *RMSE* = .03 sehr niedrig. Der Test auf Verschiedenheit der Strengeparameter ist insignifikant. Auch die deskriptiven Separationsstatistiken indizieren hierbei eine lediglich geringe Unterscheidbarkeit der Auswerter anhand ihrer Strengeparameter mit einem Separationsindex $G = 1.62$ („Separation“) und der Separationsreliabilität von $R = .72$ („Reliability“). Die jeweiligen durchschnittlichen Unterschiede in den Beurteilungen fallen daher auch mit $Urteilsmittelwert_{\text{Auswerter 1}} = 5.6$ und $Urteilsmittelwert_{\text{Auswerter 2}} = 5.5$ und ihren korrespondierenden Logit-Werten von $Strengeparameter_{\text{Auswerter 1}} = 0.48$ zu $Strengeparameter_{\text{Auswerter 2}} = 0.56$ sehr gering aus. Auswerter 1 zeigt gegenüber Auswerter 2 daher eine größere Milde in seinen Urteilen. Diese geringen *mittleren* Unterschiede dürfen jedoch nicht darüber hinwegtäuschen, dass es bei einzelnen Probandenbeurteilungen zu größeren Abweichungen kommen kann. Vergleicht man zudem die Werte beider Auswerter hinsichtlich des beobachteten Durchschnitts („Obsvd Average“) mit denen des fairen („Fair-M Avrage“), so

weisen beide Auswerter eine Strengetendenz auf, da der faire Durchschnitt jwls. leicht unter dem beobachteten liegt. Allerdings fällt die Korrektur in diesem Fall nicht groß aus, sodass der Strengeeffekt beider Beurteiler im Mittel keinen relevanten Einfluss hat. Zudem liegt die SR/ROR-Korrelation („PtBis“) als Maß der Übereinstimmung im befriedigenden Bereich. Auch ergeben sich keinerlei Hinweise auf andere Beurteilerfehler (z. B. einen Halo-Effekt). Die MNSQ-Fit-Statistiken liegen für beide Beurteiler jwls. sehr nahe an ihren Erwartungswerten und die z-standardisierten Werte fallen alle insignifikant aus. Die Modellgeltung dieser Facette kann daher insgesamt als sehr gut angesehen werden.

Tabelle 40: Statistische Fit-Analyse für die Aufgabenfacette Subtest Ideenflexibilität (N = 434)

OBSVD SCORE	OBSVD COUNT	OBSVD AVERAGE	FAIR-M AVRAGE	MEASURE	S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTBIS	ITEMS
5078	868	5.9	5.65	-.12	.03	1.05	.9	1.06	1.2	.41	1 K.-M. ZEICHNUNGEN
4578	868	5.3	5.22	.12	.03	.93	-1.3	.94	-1.0	.45	2 K.-M. VERWENDUNGEN
4828.0	868.0	5.6	5.43	.00	.03	.99	-.2	1.00	.1	.43	MEAN (COUNT: 2)
353.6	.0	.4	.31	.17	.00	.08	1.6	.08	1.6	.03	S.D. (SAMPLE)
MODEL, SAMPLE: RMSE .03 ADJ (TRUE) S.D. .17 SEPARATION 6.18 RELIABILITY .97											
MODEL, FIXED (ALL SAME) CHI-SQUARE: 39.1 D.F.: 1 SIGNIFICANCE (PROBABILITY): .00											

Anmerkung. *OBSVD SCORE*: Summe der beobachteten Beurteilungen über alle Aufgaben; *OBSVD COUNT*: Anzahl beobachteter Beurteilungen je Aufgabe; *OBSVD AVRAGE*: Durchschnittlich beobachtete Beurteilung für jede Aufgabe; *FAIR-M AVRAGE*: Fairer Durchschnitt (um Strengefehler korrigierte Aufgabenschwierigkeit in Rohwerteinheiten); *MEASURE*: Logit der jwl. Aufgabenschwierigeit; *Model S.E.:* Standardfehler der Aufgabenschwierigkeit; *INFIT/OUTFIT MNSQ*: Meansquare-Fit-Statistiken; *ZSTD*: z-standardisierte Meansquare-Fit-Statistiken; *PTBIS*: Trennschärfe der jwl. Aufgabe.
K.-M. Zeichnungen: Kategorienmenge Zeichnungen zu Teilvorlage
K.-M.Verwendungen: Kategorienmenge Verwendungen von Schaumstoffpolster

Die Nullhypothese nur unbedeutender Unterschiede in den Aufgabenschwierigkeiten wird vom Chi-Quadrat-Test verworfen. Die Größe der Schwierigkeitsunterschiede beider Aufgaben kann mit G = 6.18 („Separation“) und R = .97 („Reliability“) als sehr gut angesehen werden. Hierbei erweist sich die Aufgabe zur Produktion möglichst vieler Zeichnungen aus verschiedenen Kategorien mit einem Schwierigkeitslogit von σ_1 = -0.12 als deutlich leichter als diejenige zur Produktion vieler verschiedener Verwendungsmöglichkeiten mit σ_2 = 0.12. Bezüglich der Modellpassung der Aufgabenfacette sind beide Aufgaben modellkonform. Global betrachtet fällt der sehr geringe mittlere Modellanpassungsfehler von *RMSE* = 0.03 auf.

Darüber hinaus zeigt keine der Aufgaben nach den MNSQ-Fit-Statistiken eine Abweichung von den Modellannahmen, beide liegen nahe an ihren Erwartungswerten und auch die z-standardisierten Werte sind nicht signifikant. Die faire Schwierigkeitseinstufung („Fair-M Avrage"), d.h., der um die Strenge der Beurteiler und der Verteilung der Probandenfähigkeit bereinigte Wert, liegt für beide Aufgaben nur knapp unter dem unadjustierten („Obsvd Average") Wert. Die Trennschärfen („PtBis") beider Aufgaben sind mit Werten um .40 befriedigend. Insgesamt lässt sich also für diese Facette eine sehr gute Modellpassung feststellen.

Tabelle 41: Statistische Fit-Analyse für die Probandenfacette Subtest Ideenflexibilität ($N = 434$)

OBSVD SCORE	OBSVD COUNT	OBSVD AVERAGE	FAIR-M AVRAGE	MODEL MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	
22.2	4.0	5.6	5.58	.00	.41	.96	-.3	.98	-.3	MEAN (COUNT: 434)
6.4	.0	1.6	1.60	1.12	.04	1.33	1.4	1.36	1.4	S.D. (POPULN)
6.5	.0	1.6	1.60	1.12	.04	1.34	1.4	1.37	1.4	S.D. (SAMPLE)

MODEL, SAMPLE: RMSE .41 ADJ (TRUE) S.D. 1.05 SEPARATION 2.53 RELIABILITY .86
MODEL, FIXED (ALL SAME) CHI-SQUARE: 3755.6 D.F.: 433 SIGNIFICANCE (PROBABILITY): .00
MODEL, RANDOM (NORMAL) CHI-SQUARE: 394.5 D.F.: 432 SIGNIFICANCE (PROBABILITY): .90

Anmerkung. *OBSVD SCORE*: Mittelwert beobachteter Gesamtbeurteilungen; *OBSVD COUNT*: Mittelwert beobachteter Beurteilungen je Aufgabe; *OBSVD AVRAGE*: Durchschnittlich beobachtete Beurteilung je Aufgabe; *FAIR-M AVRAGE*: Fairer Durchschnitt beobachtete Beurteilung je Aufgabe; *MEASURE*: Mittlerer Logit der Probandenfähigkeit; *MODEL S.E.:* Mittelwert des Standardfehlers des Logits der Probandenfähigkeit; *INFIT/OUTFIT MNSQ*: Mittelwert der Meansquare-Personenfitindizes; *ZSTD*: z-standardisierte Meansquare-Personenfitindizes.

Die Verteilung der Personenfähigkeiten zur Ideenflexibilität kann nach dem Chi-Quadrat-Test als normalverteilt angenommen werden. Die Personendifferenzierungsfähigkeit der Skala fällt mit $G = 2.53$ („Separation") und $R = .86$ („Reliability") ähnlich gut aus wie diejenige der Skala Ideenflüssigkeit (s. Tabelle 38). Mit einem Mittelwert der Personenparameter von 0.00 ist diese Kreativitätsfacette für die untersuchte Probandenstichprobe weder zu schwer noch zu leicht gewesen. Der mittlere Kalibrierungsfehler ist allerdings mit *RMSE* = .41 relativ hoch. Wie bereits in der Personenfacette der Ideenflüssigkeit liegen die Mittelwerte der Personenfit-Indizes sowohl der MNSQ- als auch der z-standardisierten Werte jeweils nahe an ihren Erwartungswerten. Allerdings fällt auch hier die jeweils hohe Standardabweichung auf. Eine Inspektion der Beurteilungsmuster anhand der Personenfit-Indizes ergibt, dass die hohe

Varianz insbesondere durch zwei entgegengesetzte Antwortmustertypen bedingt ist, welche auf eine Wechselwirkung von Probanden mit Aufgabentypen hindeutet. Ein charakteristisches Beispielantwortmuster des Typs „Anwendungsflexibilität“ war dasjenige von Proband Nr. 264 mit 4 (Kategorienmenge Zeichnungen) und 11 (Kategorienmenge Verwendungen). Dieses Antwortmuster passt insofern schlecht zu den Modellannahmen, als dass die Person mit einem Personenparameter von $\theta_{264} = 7.25$ bei der als leichter kalibrierten Aufgabe weniger produzieren konnte als bei der deutlich schwereren (s. die entsprechenden Logits der Aufgabenschwierigkeiten in Tabelle 40). Ein charakteristisches Antwortmuster vom Typ „bildende Ideenflexibilität“ (in Anlehnung an den Begriff der „bildenden Künste“) lag bei Proband Nr. 370 vor mit 12 (Kategorienmenge Zeichnungen) und 4 (Kategorienmenge Verwendungen). Hier hätte die Versuchsperson $\theta_{370} = 3.79$ aufgrund ihrer sehr hohen Mengenleistung in der ersten Aufgabe den Modellvoraussagen nach mehr als 4 verschiedene Verwendungsmöglichkeiten in der zweiten Aufgabe erbringen müssen. Es wäre eine mit einer Erweiterung des MFRM als mixed Rasch-Modell zu untersuchende Frage, ob es sich hierbei um substanzielle Typenstrukturen im Sinne unterschiedlichem Mengenproduktionsverhalten handelt. Man könnte hierbei bei dem ersteren Typ etwa an eine stärker „anwendungsbezogene“ Ideenflexibilität denken, beim zweiten hingegen an eine stärker grafisch-künstlerische. Nach Kenntnis des Autors wurde jedoch das MFRM bislang noch nicht zum mixed Rasch-Modell erweitert, sodass die Beantwortung dieser Frage zukünftiger Forschung überlassen bleibt.

Insgesamt resultiert für die Kreativitätsfacette der Ideenflexibilität eine sehr gute Passung mit den Modellannahmen des MFRM in der hier vorgenommenen Modellierung. Außer den, vom Modell aber implizierten Strengeunterschieden der Beurteiler, erweisen sich auch die Aufgaben als konform mit dem MFRM. Als problematischer zeigt sich hingegen die Probandenfacette. Hier liegen zwar die Mittelwerte der Personenfitindizes nahe an ihren Erwartungswerten, allerdings sind die jwls. hohen Standardabweichungen auffällig. Die Inspektion der Fit-Indizes zeigte zwei typisch entgegengesetzte Antwortmuster hinsichtlich der Lösungsmenge bei beiden Aufgaben. Ob sich diese gegensätzlichen Antwortmuster inhaltlich als das Vorliegen einer substanzieller Mischverteilungen hindeuten, kann nur spekuliert werden, da zur Untersuchung dieser Fragestellung kein entsprechendes Mischverteilungsmodell entwickelt wurde. Für die weiteren Analysen werden daher alle Probanden als aus einer gemeinsamen Verteilung stammend angenommen.

14.1.2.8 Subtest „empiriebezogenes Denken“

Zum besseren Verständnis der nachfolgenden Analysen seien an dieser Stelle noch einmal die Grundkategorien zur Beurteilung der Probandenleistungen dargestellt (für eine detaillierte Ausführung s. Anhang D):

a) Prinzip des Gruppenvergleichs genannt und skizziert
b) Vorkehrungen zur Störvariablenkontrolle und Fehlervarianzminimierung vorgenommen
c) Prinzipien der Verallgemeinerbarkeit auf verschiedene Personen und Variablen genannt
d) Absicherung der Ergebnisse gegenüber dem Zufall genannt und skizziert
e) Externe Validität der Ergebnisse gesichert

Bei einer ersten Analyse mit den oben genannten Kriterien zeigte sich, dass Aufgabe d) eine deutlich zu schlechte Modellpassung aufwies mit MNSQ-Infit = 1.58 bzw. ZSTD-Infit = 5.2. Die Vermutung liegt daher nahe, dass dieses Item anfällig gegenüber curricularem Vorwissen ist und somit nicht empiriebezogenes Denken erfasst als vielmehr statistisch-methodisches Wissen. Daher wurde es ausgeschlossen und eine erneute Analyse vorgenommen, deren Ergebnis im Folgenden berichtet wird.

Tabelle 42: Statistische Fit-Analyse für die Beurteilerfacette Subtest „empiriebezogenes Denken“ ($N = 434$)

OBSVD SCORE	OBSVD COUNT	OBSVD AVERAGE	FAIR-M AVRAGE	MODEL MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTBIS	BEURTEILER
3570	1580	2.3	2.08	1.05	.03	.89	-3.1	.89	-2.9	.47	1
3207	1572	2.0	1.85	1.42	.03	1.09	2.3	1.07	1.5	.41	2
3388.5	1576.0	2.1	1.96	1.24	.03	.99	-.4	.98	-.7	.44	MEAN (COUNT: 2)
256.7	5.7	.2	.16	.26	.00	.14	3.9	.13	3.2	.04	S.D. (SAMPLE)

MODEL, SAMPLE: RMSE .03 ADJ (TRUE) S.D. .25 SEPARATION 7.80 RELIABILITY .98
MODEL, FIXED (ALL SAME) CHI-SQUARE: 61.9 D.F.: 1 SIGNIFICANCE (PROBABILITY): .00
RATER AGREEMENT OPPORTUNITIES: 1724 EXACT AGREEMENTS: 1016 = 58.9%

Anmerkung. *OBSVD SCORE*: Summe der beobachteten Ratings über alle Aufgaben; *OBSVD COUNT*: Anzahl beobachteter Ratings je Beobachter; *OBSVD AVERAGE*: Durchschnittliches beobachtetes Rating des jwl. Beurteilers; *FAIR-M AVERAGE*: Fairer Durchschnitt (um Strengefehler korrigierte mittlere Beurteilung); *MEASURE*: Logit-Strengewert des jwl. Beurteilers; *MODEL S.E.:* Standardfehler des Strengeparameters; *INFIT/OUTFIT MNSQ*: Meansquare-Fit-Statistiken; *ZSTD*: z-standardisierte Meansquare-Fit-Statistiken; *PTBIS*: SR/ROR-Korrelation.

Die Beurteilerfacette zeigt nach dem *RMSE* = .03 eine sehr gute Modellpassung. Die Strengeunterschiede sind dabei deutlich ausgeprägt. Der Chi-Quadrat-Test verwirft hierbei die Annahme nicht verschiedener Strengeparameter der Beurteiler. Demgemäß zeigt Beurteiler 2 mit einem Strenge-Logit von 1.42 gegenüber Beurteiler 1 mit einem von 1.05 eine ausgeprägtere Urteilsstrenge. In Einheiten der Antwortskala ausgedrückt vergab also Beurteiler 1 im Mittel Einschätzungen, die 0.23 Punkte unter denen von Beurteiler 1 lagen. Das mittlere Ausmaß des Strengeunterschieds mag numerisch nicht groß sein, allerdings verweist gerade der Separationsindex von *G* = 7.80 („Separation“) darauf, dass die Unterschiede zwischen den Beurteilern insgesamt sehr genau feststellbar sind, was auch die Separationsreliabilität von *R* = .98 indiziert. Vergleicht man zudem die unadjustierten mittleren Beurteilungen („Obsvd Average“) der Beurteiler mit den adjustierten („Fair-M Avrage“), so liegen letztere unter den ersteren, sodass offensichtlich beide Beurteiler einen Strengefehler zeigten.
Hinsichtlich Beurteiler 1 lässt sich über die MNSQ-Fit-Statistiken und deren z-standardisierte Werte ein tendenzieller Halo-Effekt diagnostizieren, da sowohl Infit als auch Outfit mit jwls. Werten von .89 unter ihren Erwartungswerten liegen und die z-standardisierten Werte signifikant ausfallen. Beurteiler 1 neigt also dazu, einer Person über alle Aufgaben hinweg ähnliche Wertse zu vergeben, kann also tendenziell schlechter zwischen den Aufgaben unterscheiden. Allerdings liegen die MNSQ-Werte lediglich nur geringfügig unter ihren Erwartungswerten und somit noch im kritischen Intervall von 0.80 bis 1.20. Das Ausmaß dieser Modellabweichung durch den Halo-Effekt ist daher als gering anzusehen. Gleiches gilt für Beurteiler 2, welcher nach dem z-standardisierten Infit-Wert eine Tendenz zu zufälligeren Urteilen zeigt. Allerdings liegt nach dem MNSQ-Infit von 1.09 dieser Wert ebenfalls nahe am Erwartungswert und im Toleranzbereich einer guten Modellpassung.
Nach der SR/ROR-Korrelation („PtBis“) kann die Übereinstimmung als befriedigend angesehen werden.
Die Modellpassung dieser Facette kann insgesamt als gut angesehen werden, da zum einen der mittlere Anpassungsfehler sehr gering ausfällt und sich zum anderen die Modellabweichungen der Beurteiler als gering erweisen.

Tabelle 43: Statistische Fit-Analyse für die Aufgabenfacette Subtest „empiriebezogenes Denken“ (*N* = 434)

OBSVD SCORE	OBSVD COUNT	OBSVD AVERAGE	FAIR-M AVRAGE	MODEL MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	ITEMS
2017	788	2.6	2.43	-.66	.04	.97	-.5	.91	-1.8	1 GRUPPENVGL.
1587	788	2.0	1.84	.19	.05	.92	-1.5	.92	-1.3	2 STÖRV.-KONTROLLE
1805	788	2.3	2.13	-.26	.04	1.12	2.2	1.19	3.3	3 VERALLGEMEINERBARK.
1368	788	1.7	1.57	.73	.05	.91	-1.6	.90	-1.4	4 EXTERNE VALIDITÄT
1694.3	788.0	2.2	1.99	.00	.05	.98	-.4	.98	-.4	MEAN (COUNT: 4)
279.5	.0	.4	.37	.60	.00	.10	1.8	.14	2.5	S.D. (SAMPLE)

MODEL, SAMPLE: RMSE .05 ADJ (TRUE) S.D. .60 SEPARATION 12.74 RELIABILITY .99
MODEL, FIXED (ALL SAME) CHI-SQUARE: 473.8 D.F.: 3 SIGNIFICANCE (PROBABILITY): .00

Anmerkung. *OBSVD SCORE*: Summe der beobachteten Beurteilungen über alle Aufgaben; *OBSVD COUNT*: Anzahl beobachteter Beurteilungen je Aufgabe; *OBSVD AVERAGE*: Durchschnittlich beobachtete Beurteilung für jede Aufgabe; *FAIR-M AVRAGE*: Fairer Durchschnitt (um Strengefehler korrigierte Aufgabenschwierigkeit in Rohwerteinheiten); *MEASURE*: Logit der jwl. Aufgabenschwierigeit; *Model S.E.:* Standardfehler der Aufgabenschwierigkeit; *INFIT/OUTFIT MNSQ*: Meansquare-Fit-Statistiken; *Zstd*: z-standardisierte Meansquare-Fit-Statistiken; *PTBIS*: Trennschärfe der jwl. Aufgabe.

Der Chi-Quadrat-Test zur Hypothese nicht unterschiedlicher der Itemparameter zeigt eine signifikante Variation der Itemschwierigkeiten an. Der Separationsindex („Separation“) fällt mit $G = 12.74$ außerordentlich hoch aus, ebenso die Separationsreliabilität mit $R = .99$. Offensichtlich decken die Items einen großen Bereich der latenten Variable ab und vermögen in verschiedenen Fähigkeitsbereichen zu differenzieren, was auch der Range der Itemschwierigkeiten von -0.66 bis 0.73 noch einmal unterstreicht. Der mittlere Kalibrierungsfehler fällt mit *RMSE* = .05 gering aus. Die Itemfitindizes diagnostizieren lediglich für die Aufgabe zur Nennung von Vorkehrungen zur Verallgemeinerbarkeit der Aussagen einen signifikanten Missfit, sowohl nach Infit als auch nach Outfit. Diese Aufgabe zeigt relativ gesehen mit MNSQ-Werten von 1.12 bzw. 1.19 den größten Zufallsanteil. Beide MNSQ-Werte liegen jedoch im Intervall einer guten Modellpassung. Dennoch wäre in einer Testrevision der Beurteilungsmaßstab dieser Aufgabe nochmals hinsichtlich seiner Eindeutigkeit zu verbessern. Die faire Schwierigkeitseinstufung („Fair-M Avrage“), d.h., der um die Strenge der Beurteiler und der Verteilung der Probandenfähigkeit bereinigte Wert, liegt für alle drei Aufgaben nur knapp unter dem unadjustierten Wert („Obsvd Average“). Relativ am stärksten muss hierbei die mittlere Schwierigkeitseinschätzung der Aufgabe zur Nennung und Skizzierung des Gruppenvergleichsprinzips von 2.6 auf 2.43 adjustiert werden.

Die Modellpassung dieser Facette ist insgesamt gesehen als gut anzusehen. Zum einen fällt der mittlere Kalibrierungsfehler gering aus, zum anderen erweisen sich nach den MNSQ-Fit-Statistiken die einzelnen Aufgaben als modellkonform.

Tabelle 44: Statistische Fit-Kontrolle für die Personenfacette Subtest „empiriebezogenes Denken“ ($N = 434$)

OBSVD SCORE	OBSVD COUNT	OBSVD AVERAGE	FAIR-M AVRAGE	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	
16.4	8.0	2.1	2.03	-.34	.62	.98	-.1	.98	.0	MEAN (COUNT: 433)
6.1	.3	.8	.77	1.59	.41	.56	1.1	.55	1.1	S.D. (SAMPLE)

WITH EXTREMES, MODEL, SAMPLE: RMSE .74 ADJ (TRUE) S.D. 1.40 SEPARATION 1.89 RELIABILITY .78

WITHOUT EXTREMES, MODEL, SAMPLE: RMSE .52 ADJ (TRUE) S.D. 1.08 SEPARATION 2.06 RELIABILITY .81
WITH EXTREMES, MODEL, FIXED (ALL SAME) CHI-SQUARE: 2073.9 D.F.: 432 SIGNIFICANCE (PROBABILITY): .00
WITH EXTREMES, MODEL, RANDOM (NORMAL) CHI-SQUARE: 316.5 D.F.: 431 SIGNIFICANCE (PROBABILITY): 1.00

Anmerkung. *OBSVD SCORE*: Mittelwert beobachteter Gesamtbeurteilungen; *OBSVD COUNT*: Mittelwert beobachteter Beurteilungen je Aufgabe; *OBSVD AVERAGE*: Durchschnittlich beobachtete Beurteilung je Aufgabe; *FAIR-M AVRAGE*: Fairer Durchschnitt beobachtete Beurteilung je Aufgabe; *MEASURE*: Mittlerer Logit der Probandenfähigkeit; *MODEL S.E.:* Mittelwert des Standardfehlers des Logits der Probandenfähigkeit; *INFIT/OUTFIT MNSQ*: Mittelwert der Meansquare-Personenfitindizes; *Zstd*: z-standardisierte Meansquare-Personenfitindizes.

Aus der Perspektive der Parameterschätzung als problematisch erwies sich für die Probandenfacette das Vorliegen von Nulllösungen bei 37 Probanden. Solchen Minimumwerten können mathematisch keine Schätzungen auf der latenten Variablen zugeordnet werden, da sie einen unendlichen Wert implizieren und führen daher zu verfälschten Schätzungen der Modellparameter. Das Programm Facets schätzte daher die Personenparameter und Personen-Fit-Statistiken ohne diese Extremwerte, die übrigen Werte wie *RMSE* etc. werden jedoch im Folgenden sowohl mit als auch ohne Extremwerte ausgegeben, um dem Leser eine Abschätzung der Verzerrung zu ermöglichen. Diskutiert werden jedoch nur die Ergebnisse ohne Extremwerte.

Die Verteilung der Personenfähigkeiten der des Subtests „empiriebezogenes Denken“ kann nach dem Chi-Quadrat-Test als normalverteilt angenommen werden. Die Hypothese auf nicht unterschiedliche Fähigkeitswerte wird verworfen. Der Separationsindex mit $G = 2.06$ („Separation“) und die Separationsreliabilität mit $R = .81$ verweisen auf eine befriedigende

Unterscheidungsfähigkeit der Probanden. Der Mittelwert der Personenparameter liegt bei -.0.34. Insgesamt war dieser Test also für die Probandenstichprobe eher zu schwer. Dies wird in Einheiten der Ratingskala ebenso offensichtlich, betrachtet man die niedrige mittlere Beurteilung („Obsvd Average“) von 2.1. Der adjustierte faire Durchschnitt („Fair-M Avrage“) liegt mit 2.03 nur knapp darunter. Wie bereits in den Analysen der Kreativitätsfacetten zeigt die Probandenfacette den relativ größten Kalibrierungsfehler mit *RMSE* = .52. Die Mittelwerte der Personenfit-Indizes liegen nahe an ihren jeweiligen Erwartungswerten, allerdings legt die Betrachtung der Standardabweichungen eine Analyse der Personenfitindizes nahe. Sie ergab, dass die Varianz in den Personenfit-Statistiken überwiegend durch Antwortmuster produziert worden waren, die am ehesten als Symptom von „Achtlosigkeit“ beschrieben werden können und sowohl durch hohe Infit- als auch hohe Outfit-Werte gekennzeichnet war. Ein hierfür typisches Beurteilungsmuster (welches zudem von beiden Beurteilern übereinstimmend vergeben wurde) war 5,1,1,1 (Proband 51). Nach Modellerwartungen hätte diese Person mit einem Personenparameter von $\theta_{51} = 0.31$ gerade in der zweiten und dritten Aufgabe bessere Leistungen erbringen müssen, da deren Schwierigkeitswerte mit $\sigma_2 = 0.19$ und $\sigma_3 = -0.26$ (vgl. Tabelle 43) deutlich unter Fähigkeitswert des Probanden lagen. Ein möglicher Grund für diese „Achtlosigkeit“ könnte eine geringere Testmotivation sein, zumal diese Aufgabe eine der letzten der Testbatterie überhaupt war.
Die Beurteilung der Modellgültigkeit dieser Facette ergibt insgesamt eine befriedigende Passung. Als problematisch erweist sich das Vorliegen von Antwortmustern, die auf Achtlosigkeit bei der Beantwortung der Teilaufgaben hinweisen und den mittleren Kalibrierungsfehler negativ beeinflussten. Die Reliabilität der Personenunterschiede kann anhand des Separationsindexes und der Separationsreliabilität als befriedigend angesehen werden.

14.1.2.8.1 Exploratorische Bias-Analyse Subtest „empiriebezogenes Denken“

Ein letzter Evaluationsschritt, der auch die Weiterverarbeitung dieser Daten hinsichtlich der Untersuchung von Leistungstestunterschieden zwischen den Teilnehmern aus verschiedenen Studienfächern betraf, war die explorative Analyse von Interaktionseffekten zwischen Beurteilern und Probandengruppen. Dies ist deshalb unerlässlich, weil etwaige signifikante Mittelwertsunterschiede zwischen Studierenden des ersten Semesters in Psychologie und denen anderer Studienfächer inhaltlich bedeutungslos wären, sollten etwa die Beurteiler die Gruppe der Psychologen milder beurteilt haben. Eine Anfälligkeit der Beurteiler in Bezug auf derartige differenzielle Strenge- oder Mildeeffekte würde die Objektivität des Verfahrens korrumpieren.

Daher wurde das bestehende Modell (Gleichung (36)) um einen Interaktionsterm Beurteiler × Studienfach erweitert. Hierzu wurde eine Dummy-Variable gebildet, welche Psychologiestudenten und Teilnehmer aller anderen Studienfächer codierte. Sollte hierbei der Interaktionsterm signifikant werden, stünde ein z-Wert größer als 1.96 für eine signifikant größere Strenge des Beurteilers gegenüber dieser Gruppe, wohingegen ein Wert kleiner als -1.96 eine größere Milde dieser Gruppe gegenüber indizieren würde. Die Beurteilung würde in diesen Fällen einer systematischen Verzerrung (Bias) unterliegen. Tabelle 45 gibt einen Überblick über die Ergebnisse der Bias-Analyse.

Tabelle 45: Bias-Analyse: Beurteiler × Studienfach-Interaktion Subtest „empiriebezogenes Denken“ ($N = 434$)

OBSVD SCORE	EXP. SCORE	OBSVD COUNT	OBS-EXP AVERAGE	BIAS SIZE	MODEL S.E.	Z	P-VALUE	BEURTEILER	GRUPPE
833	830.2	512	.01	.02	.08	.21	.83	1	1 PSYCHOLOGEN
838	836.2	488	.00	.01	.07	.13	.89	2	2 ANDERE FÄCHER
1027	1030.1	536	-.01	-.01	.06	-.20	.84	1	2 ANDERE FÄCHER
816	818.1	548	.00	-.01	.08	-.17	.86	2	1 PSYCHOLOGEN
878.5	878.7	521.0	.00	.00	.07	-.01		MEAN (COUNT: 4)	
99.4	101.3	26.6	.01	.02	.01	.21		S.D. (SAMPLE)	

FIXED (ALL = 0) CHI-SQUARE: .1 D.F.: 4 SIGNIFICANCE (PROBABILITY): 1.00

Anmerkung. *OBSVD SCORE*: Mittelwert beobachteter Gesamtbeurteilungen; *EXP. SCORE*: Nach den Modellannahmen *ohne* Interaktion erwarteter Rohwert; *OBSVD COUNT*: Beobachtete Beurteilung je Fach und Beurteiler; *OBSV-EXP AVERAGE*: Differenz beobachteter mittlerer Beurteilungen zur mittleren vom Modell ohne Interaktion erwarteten Beurteilung; *BIAS-SIZE*: Größe der Verzerrung der Beurteilung durch Interaktion in Logit-Einheit; *MODEL S.E.:* Mittelwert des Standardfehlers des Logits der Verzerrung; *z*: z-Wert zur Hypothese, dass kein Bias vorliegt; *P-VALUE*: p-Wert zur Hypothese, dass kein Bias vorliegt; *Beurteiler*: Dummy-Code der Beurteiler; *GRUPPE*: Dummy-Code der Studienfachgruppe.

Bereits der Chi-Quadrat-Test zur Hypothese, dass die Interaktionseffekte nur unwesentlich von Null verschieden sind, zeigt an, dass kein differentieller Strenge- oder Mildeeffekt vorliegt. Keiner der Bias-Terme („Bias Size“) erweist sich als signifikant. Ebenso zeigen sich auf rein deskriptiver Ebene der Differenzen der beobachteten („Obsvd Score“) und der erwarteten Beurteilungsrohwerte („Exp. Score“) keinerlei Effekte differenzieller Strenge oder Milde gegenüber Studierenden der Psychologie und anderer Fächer. Eine Bevorzugung von Psychologiestudierenden kann damit ausgeschlossen werden. Die Objektivität des Beurteilers

wurde also durch das Wissen um die Zugehörigkeit des Probanden zum Studienfach Psychologie gegenüber anderen Fächern nicht beeinflusst.

Insgesamt betrachtet kann eine gute Modellpassung der Aufgabe zum empirischen Denken festgestellt werden. Mit Ausnahme der Teilaufgabe, eine signifikanzstatistische Absicherung der Gruppenunterschiede zu nennen und zu skizzieren, erwiesen sich die übrigen Aufgaben als modellkonform. Hierbei müsste allerdings die Teilaufgabe zur Verallgemeinerbarkeit der Ergebnisse im Hinblick auf die Klarheit der Beurteilungskriterien optimiert werden, da sie die relativ größte Modellabweichung zeigte, wobei diese von ihrer Effektstärke her betrachtet nicht beträchtlich war. Die Differenzierungsfähigkeit der Aufgabenfacette ist als sehr gut anzusehen, da zum einen die Separationsstatistiken sehr hoch lagen und zum anderen Teilaufgaben über einen breiten Bereich der Logitskala streuten. Die Strengeunterschiede der Beurteiler waren substanziell und belegten damit die Notwendigkeit des hierzu in der Modellgleichung spezifizierten Strengeparameters. Wenn auch die Differenz zwischen den Beurteilern nicht sonderlich groß war, so darf dies nicht auf weitere potenzielle Beurteiler generalisiert werden. Im Rahmen von High-Stakes-Testungen müssten derartige Aufgabentypen von mehr als zwei Personen beurteilt werden, was größere Strengedifferenzen sehr viel wahrscheinlicher macht, zumal mit weiteren Beobachterfehlern zu rechnen ist. In diesem Zusammenhang zeigte Beurteiler 1 die Tendenz zu einem Halo-Effekt, dessen Auswirkungen allerdings sehr gering waren.
Für die Probandenfacette zeigte sich wie schon bei den Kreativitätsfacetten die relativ größte Modellabweichung. Diese konnte insbesondere durch achtlose Antworten einiger Probanden erklärt werden. Probleme der Testmotivation mögen hierfür eine Ursache gewesen sein.

14.1.3. Fazit zu den Ergebnissen der Modellgeltungstests

Aufseiten der Subtests im Multiple-Choice-Format bedarf jede der sechs Testvorversionen einer gründlichen Testrevision. Als problematisch erwies sich für die Subtests verbale Analogien, Zahlenreihen und Matrizen das Vorliegen von Speed-Effekten, welche die Eindimensionalität der Tests zerstörten. Zwar war die Größe der Modellabweichungen einzelner Items durch Speed-Effekte laut der MNSQ-Item-Fit-Indizes nicht bedeutsam, jedoch waren zu viele Items davon betroffen, sodass der Speed-Effekt im Weiteren berücksichtigt werden musste. Zudem schlug sich der Effekt in einer besser passenden Zweiklassen-Lösung des mixed Rasch-Modells mit jwls. einer bearbeitungsschnelleren und einer

bearbeitungslangsameren Klasse nieder, was den Befund auch auf Personenebene unterstützte. Eine getrennte Auswertung in weiteren Analysen hinsichtlich kriteriumsbezogener Validitäten wäre zwar über das mixed Rasch-Modell möglich gewesen, hätte allerdings die Vergleichbarkeit dieser Ergebnisse und deren Generalisierbarkeit erschwert. Daher wurden die jeweils von der zu strikten Testzeitbegrenzung betroffenen Subtests mit einem Rasch-Modell für Speed-Effekte ausgewertet, um deren Einfluss kontrollieren zu können. Hierdurch ergab sich eine jwls. deutlich verbesserte Modellgeltung.
Bei den nicht von Speed-Effekten betroffenen Subtests Even-Odd-One-Out und Zahlenmatrizen zeigten einige Items meist aufgrund zu geringer Trennschärfe eine schlechte Modellpassung und mussten ausgesondert werden. Im Subtest Zahlenmatrizen und SPARK waren zudem die Itemantworten auf einigen Items nicht lokal stochastisch voneinander unabhängig.

Hinsichtlich der Personenseparationsreliabilität zeigten die Subtests eine ausreichende bis befriedigende Messgenauigkeit. Die entsprechenden Kennwerte aufseiten der Itemseparation waren hingegen durchweg sehr gut. Da allerdings die hier konstruierten Tests für den Selektionskontext konzipiert sind, wäre die Personenseparationsreliabilität durch geeignete Maßnahmen wie der Ersetzung zu leichter Items durch schwerere und einer Erhöhung der Testzeit zu verbessern. Allerdings darf dies nicht zulasten der hier bereits sehr guten Itemseparationsstatistiken gehen, will man das zugrunde liegende Konstrukt als Skala und nicht als eng umgrenzten Bereich erfassen.

Seitens der Aufgaben im freien Antwortformat ergibt sich für die Kreativitätsfacetten „Ideenflüssigkeit“ und „Ideenflexibilität“ insgesamt betrachtet eine befriedigende bis gute Modellgeltung. Als problematischer erwies sich allerdings die Aufgabe zur „Produktion möglichst vieler Verwendungen“ für den Begriff „Schaumstoffpolster“, weil hier vermutlich spezifische Wissensstrukturen die Eindimensionalität zerstörten und somit nicht alleine Ideenflüssigkeit erfasst wurde. Demgegenüber war die Eindimensionalität der Aufgaben zur Facette Ideenflexibilität deutlich unproblematischer, da hier keine Modellabweichungen auftraten. Die in beiden Facetten auftretende Personenheterogenität war im Falle der Ideenflüssigkeit überwiegend durch die Aufgabe zum Begriff „Schaumstoffpolster“ bedingt, im Falle der Ideenflexibilität jedoch struktureller, d.h. in Form von zwei möglicherweise vorhandenen Typenstrukturen. Dies wäre jedoch mit entsprechenden Mischverteilungsmodellen nachzuweisen, die nach Kenntnis des Autors noch nicht für das MFRM entwickelt wurden. Die Notwendigkeit der Korrektur des Strenge- bzw. Mildefehlers der Beurteiler durch den im

Modell spezifizierten Strengeparameter zeigte sich für beide Kreativitätsfacetten, wenn auch der Einfluss dieser Urteilerfehler auf die durchschnittlichen Beurteilungen nicht groß war.

Bei der Aufgabe zum empiriebezogenen Denken erwies sich die Teilaufgabe zur inferenzstatistischen Absicherung als nicht modellkonform und musste aus dem Test entfernt werden. Danach zeigte sich eine insgesamt gute Modellanpassung der Aufgabenfacette. Wie durch die Modellspezifikation hypothetisiert, waren die Beurteiler unterschiedlich streng. Eine explorative Analyse differenzieller Strengeeffekte gegenüber Studierenden anderer Fächer als denen der Psychologie ergab keinen Hinweis auf eine Bevorteilung der letzteren. Dieser Nachweis ist gerade hinsichtlich valider, d.h. unverzerrter Leistungsvergleiche zwischen Studienfächern in dieser Aufgabe wesentlich. Als problematischer erwies sich wie bereits bei den Kreativitätsaufgaben die Probandenfacette. Hierbei waren Antwortmuster identifizierbar, die am ehesten auf Achtlosigkeit oder Motivationsprobleme zurückgeführt werden können. Vermutlich zeigt sich hierin ein Effekt der Länge des Gesamtverfahrens.

15. Kreuzvalidierungen der Skalenanalysen

Die im vorigen Kapitel dargestellten Modellgeltungsanalysen und die damit verbundenen Itemselektionen bergen für die weiterhin durchzuführende Kriteriumsvalidierungen der Skalen an derselben Stichprobe bzw. Teilmengen hieraus die Gefahr, dass die Itemselektionen zu sehr an die Gegebenheiten der Stichprobe angepasst sind. Die Verallgemeinerbarkeit der Validierungsergebnisse der Skalen fiele dadurch in einem unbekannten Ausmaß unsicher aus. Eine Kreuzvalidierung der Itemselektionsergebnisse an einer weiteren Stichprobe vom selben Umfang war jedoch aus zeitlichen, finanziellen und nicht zuletzt organisationellen Gründen nicht möglich. Daher wurde im vorliegenden Fall das Vorgehen einer Kreuzvalidierung der Skalenanalysen anhand einer Zufallsaufteilung der Gesamtstichprobe ($N = 434$) in zwei Stichproben von annähernd gleichem Umfang ($n_A = 207$, $n_B = 227$) gewählt. Deren skalenanalytischen Befunde werden zentral zum einen mit der Gesamtstichprobe verglichen, zum anderen auch in Bezug auf ihre Übereinstimmungen miteinander.
Der so durchgeführte Vergleich der Skalenanalyse-Ergebnisse fokussiert sowohl auf die Frage nach der Übereinstimmung hinsichtlich der *Identifizierung passender* bzw. *nicht passender Items* anhand der Itemfitindizes als auch auf die *Invarianz der Itemparameter* und auf die *Konstanz der Ergebnisse der Personenhomogenität*.

Prinzipiell problematisch an dieser Form der Kreuzvalidierung ist die Verringerung der Teststärke durch die Halbierung der Personenstichprobe, insbesondere bei der Identifikation nicht modellkonformer Items über die z-standardisierten Meansquare-Fit-Indizes Infit und Outfit: in der Gesamtanalyse als modellinkonform identifizierte Items haben in den kleineren Zufallsstichproben eine geringere Wahrscheinlichkeit von dem Modellerwartungen als abweichend erkannt zu werden. Diesem Problem soll im vorliegenden Fall dadurch begegnet werden, dass bei der Itemauswahl nicht an erster Stelle die z-standardisierten Meansquare--Fit-Statistiken verglichen werden, sondern die in dieser Arbeit ohnehin präferierten unstandardisierten Statistiken, welche als Effektstärkemaße der Modellabweichungen definiert sind. Wie in der Gesamtanalyse gilt hierbei der Toleranzbereich guter Modellpassung von 0.8 bis 1.2. Als Maß globaler Übereinstimmung zwischen den einzelnen Stichproben dient die Korrelation zwischen den unstandardisierten Meansquare-Fit-Statistiken. Für die Überprüfung der Invarianz der Itemparameter dient die Korrelation der Itemparameter zwischen den einzelnen Stichproben. Zum Vergleich der Ergebnisse der Personenhomogenität über das mixed Rasch-Modell fungieren (wie bereits in der Gesamtanalyse) an erster Stelle die informationstheoretischen Maße BIC und CAIC und ferner der Likelihood-Quotiententests. Da das mixed Rasch-Modell noch nicht für Aufgaben im freien Antwortformat generalisiert worden ist, wird bei den Aufgabenstellungen zu Kreativitätsfacetten und dem „empiriebezogenem Denken“ auf die Übereinstimmung der Item-Fit-Indizes und der Itemparameter fokussiert.

Angesichts des ohnehin großen Umfangs der Skalenanalysen soll die Darstellung der Kreuzvalidierungsergebnisse in der gebotenen Kürze und somit überblicksartig erfolgen.

15.1 Kreuzvalidierungsanalysen zum Subtest „verbale Analogien“

Tabelle 46 gibt einen vergleichenden Überblick über die Kreuzvalidierungsergebnisse zur Modellpassung auf Itemebene.

Tabelle 46: Item-Fit-Kreuzvalidierungsergebnisse Subtest „verbale Analogien“ (N_{Gesamt} = 434, n_A = 207, n_B = 227)

Item	MNSQ Infit A	MNSQ Infit B	MNSQ Infit Gesamt	ZSTD Infit A	ZSTD Infit B	ZSTD Infit Gesamt	MNSQ Outfit A	MNSQ Outfit B	MNSQ Outfit Gesamt	ZSTD Outfit A	ZSTD Outfit B	ZSTD Outfit Gesamt
A1	1.13	1.04	1.08	0.8	0.3	0.7	1.77	1.47	1.60	2.1	1.4	2.3
A2	1.18	1.15	1.17	2.3	2.1	3.2	1.26	1.26	1.28	1.9	2.1	3.1
A3	1.21	1.04	1.11	3.5	0.7	3.0	1.35	1.04	1.18	4.1	0.5	3.2
A4	1.12	1.13	1.12	1.6	2.0	2.5	1.12	1.19	1.16	1.1	1.8	2.0
A5	1.13	1.10	1.12	1.6	1.7	2.6	1.16	1.12	1.15	1.3	1.5	2.1
A6	1.10	1.18	1.14	1.8	3.4	3.6	1.16	1.29	1.23	2.0	3.6	4.0
A7	0.99	1.10	1.05	0.0	1.1	0.8	1.04	1.20	1.12	0.3	1.5	1.3
A8	1.07	1.09	1.08	1.1	1.2	1.5	1.23	1.17	1.20	2.1	1.4	2.4
A9	0.98	1.08	1.03	-0.3	1.1	0.6	1.02	1.15	1.09	0.2	1.2	1.0
A10	0.95	0.93	0.94	-0.8	1.4	-1.5	0.93	0.91	0.92	-0.7	-1.2	-1.3
A11	0.79	0.84	0.81	-2.4	1.7	-3.0	0.65	0.75	0.69	-2.5	-1.7	-3.1
A12	1.14	1.03	1.08	1.9	0.5	1.5	1.31	1.06	1.17	2.4	0.5	1.9
A13	1.00	0.99	1.00	0.0	-0.1	0.0	0.91	0.89	0.94	-0.6	-0.7	-0.5
A14	0.92	0.88	0.90	-1.0	1.9	-2.0	0.84	0.81	0.83	-1.2	-1.8	-2.2
A15	0.77	0.82	0.79	-3.9	3.6	-5.3	0.69	0.76	0.73	-3.3	-3.0	-4.5
A16	0.76	0.82	0.79	-3.2	2.3	-3.8	0.65	0.71	0.68	-2.8	-2.2	-3.6
A17	0.99	0.94	0.96	0.2	-0.2	-0.1	0.63	0.60	0.60	-0.2	-1.1	-1.1
A18	0.85	0.89	0.87	-0.7	-0.6	-0.9	0.51	0.59	0.56	-1.6	-1.5	-2.3
A19	0.81	0.87	0.84	-1.2	-0.8	-1.4	0.66	0.63	0.65	-1.3	-1.4	-2.0
A20	0.85	0.92	0.89	-0.6	-0.3	-0.7	0.45	0.60	0.53	-1.7	-1.2	-2.1
r_{AB}	.86						.91					
$r_{Gesamt\ A}$			.96						.97			
$r_{Gesamt\ B}$			.96						.97			

Anmerkung. MNSQ Infit A: Meansquare-Infit-Statistiken in Zufallsstichprobe A; MNSQ Infit B: Meansquare-Infit-Statistiken in Zufallsstichprobe B; MNSQ Infit Gesamt: Meansquare-Infit-Statistiken in Gesamtstichprobe; ZSTD Infit A: z-standardisierte Meansquare-Infit-Statistiken in Zufallsstichprobe A; ZSTD Infit B: z-standardisierte Meansquare-Infit-Statistiken in Zufallsstichprobe B; ZSTD Infit Gesamt: z-standardisierte Meansquare-Infit-Statistiken in Gesamtstichprobe;
MNSQ Outfit A: Meansquare-Outfit-Statistiken in Zufallsstichprobe A, MNSQ Outfit B: Meansquare-Outfit-Statistiken in Zufallsstichprobe B; MNSQ Outfit Gesamt: Meansquare-Outfit-Statistiken in Gesamtstichprobe; ZSTD Outfit A: z-standardisierte Meansquare-Outfit-Statistiken in Zufallsstichprobe A; ZSTD Outfit B: z-standardisierte Meansquare-Outfit-Statistiken in Zufallsstichprobe B; ZSTD Outfit Gesamt: z-standardisierte Meansquare-Outfit-Statistiken in Gesamtstichprobe;
r_{AB}: Korrelation der Meansquare-Fit-Statistiken aus Zufallsstichprobe A und B; $r_{GesamtA}$: Korrelation der Meansquare-Fit-Statistiken aus Gesamtstichprobe mit Zufallsstichprobe A; $r_{GesamtB}$: Korrelation der Meansquare-Fit-Statistiken aus Gesamtstichprobe mit Zufallsstichprobe B

Die Korrelationen der Meansquare-Itemfit-Statistiken fallen sowohl zwischen den Zufallsstichproben als auch verglichen mit der Gesamtstichprobe jeweils sehr hoch aus. Bei den z-standardisierten Meansquare-Werte tendiert die Gesamtanalyse wegen der stichprobengrößenbedingten höheren Teststärke zu einer strengeren Identifikation von Modellabweichungen einzelner Items.
Auch die Korrelationen der Itemparameter zwischen den Stichproben nahmen mit $r_{AB} = .97$ und $r_{Gesamt\,A} = .99$ bzw. $r_{Gesamt} = .99$ jeweils sehr hohe Werte an, wodurch die Itemparameterinvarianz über die jeweiligen Stichproben hinweg betrachtet angenommen werden kann.

Im Weiteren muss der Frage nachgegangen werden, ob sich der Speed-Effekt, der sich in diesem Subtest in der Gesamtanalyse als bessere Passung der Zweiklassen- gegenüber der Einklassenlösung des mixed Rasch-Modells zeigte, auch in den beiden Zufallsstichproben ergibt. Tabelle 47 stellt hierzu die Ergebnisse der mixed Rasch-Analysen vergleichend gegenüber.

Tabelle 47: Kreuzvalidierungsergebnisse der mixed Rasch-Analysen Subtest „verbale Analogien“

Stichprobe	**Klassenanzahl**	**BIC**	**CAIC**	**Log L**	**-2(log(cL_{RM}) -log(cL_{2KL}))**
A	1	4317.63	4338.63	-2102.82	189.06**
	2	4245.88	4288.88	-2008.29	
B	1	4818.55	4839.55	-2352.31	251.62**
	2	4686.28	4729.28	-2226.5	
Gesamtstichprobe	1	8799.42	8820.42	-4335.94	444.9**
	2	8488.11	8531.11	-4113.49	

Anmerkung: Log L: Logarithmierte Likelihood der Daten unter den jwl. Modellannahmen; BIC: Bayes Information Criterion; CAIC: Consistent Akaikes Information Criterion; cL_{RM}: Bedingte Likelihood des dichotomen Rasch-Modells; cL_{2KL}: Bedingte Likelihood des mixed Rasch-Modells in zwei Klassen; -2(log(cL_{RM})-log(cL_{2KL})): Bedingter Likelihood-Quotiententest;
**: $p < .01$

Wie in der Gesamtstichprobe resultiert im Vergleich von BIC und CAIC in beiden Zufallsstichproben eine bessere Modellpassung der Zweiklassenlösung, die sich zudem als signifikant erweist. Jedoch beantwortet dies noch nicht die Frage, ob die jeweiligen Zweiklassenlösungen von derselben Art der Personenheterogenität wie in der Gesamtstichprobe sind. Hierzu müssen

wie bereits in der Gesamtstichprobenanalyse die Itemparameter der beiden Klassen in den Zufallsstichproben miteinander verglichen werden. Tabelle 48 zeigt die Ergebnisse.

Tabelle 48: Itemparametervergleich mixed Rasch-Analysen Subtest „verbale Analogien“

Item	**Itemparameter Klasse 1A**	**Itemparameter Klasse 2A**	**Itemparameter Klasse 1B**	**Itemparameter Klasse 2B**
A1	-3.76	-1.70	-3.80	-2.30
A2	-0.43	1.42	-0.54	1.19
A3	-1.13	0.14	-1.25	-0.14
A4	-0.27	0.72	-0.67	1.00
A5	-2.10	-0.91	-1.98	0.14
A6	-1.16	0.44	-1.52	0.24
A7	-2.17	-1.42	-2.39	-1.01
A8	-2.01	-0.15	-2.33	-0.76
A9	-2.03	-1.03	-2.37	-0.60
A10	-1.52	-0.85	-1.46	-0.40
A11	-2.18	-2.36	-2.11	-2.42
A12	-0.25	0.89	0.09	0.82
A13	0.30	0.72	0.69	0.77
A14	0.42	0.18	0.73	-0.31
A15	0.73	-1.44	1.82	-2.31
A16	2.85	-1.40	3.15	-0.30
A17	3.81	3.07	3.76	1.91
A18	3.77	1.28	3.75	1.41
A19	3.26	0.90	2.93	1.31
A20	3.86	1.52	3.47	1.77
Klassen-größe	70%	30%	59%	41%

In beiden Zufallsstichproben lassen sich über die Betrachtung der Itemparameter, wie bereits in der Gesamtstichprobenanalyse, zwei Personenklassen identifizieren, wovon die jeweils erste (Klasse 1A bzw. Klasse 1B) die bearbeitungslangsamere ist, die jeweils zweite (Klasse 2A bzw. Klasse 2B) hingegen die bearbeitungsschnellere. Deutlich erkennbar ist dies an den höheren Itemschwierigkeiten der Items in der jeweils ersten Klasse (der somit bearbeitungs-langsameren) gegenüber der zweiten (der bearbeitungsschnelleren) ab Item A12.

Auf Grundlage der gelungenen Kreuzreplikation der Personenheterogenität ist es wie bereits in der Analyse der Gesamtstichprobe nötig, die dort vorgenommene Reformulierung dieses Subtests als Rasch-Modell mit einer zusätzlich spezifizierten Speed-Komponente nach Gleichung (47) und der daraus resultierenden Itemauswahl zu kreuzvalidieren. Die Modell-parameter wurden daher auch für beide Zufallsstichproben berechnet. Die Ergebnisse zeigt Tabelle 49.

Tabelle 49: Item-Fit-Kreuzvalidierungsergebnisse Subtest „verbale Analogien“ nach Modellreformulierung (N_{Gesamt} = 434, n_A = 207, n_B = 227)

Item	MNSQ Infit A	MNSQ Infit B	MNSQ Infit Gesamt	ZSTD Infit A	ZSTD Infit B	ZSTD Gesamt	MNSQ Outfit A	MNSQ Outfit B	MNSQ Outfit Gesamt	ZSTD Outfit A	ZSTD Outfit B	ZSTD Outfit Gesamt
A1	1.03	0.97	1.05	0.2	-0.1	0.4	1.25	1.22	1.30	0.7	0.7	1.1
A2	1.08	1.07	1.12	1.3	1.2	**2.6**	1.13	1.09	1.17	0.5	0.4	0.7
A3	1.03	1.01	1.04	0.9	0.2	1.4	1.03	1.00	1.04	0.2	0.1	0.3
A4	1.06	1.05	1.08	1.0	1.0	1.5	1.08	1.06	1.11	0.3	0.3	0.5
A5	1.03	1.03	1.08	0.5	0.7	1.5	1.04	1.03	1.11	0.2	0.2	0.5
A6	1.09	1.08	1.12	2.5	2.4	**4.6**	1.11	1.07	1.14	0.4	0.3	0.6
A7	1.00	1.03	1.03	0.0	0.4	0.5	1.02	1.06	1.07	0.2	0.3	0.3
A8	1.07	1.07	1.07	1.3	0.9	1.6	1.11	1.10	1.12	0.4	0.4	0.5
A9	1.02	1.03	1.02	0.3	0.4	0.4	1.04	1.03	1.04	0.2	0.2	0.2
A10	1.01	0.99	1.00	0.2	-0.2	-0.1	1.02	1.00	1.01	0.2	0.1	0.1
A11	0.9	0.89	0.87	-1	-1.3	-1.9	0.82	0.83	0.81	-0.4	-0.4	-0.7
A12	1.05	0.99	1.01	0.8	-0.1	0.2	1.06	0.97	0.99	0.3	0.0	0.1
A13	0.99	0.98	0.95	-0.1	-0.3	-0.8	0.97	0.95	0.92	0.0	-0.1	-0.3
A14	0.96	0.94	0.93	-0.7	-1.6	-1.9	0.94	0.92	0.9	0.0	-0.1	-0.3
A15	0.81	0.89	0.82	-5.1	-2.6	**-6.8**	0.8	0.87	0.79	-0.5	-0.3	-0.8
A16	0.82	0.88	0.83	-2.6	-2.7	**-4.2**	0.78	0.85	0.79	-0.6	-0.3	-0.9
A17	1.01	0.97	0.98	0.2	-0.2	-0.1	1.02	0.95	0.92	0.2	0.0	-0.2
A18	0.9	0.92	0.89	-0.4	-0.6	-1.0	0.69	0.76	0.77	-0.7	-0.6	-0.9
A19	0.9	0.91	0.87	-0.8	-0.8	-1.6	0.85	0.87	0.82	-0.3	-0.2	-0.6
A20	0.89	0.95	0.93	-0.7	-0.4	-0.6	0.82	0.80	0.84	-0.4	-0.3	-0.5
r_{AB}		.91						.96				
$r_{GesamtA}$			.95						.95			
$r_{GesamtB}$			.96						.97			

Anmerkung. MNSQ Infit A: Meansquare-Infit-Statistiken in Zufallsstichprobe A; MNSQ Infit B: Meansquare-Infit-Statistiken in Zufallsstichprobe B; MNSQ Infit Gesamt: Meansquare-Infit-Statistiken in Gesamtstichprobe; ZSTD Infit A: z-standardisierte Meansquare-Infit-Statistiken in Zufallsstichprobe A; ZSTD Infit B: z-standardisierte Meansquare-Infit-Statistiken in Zufallsstichprobe B; ZSTD Infit Gesamt: z-standardisierte Meansquare-Infit-Statistiken in Gesamtstichprobe;
MNSQ Outfit A: Meansquare-Outfit-Statistiken in Zufallsstichprobe A, MNSQ Outfit B: Meansquare-Outfit-Statistiken in Zufallsstichprobe B; MNSQ Outfit Gesamt: Meansquare-Outfit-Statistiken in Gesamtstichprobe; ZSTD Outfit A: z-standardisierte Meansquare-Outfit-Statistiken in Zufallsstichprobe A; ZSTD Outfit B: z-standardisierte Meansquare-Outfit-Statistiken in Zufallsstichprobe B; ZSTD Outfit Gesamt: z-standardisierte Meansquare-Outfit-Statistiken in Gesamtstichprobe;
r_{AB}: Korrelation der Meansquare-Fit-Statistiken aus Zufallsstichprobe A und B; $r_{GesamtA}$: Korrelation der Meansquare-Fit-Statistiken aus Gesamtstichprobe mit Zufallsstichprobe A; $r_{GesamtB}$: Korrelation der Meansquare-Fit-Statistiken aus Gesamtstichprobe mit Zufallsstichprobe B

Wie anhand der jeweils sehr hohen Korrelationen zwischen den Meansquare-Werten der Zufallsstichproben und der Gesamtstichprobe ersichtlich, resultieren jeweils identische Ergebnisse der Item-Modellpassung. Die in Kapitel 14.1.2.1 anhand der Gesamtstichprobe über die signifikanten z-standardisierten Infit-Werte vorgenommen Itemselektion (in der Tabelle fett markiert) erweist sich im Falle von Item A2 wegen höherer Teststärke zudem als strenger.

Auch die Invarianz der Itemparameter im Vergleich über die Stichproben kann angenommen werden, da deren Korrelationen mit $r_{A\,B} = .96$ und $r_{Gesamt\,A} = .98$ bzw. $r_{Gesamt\,B} = .99$ jeweils sehr hoch ausfielen.

15.2 Kreuzvalidierungsanalysen zum Subtest „Odd-Even-One-Out"

Tabelle 50 gibt einen vergleichenden Überblick über die Kreuzvalidierungsergebnisse der Item-Modellpassung.

Tabelle 50: Item-Fit-Kreuzvalidierungsergebnisse Subtest „Odd-Even-One-Out"($N_{Gesamt} = 434$, $n_A = 207$, $n_B = 227$)

Item	MNSQ Infit A	MNSQ Infit B	MNSQ Infit Gesamt	ZSTD Infit A	ZSTD Infit B	ZSTD Infit Gesamt	MNSQ Outfit A	MNSQ Outfit B	MNSQ Outfit Gesamt	ZSTD Outfit A	ZSTD Outfit B	ZSTD Outfit Gesamt
O21	1.20	1.16	1.17	2.3	2.4	**3.2**	2.33	2.65	2.05	9.5	9.6	**9.5**
O22	0.80	0.82	0.88	-0.5	-0.4	-0.2	0.78	0.72	0.77	-0.2	-0.4	-0.2
O23	0.95	1.1	1.03	-0.5	1.2	0.5	0.93	1.07	1.01	-0.5	0.6	0.1
O24	1.08	1.05	1.07	1.5	0.9	1.7	1.23	1.12	1.18	2.7	1.1	**2.6**
O25	1.02	1.04	1.03	0.3	0.6	0.6	1.05	1.02	1.03	0.5	0.2	0.4
O26	0.82	0.81	0.86	-0.3	-0.4	-0.2	0.77	0.75	0.67	-1.1	-1.3	-0.3
O27	1.00	1.04	1.02	0.1	0.5	0.4	1.03	1.05	1.04	0.3	0.4	0.5
O28	1.10	1.06	1.07	0.6	0.4	0.6	1.40	1.23	1.31	1.5	0.9	1.5
O29	1.08	1.04	1.05	1.7	0.8	1.7	1.11	1.03	1.07	1.7	0.4	1.4
O30	1.00	0.96	0.98	-0.1	-0.7	-0.7	1.02	0.94	0.98	0.3	-0.7	-0.4
O31	1.09	1.15	1.10	1.1	3.1	**3.1**	1.06	1.18	1.12	0.9	2.2	**2.4**
O32	0.88	0.83	0.85	-0.3	-0.6	-0.7	0.56	0.69	0.64	-1.0	-0.7	-1.3
O33	1.03	0.99	1.01	0.6	-0.1	0.2	1.04	1.05	1.04	0.5	0.5	0.7
O34	0.95	1.02	0.99	-0.4	0.2	-0.1	0.96	1.15	1.06	-0.2	0.9	0.5
O35	0.99	0.91	0.95	-0.2	-1.9	-1.7	0.95	0.87	0.91	-0.8	-1.6	-1.8
O36	0.85	0.84	0.85	-1.6	-1.6	**-2.2**	0.76	0.82	0.79	-1.9	-1.2	**-2.2**
O37	0.90	0.89	0.89	-2.2	-2.4	**-3.2**	0.87	0.88	0.88	-2.0	-1.6	**-2.5**
O38	0.93	0.97	0.96	-0.5	-0.3	-0.5	0.84	0.93	0.89	-0.9	-0.3	-0.8
O39	0.93	0.95	0.94	-0.6	-0.5	-0.8	0.87	0.79	0.83	-0.8	-1.1	-1.4
O40	0.94	0.87	0.90	-0.9	-2.2	**-2.2**	0.91	0.80	0.85	-0.9	-1.8	**-2.0**
r_{AB}	.86						.95					
$r_{GesamtA}$			.93						.98			
$r_{GesamtB}$			.96						.96			

Anmerkung. MNSQ Infit A: Meansquare-Infit-Statistiken in Zufallsstichprobe A; MNSQ Infit B: Meansquare-Infit-Statistiken in Zufallsstichprobe B; MNSQ Infit Gesamt: Meansquare-Infit-Statistiken in Gesamtstichprobe; ZSTD Infit A: z-standardisierte Meansquare-Infit-Statistiken in Zufallsstichprobe A; ZSTD Infit B: z-standardisierte Meansquare-Infit-Statistiken in Zufallsstichprobe B; ZSTD Infit Gesamt: z-standardisierte Meansquare-Infit-Statistiken in Gesamtstichprobe; MNSQ Outfit A: Meansquare-Outfit-Statistiken in Zufallsstichprobe A, MNSQ Outfit B: Meansquare-Outfit-Statistiken in Zufallsstichprobe B; MNSQ Outfit Gesamt: Meansquare-Outfit-Statistiken in Gesamtstichprobe; ZSTD Outfit A: z-standardisierte Meansquare-Outfit-Statistiken in Zufallsstichprobe A; ZSTD Outfit B: z-standardisierte Meansquare-Outfit-Statistiken in Zufallsstichprobe B; ZSTD Outfit Gesamt: z-standardisierte Meansquare-Outfit-Statistiken in Gesamtstichprobe; r_{AB}: Korrelation der Meansquare-Fit-Statistiken aus Zufallsstichprobe A und B; $r_{GesamtA}$: Korrelation der Meansquare-Fit-Statistiken aus Gesamtstichprobe mit Zufallsstichprobe A; $r_{GesamtB}$: Korrelation der Meansquare-Fit-Statistiken aus Gesamtstichprobe mit Zufallsstichprobe B

Die Korrelationen zwischen den Meansquare-Fit-Statistiken der einzelnen Stichproben fallen sowohl für die Infit- und Outfit-Maße sehr hoch aus und indizieren somit eine sehr gute Übereinstimmung. Die Itemauswahl anhand signifikanter z-standardisierter Meansquare-Werte anhand der Gesamtstichprobe korrespondiert ebenfalls sehr stark mit den entsprechenden Werten in den beiden Zufallsstichproben. Die Itemselektion anhand der Gesamtstichprobe kann somit über die Kreuzvalidierungsstichproben bestätigt werden.

Die Korrelationen der Itemparameter lagen für alle zu vergleichenden Stichproben bei $r = .99$; die Forderung der Invarianz der Itemparameter über die hier verglichenen Stichproben ist somit erfüllt.

Der Nachweis der Invarianz der Itemparameter in der hier gewählten Zufallsaufteilung zeigt jedoch lediglich, dass sich eine Personenheterogenität nicht in der hier gewählten Stichprobenaufteilung niederschlägt (Stelzl, 1979). Wie auch anhand der Gesamtstichprobe in Kapitel 14.1.2.2 geschehen, muss die dort vorgenommene Personenhomogenitäts-Analyse nach dem strengeren Prüfkriterium durch das mixed Rasch-Modell ebenfalls in beiden Zufallsstichproben durchgeführt werden. Diese Ergebnisse zeigt Tabelle 51.

Tabelle 51: Kreuzvalidierungsergebnisse der mixed Rasch-Analysen Subtest „Odd-Even-One-Out“

Stichprobe	**Klassenanzahl**	**BIC**	**CAIC**	**Log L**	**-2(log(cL_{RM}) -log(cL_{2KL}))**
A	1	4319.14	4340.14	-2103.58	96.90 **
	2	4339.57	4382.57	-2055.13	
B	1	4702.79	4723.79	-2294.43	123.38**
	2	4698.76	4741.76	-2232.74	
Gesamtstichprobe	1	8933.14	8954.14	-4402.80	23.94 n.s.
	2	9042.67	9085.67	-4390.83	

Anmerkung: Log L: Logarithmierte Likelihood der Daten unter den jwl. Modellannahmen; BIC: Bayes Information Criterion; CAIC: Consistent Akaikes Information Criterion; cL_{RM}: Bedingte Likelihood des dichotomen Rasch-Modells; cL_{2KL}: Bedingte Likelihood des mixed Rasch-Modells in zwei Klassen; -2(log(cL_{RM})-log(cL_{2KL})): Bedingter Likelihood-Quotiententest; **: $p < .01$

Der Likelihood-Quotiententest fällt in beide Zufallsstichproben zugunsten der Zweiklassen--Lösung signifikant aus. Die hierdurch bessere Modellanpassung an die Daten wird allerdings über die in der Zweiklassenlösung jeweils höherer Parameteranzahl „erkauft“ und demgemäß

sowohl vom BIC als auch CAIC als Überparametrisierung identifiziert. Beide Indizes kommen daher zu einer Bevorzugung der in ihren Annahmen sparsameren Einklassenlösung in beiden Zufallsstichproben und stimmen hierin auch mit dem Ergebnis aus der Gesamtstichprobe überein.

15.3 Kreuzvalidierungsanalysen zum Subtest „Zahlenreihen“

Eine Übersicht über die Modellgeltung der Items in den zu vergleichenden Stichproben liefert Tabelle 52.

Tabelle 52: Item-Fit-Kreuzvalidierungsergebnisse Subtest „Zahlenreihen“ (N_{Gesamt} = 434, n_A = 207, n_B = 227)

Item	MNSQ Infit A	MNSQ Infit B	MNSQ Infit Gesamt	ZSTD Infit A	ZSTD Infit B	ZSTD Infit Gesamt	MNSQ Outfit A	MNSQ Outfit B	MNSQ Outfit Gesamt	ZSTD Outfit A	ZSTD Outfit B	ZSTD Outfit Gesamt
ZR41	1.05	1.13	1.08	0.3	0.5	0.7	1.71	1.84	1.60	0.9	1.3	2.3
ZR42	1.23	1.12	1.17	0.9	0.5	3.2	1.30	1.23	1.28	2.0	1.9	3.1
ZR43	1.05	1.09	1.11	0.3	0.5	3.0	1.16	1.13	1.18	2.1	2.7	3.2
ZR44	1.05	1.08	1.12	0.3	0.7	2.5	1.18	1.12	1.16	0.6	0.3	2.0
ZR45	0.99	1.02	1.12	0.0	0.2	2.6	1.10	1.09	1.15	0.4	0.4	2.1
ZR46	1.10	1.14	1.14	1.2	1.6	3.6	1.34	1.45	1.23	2.1	2.3	4.0
ZR47	1.07	1.03	1.05	1.0	0.6	0.8	1.26	1.35	1.12	2.1	3.0	1.3
ZR48	0.97	0.94	1.08	-0.2	-0.7	1.5	1.10	1.08	1.20	0.5	0.4	2.4
ZR49	0.90	0.90	1.03	-1.3	-1.3	0.6	1.02	1.05	1.09	0.1	0.3	1.0
ZR50	1.01	1.01	0.94	0.1	0.2	-1.5	0.92	0.94	0.92	-0.7	-0.6	-1.3
ZR51	0.80	0.79	0.81	-2.9	-3.2	-3.0	0.73	0.74	0.69	-2.4	-2.4	-3.1
ZR52	0.90	0.99	1.08	-1.2	-0.1	1.5	1.10	1.09	1.17	0.9	0.7	1.9
ZR53	0.90	1.00	1.00	-0.5	0.1	0.0	0.89	0.96	0.94	-1.0	-0.2	-0.5
ZR54	0.95	0.95	0.90	-0.7	-0.8	-2.0	0.91	0.88	0.83	-0.8	-1.1	-2.2
ZR55	0.79	0.74	0.79	-1.8	-2.1	-5.3	0.81	0.90	0.73	-1.2	-0.6	-4.5
ZR56	0.69	0.76	0.79	-5.4	-3.0	-3.8	0.59	0.75	0.68	-4.1	-2.1	-3.6
ZR57	0.94	0.90	0.96	-0.3	-0.7	-0.1	0.71	0.70	0.60	-0.8	-0.9	-1.1
ZR58	0.79	0.81	0.87	-4.3	-4.8	-0.9	0.69	0.77	0.56	-2.7	-2.4	-2.3
ZR59	0.86	0.85	0.84	-0.8	-0.6	-1.4	0.67	0.70	0.65	-2.2	-2.0	-2.0
ZR60	0.97	1.00	0.89	-0.4	0.0	-0.7	0.77	0.66	0.53	-1.9	-2.4	-2.1
r_{AB}	.91						.97					
$r_{GesamtA}$			.82						.95			
$r_{GesamtB}$			.86						.92			

Anmerkung. MNSQ Infit A: Meansquare-Infit-Statistiken in Zufallsstichprobe A; MNSQ Infit B: Meansquare-Infit-Statistiken in Zufallsstichprobe B; MNSQ Infit Gesamt: Meansquare-Infit-Statistiken in Gesamtstichprobe; ZSTD Infit A: z-standardisierte Meansquare-Infit-Statistiken in Zufallsstichprobe A; ZSTD Infit B: z-standardisierte Meansquare-Infit-Statistiken in Zufallsstichprobe B; ZSTD Infit Gesamt: z-standardisierte Meansquare-Infit-Statistiken in Gesamtstichprobe; MNSQ Outfit A: Meansquare-Outfit-Statistiken in Zufallsstichprobe A, MNSQ Outfit B: Meansquare-Outfit-Statistiken in Zufallsstichprobe B; MNSQ Outfit Gesamt: Meansquare-Outfit-Statistiken in Gesamtstichprobe; ZSTD Outfit A: z-standardisierte Meansquare-Outfit-Statistiken in Zufallsstichprobe A; ZSTD Outfit B: z-standardisierte Meansquare-Outfit-Statistiken in Zufallsstichprobe B; ZSTD Outfit Gesamt: z-standardisierte Meansquare-Outfit-Statistiken in Gesamtstichprobe; r_{AB}: Korrelation der Meansquare-Fit-Statistiken aus Zufallsstichprobe A und B; $r_{GesamtA}$: Korrelation der Meansquare-Fit-Statistiken aus Gesamtstichprobe mit Zufallsstichprobe A; $r_{GesamtB}$: Korrelation der Meansquare-Fit-Statistiken aus Gesamtstichprobe mit Zufallsstichprobe B

Die Korrelationen zwischen den Meansquare-Fit-Werten der einzelnen Stichproben verweisen auf hohe bis sehr hohe Ergebnis-Übereinstimmungen. Die Korrelationen dieser Werte zwischen den Zufallsstichproben liegen numerisch nur etwas niedriger als jwls. mit der Gesamtstichprobe.
Die Überprüfung der Invarianz der Itemparameter über die drei Stichproben ergab eine sehr hohe Ergebnisübereinstimmung mit Korrelationen von jwls. $r = .99$.
Wie bereits für die Skala „verbale Analogien“ der Fall, ist jedoch auch hier die für eine Kreuzvalidierung kritische Frage, ob sich der anhand der Gesamtstichprobe identifizierte Speed-Effekt (vgl. Kap. 14.1.2.3) auch jeweils in den Zufallsstichproben nachweisen lässt und zu einer identischen Itemselektion durch eine Modellreformulierung nach Gleichung (47) mit einem zusätzlichen Parameter zur Kontrolle der Speed-Komponenten (vgl. Kapitel 14.1.2.3) führt. Dies ist Gegenstand der Kreuzvalidierungsanalysen der folgenden Personen-homogenitätsannahme.

In Bezug auf die Überprüfung der Personenhomogenität ergibt sich nach der folgenden Tabelle 53 auch für die Zufallstichproben nach allen Kriterien eine bessere Modellgeltung der Zwei- gegenüber der Einklassenlösung. Sowohl BIC als auch CAIC nehmen jeweils niedrigere Werte als die Einklassenlösung an und der Likelihood-Quotiententest weist die Personenheterogenität als substanziell aus.

Tabelle 53: Kreuzvalidierungsergebnisse der mixed Rasch-Analysen Subtest „Zahlenreihen“

Stichprobe	**Klassenanzahl**	**BIC**	**CAIC**	**Log L**	**$-2(\log(cL_{RM})-\log(cL_{2KL}))$**
A	1	3874.31	3895.31	-1881.16	152.36**
	2	3839.28	3882.28	-1804.98	
B	1	4178.63	4199.63	-2032.36	104.4**
	2	4193.98	4236.98	-1980.35	
Gesamtstichprobe	1	7963.19	7984.19	-3917.83	231.1**
	2	7865.7	7908.7	-3802.28	

Anmerkung: Log L: Logarithmierte Likelihood der Daten unter den jwl. Modellannahmen; BIC: Bayes Information Criterion; CAIC: Consistent Akaikes Information Criterion; cL_{RM}: Bedingte Likelihood des dichotomen Rasch-Modells; cL_{2KL}: Bedingte Likelihood des mixed Rasch-Modells in zwei Klassen; $-2(\log(cL_{RM})-\log(cL_{2KL}))$: Bedingter Likelihood-Quotiententest;
**: $p < .01$

Um der Frage nachzugehen, ob sich in den Zufallsstichproben die gleiche Form der Personenheterogenität wie in der Gesamtstichprobe als Aufteilung in eine bearbeitungslangsamere und eine bearbeitungsschnellere Klasse widerspiegelt, müssen die Itemparameter in den jeweiligen Klassen der Zufallsstichprobe verglichen werden, wie sie Tabelle 54 wiedergibt.

Tabelle 54: Itemparametervergleich mixed Rasch-Analysen Subtest „Zahlenreihen“

Item	**Itemparameter Klasse 1A**	**Itemparameter Klasse 2A**	**Itemparameter Klasse 1B**	**Itemparameter Klasse 2B**
ZR41	-1.99	-3.66	-2.12	-3.87
ZR42	-1.52	-3.94	-2.97	-3.15
ZR43	-1.92	-2.75	-1.64	-2.86
ZR44	-1.51	-2.52	-1.40	-2.68
ZR45	-1.16	-2.27	-1.38	-2.59
ZR46	1.78	1.13	2.14	0.93
ZR47	0.60	-0.53	0.40	-0.45
ZR48	-1.84	-1.77	-1.56	-2.20
ZR49	-0.17	-0.90	-0.39	-0.97
ZR50	-0.47	-1.45	-0.46	-1.60
ZR51	1.56	0.13	0.84	0.43
ZR52	-0.86	-1.41	-0.81	-1.22
ZR53	2.79	3.32	3.52	2.66
ZR54	-0.20	0.49	0.12	-0.12
ZR55	0.74	2.38	1.43	1.91
ZR56	-1.60	1.46	-0.72	0.88
ZR57	1.34	2.06	0.94	3.30
ZR58	-0.19	3.58	-0.24	3.74
ZR59	3.60	4.40	3.52	4.45
ZR60	1.01	2.25	0.76	3.40
Klassen-größe	53%	47%	57%	43%

In Übereinstimmung mit den Ergebnissen der Gesamtstichprobe (vgl. Kap. 14.1.2.3) ist Klasse 1 jeweils die bearbeitungsschnellere, erkennbar an den im Vergleich zur bearbeitungslangsameren Klasse 2 jeweils geringeren Itemschwierigkeiten ab ZR54 (bezüglich Klasse 1A) bzw. ZR55 (bezüglich Klasse 1B).

Diese Übereinstimmung der Befunde in Bezug auf die Gesamtanalyse verweist auf die Notwendigkeit einer Reformulierung des Testmodells zur Kontrolle der Speed-Komponente nach Gleichung (47) auch in beiden Zufallsstichproben mit einem Vergleich der hieraus resultierenden Itemselektionen. Die Ergebnisse dieser Analyse zeigt Tabelle 55.

Tabelle 55: Item-Fit-Kreuzvalidierungsergebnisse Subtest „Zahlenreihen“ nach Modellreformulierung (N_{Gesamt} = 434, n_A = 207, n_B = 227)

Item	MNSQ Infit A	MNSQ Infit B	MNSQ Infit Gesamt	ZSTD Infit A	ZSTD Infit B	ZSTD Infit Gesamt	MNSQ Outfit A	MNSQ Outfit B	MNSQ Outfit Gesamt	ZSTD Outfit A	ZSTD Outfit B	ZSTD Outfit Gesamt
ZR41	1.05	1.06	1.04	0.3	0.3	0.3	1.26	1.2	1.44	0.8	0.6	1.5
ZR42	1.1	1.02	1.04	0.5	0.2	0.2	1.40	1.32	1.45	1.0	0.7	1.5
ZR43	1.06	1.03	1.02	0.3	0.2	0.2	1.29	1.3	1.43	0.8	0.8	1.5
ZR44	1.04	1.03	1.03	0.3	0.3	0.3	1.05	1.04	1.16	0.2	0.2	0.6
ZR45	1.07	1.08	1.06	0.5	0.5	0.6	1.33	1.39	1.38	0.9	1.0	1.3
ZR46	0.97	1.03	1.02	-0.4	0.5	0.5	0.99	1.11	1.10	0.1	0.4	0.5
ZR47	1.24	1.21	1.26	**4.1**	**3.8**	**6.1**	1.29	1.28	1.33	0.9	0.8	1.2
ZR48	0.98	1.01	0.95	0.0	0.1	-0.4	0.93	1.09	0.92	-0.1	0.4	-0.2
ZR49	1.00	0.99	1.02	0.0	-0.1	0.4	1.03	0.98	1.05	0.2	0.1	0.3
ZR50	1.04	1.04	1.06	0.4	0.5	0.9	1.10	1.07	1.15	0.4	0.3	0.7
ZR51	1.24	1.15	1.24	**4.1**	**2.9**	**5.8**	1.32	1.17	1.31	0.9	0.6	1.2
ZR52	1.00	0.98	0.99	0.1	-0.2	-0.2	1.09	0.98	1.05	0.4	0	0.3
ZR53	0.97	0.99	0.96	-0.1	0.0	-0.2	0.91	0.93	0.98	-0.2	-0.2	0.0
ZR54	0.92	0.92	0.91	-1.6	-1.6	-1.8	0.91	0.89	0.89	-0.1	-0.3	-0.4
ZR55	0.88	0.93	0.91	*-2.3*	-1.0	-1.7	0.85	0.9	0.87	-0.4	-0.2	-0.5
ZR56	0.72	0.83	0.73	**-5.7**	**-3.7**	**-7.5**	0.69	0.81	0.71	-0.9	-0.5	-1.2
ZR57	0.98	0.94	0.98	-0.2	-0.9	-0.4	0.96	0.92	0.96	0	-0.1	-0.1
ZR58	0.78	0.75	0.72	**-4.8**	**-5.3**	**-8.5**	0.76	0.73	0.70	-0.7	-0.8	-1.3
ZR59	1.03	1.01	0.97	0.2	0.1	-0.1	1.15	1.03	0.91	0.5	0.2	-0.3
ZR60	0.91	0.92	0.93	-1.5	-1.5	-1.5	0.90	0.89	0.89	-0.2	-0.2	-0.4
r_{AB}		.94						.92				
$r_{GesamtA}$			.97						.93			
$r_{GesamtB}$			.96						.93			

Anmerkung. MNSQ Infit A: Meansquare-Infit-Statistiken in Zufallsstichprobe A; MNSQ Infit B: Meansquare-Infit-Statistiken in Zufallsstichprobe B; MNSQ Infit Gesamt: Meansquare-Infit-Statistiken in Gesamtstichprobe; ZSTD Infit A: z-standardisierte Meansquare-Infit-Statistiken in Zufallsstichprobe A; ZSTD Infit B: z-standardisierte Meansquare-Infit-Statistiken in Zufallsstichprobe B; ZSTD Infit Gesamt: z-standardisierte Meansquare-Infit-Statistiken in Gesamtstichprobe;
MNSQ Outfit A: Meansquare-Outfit-Statistiken in Zufallsstichprobe A, MNSQ Outfit B: Meansquare-Outfit-Statistiken in Zufallsstichprobe B; MNSQ Outfit Gesamt: Meansquare-Outfit-Statistiken in Gesamtstichprobe; ZSTD Outfit A: z-standardisierte Meansquare-Outfit-Statistiken in Zufallsstichprobe A; ZSTD Outfit B: z-standardisierte Meansquare-Outfit-Statistiken in Zufallsstichprobe B; ZSTD Outfit Gesamt: z-standardisierte Meansquare-Outfit-Statistiken in Gesamtstichprobe;
r_{AB}: Korrelation der Meansquare-Fit-Statistiken aus Zufallsstichprobe A und B; $r_{GesamtA}$: Korrelation der Meansquare-Fit-Statistiken aus Gesamtstichprobe mit Zufallsstichprobe A; $r_{GesamtB}$: Korrelation der Meansquare-Fit-Statistiken aus Gesamtstichprobe mit Zufallsstichprobe B

Die Korrelationen zwischen den Meansquare-Fit-Statistiken der jeweiligen Stichproben fallen jeweils sehr hoch aus. Auch in der Modellreformulierung ist somit eine hohe Übereinstimmung der Ergebnisse zwischen den Stichproben nachgewiesen. Legt man zudem die z-standardisierten Meansquare-Fit-Statistiken als Itemselektionskriterium zugrunde, wie dies in der Gesamtstichprobenanalyse geschah (vgl. Kap. 14.1.2.3), so gelangt man auch hierüber zu einer identischen Itemauswahl (s. die in der Tabelle fett gedruckten Werte). Lediglich Item ZR55 zeigt in der Zufallsstichprobe A abweichend von Zufallsstichprobe B und der Gesamtstichprobe einen signifikanten Infit (in der Tabelle kursiv gesetzt). Dieser fällt jedoch zum einen in seiner Größe der Modellabweichung mit $MNSQ_{Infit} = 0.88$ gering aus, zum anderen wird er weder in der Zufallsstichprobe B noch in der Gesamtstichprobe als signifikant von den Modellerwartungen abweichend ausgewiesen.

Die Korrelation der Itemparameter zwischen den Stichproben fiel mit $r = .99$ jeweils sehr hoch aus. Die Invarianz der Itemparameter über die Stichproben hinweg betrachtet ist somit gegeben.

15.4 Kreuzvalidierungsanalysen zum Subtest „Zahlenmatrizen“

Einen Überblick über die Ergebnisse zu Kreuzvalidierungen der Item-Modellpassung in den jeweiligen Stichproben liefert Tabelle 56.

Tabelle 56: Item-Fit-Kreuzvalidierungsergebnisse Subtest „Zahlenmatrizen“ ($N_{Gesamt} = 434$, $n_A = 207$, $n_B = 227$) (Fortsetzunng der Tabelle auf folgender Seite)

Item	MNSQ Infit A	MNSQ Infit B	MNSQ Infit Gesamt	ZSTD Infit A	ZSTD Infit B	ZSTD Infit Gesamt	MNSQ Outfit A	MNSQ Outfit B	MNSQ Outfit Gesamt	ZSTD Outfit A	ZSTD Outfit B	ZSTD Outfit Gesamt
ZM61	1.01	1.04	1.03	0.1	0.5	0.5	1.61	1.28	1.42	1.3	1.0	1.7
ZM62	0.99	0.93	0.93	0.0	-0.5	-0.5	2.44	2.94	2.52	**2.2**	**2.9**	**2.8**
ZM63	0.97	1.04	1.01	-0.2	0.3	0.2	1.50	1.34	1.42	1.3	0.3	1.2
ZM64	0.87	0.88	0.87	**-1.5**	**-1.5**	**-2.2**	0.83	0.82	0.82	-0.4	-0.7	-1.0
ZM65	0.86	0.78	0.82	**-2.0**	**-3.2**	**-3.8**	0.82	0.88	0.80	-1.1	-1.2	-1.5
ZM66	0.93	0.83	0.88	**-0.8**	**-1.9**	**-2.0**	0.75	0.70	0.76	-0.6	-1.1	-1.2
ZM67	1.21	1.23	1.23	0.5	0.6	0.6	*3.01*	*3.05*	*2.41*	*1.1*	*1.4*	*1.2*
ZM68	1.13	1.15	1.15	1.2	1.3	1.8	1.45	1.59	1.53	1.1	1.4	1.8
ZM69	0.99	1.01	1.00	-0.1	0.2	0.0	1.05	1.25	1.15	0.3	1.5	1.3
ZM70	1.01	0.88	0.94	0.1	-0.7	-0.5	1.03	1.20	1.12	0.2	0.6	0.5
ZM71	1.11	1.13	1.11	0.4	0.6	0.6	0.77	0.74	0.79	-0.3	-0.3	-0.2
ZM72	1.12	1.15	1.13	1.0	1.4	1.6	1.67	1.95	1.73	**1.9**	**2.8**	**3.1**
ZM73	0.99	0.91	0.94	0.0	-0.7	-0.6	1.17	1.33	1.21	0.9	1.1	1.0

Item	MNSQ Infit A	MNSQ Infit B	MNSQ Infit Gesamt	ZSTD Infit A	ZSTD Infit B	ZSTD Infit Gesamt	MNSQ Outfit A	MNSQ Outfit B	MNSQ Outfit Gesamt	ZSTD Outfit A	ZSTD Outfit B	ZSTD Outfit Gesamt
ZM74	1.01	1.17	1.09	0.1	1.0	0.8	1.59	1.79	1.67	1.2	1.4	1.7
ZM75	0.95	0.93	0.94	-0.5	-0.9	-1.1	1.32	1.61	1.52	**1.4**	**3.0**	**3.3**
ZM76	1.00	0.97	0.98	0.0	-0.3	-0.3	1.11	0.76	0.94	0.6	-1.1	-0.4
ZM77	0.79	0.69	0.73	-0.5	-1.0	-1.2	0.24	0.16	0.20	-0.8	-1.0	-1.5
ZM78	0.91	1.06	0.99	-0.1	0.3	0.0	*2.01*	*2.07*	*2.06*	*0.9*	*1.0*	*1.3*
ZM80	0.72	0.74	0.71	-0.5	-0.7	-0.8	0.01	0.54	0.46	0.6	0.0	-0.2
r_{AB}	.88						.95					
$r_{GesamtA}$			.93						.96			
$r_{GesamtB}$			.96						.98			

Anmerkung. MNSQ Infit A: Meansquare-Infit-Statistiken in Zufallsstichprobe A; MNSQ Infit B: Meansquare-Infit-Statistiken in Zufallsstichprobe B; MNSQ Infit Gesamt: Meansquare-Infit-Statistiken in Gesamtstichprobe; ZSTD Infit A: z-standardisierte Meansquare-Infit-Statistiken in Zufallsstichprobe A; ZSTD Infit B: z-standardisierte Meansquare-Infit-Statistiken in Zufallsstichprobe B; ZSTD Infit Gesamt: z-standardisierte Meansquare-Infit-Statistiken in Gesamtstichprobe;
MNSQ Outfit A: Meansquare-Outfit-Statistiken in Zufallsstichprobe A, MNSQ Outfit B: Meansquare-Outfit-Statistiken in Zufallsstichprobe B; MNSQ Outfit Gesamt: Meansquare-Outfit-Statistiken in Gesamtstichprobe; ZSTD Outfit A: z-standardisierte Meansquare-Outfit-Statistiken in Zufallsstichprobe A; ZSTD Outfit B: z-standardisierte Meansquare-Outfit-Statistiken in Zufallsstichprobe B; ZSTD Outfit Gesamt: z-standardisierte Meansquare-Outfit-Statistiken in Gesamtstichprobe;
r_{AB}: Korrelation der Meansquare-Fit-Statistiken aus Zufallsstichprobe A und B; $r_{GesamtA}$: Korrelation der Meansquare-Fit-Statistiken aus Gesamtstichprobe mit Zufallsstichprobe A; $r_{GesamtB}$: Korrelation der Meansquare-Fit-Statistiken aus Gesamtstichprobe mit Zufallsstichprobe B

Für die Korrelationen der Meansquare-Fit-Statistiken zwischen den Stichproben resultieren sehr hohe Werte. Auch im Hinblick auf die z-standardisierten Infit- und Outfit-Werte ergeben sich identische Itemselektionen, wobei im Falle der Items ZM64 und ZM66 die Analyse nach der Gesamtstichprobe zu einer strengeren Itemselektion führt. Wie schon in der Gesamtstichproben-Itemanalyse (s. Kap. 14.1.2.4) werden die Items ZM67 und ZM78 wegen relativ hoher Standardfehler ($SE_{ZM67\ in\ A} = 1.05$, $SE_{ZM78\ in\ A} = 0.51$; $SE_{ZM67\ in\ B} = 0.57$, $SE_{ZM78\ in\ B} = .38$) und der hierdurch verringerten Teststärke trotz sehr hoher Meansquare--Outfit-Werte auch in den Zufallsstichproben nicht als signifikant schlecht passend ausgewiesen (s. die entsprechenden kursiv gesetzten Werte in der Tabelle). Die Meansquare-Outfit--Statistiken als Effektstärkemaße der Modellabweichung hingegen führen hier zu identischen Itemselektionen, wie sie anhand der Itemanalyse in der Gesamtstichprobe durchgeführt wurden (s. Kap. 14.1.2.4).

Auch hinsichtlich der Konstanz der Itemparameter über die Stichproben hinweg betrachtet ergab sich eine sehr hohe Übereinstimmung der Ergebnisse, da deren Korrelationen mit $r_{AB} = .98$ sowie $r_{Gesamt\ A} = .99$ und $r_{Gesamt\ B} = .99$ sehr hoch ausfielen.

Das Ergebnis der Personenhomogenitätsüberprüfung nach Tabelle 57 zeigt sowohl nach dem BIC als auch nach dem CAIC eine Bevorzugung der Einklassenlösung. Zwar fällt der Likelihood-Quotiententest zum Vergleich der Einklassen- gegenüber der Zweiklassenlösung in beiden Zufallsstichproben signifikant aus, aber BIC und CAIC weisen die signifikant bessere Modellpassung im Sinne des Einfachheitskriteriums als überparametrisiert aus.

Tabelle 57: Kreuzvalidierungsergebnisse der mixed Rasch-Analysen Subtest „Zahlenmatrizen“

Stichprobe	**Klassenanzahl**	**BIC**	**CAIC**	**Log L**	**-2(log(cL_{RM}) -log(cL_{2KL}))**
A	1	2963.07	2984.07	-1425.54	63.74**
	2	3016.64	3059.64	-1393.67	
B	1	3401.83	3422.83	-1643.95	84.40**
	2	3436.76	3479.76	-1601.75	
Gesamtstichprobe	1	6283.10	6304.10	-3077.78	4.76 n.s.
	2	6271.94	6314.94	-3005.40	

Anmerkung: Log L: Logarithmierte Likelihood der Daten unter den jwl. Modellannahmen; BIC: Bayes Information Criterion; CAIC: Consistent Akaikes Information Criterion; cL_{RM}: Bedingte Likelihood des dichotomen Rasch-Modells; cL_{2KL}: Bedingte Likelihood des mixed Rasch-Modells in zwei Klassen; -2(log(cL_{RM})-log(cL_{2KL})): Bedingter Likelihood-Quotiententest;
**: $p < .01$; n.s.: nicht signifikant

15.5 Kreuzvalidierungsanalysen zum Subtest „Matrizen“

Die Kreuzvalidierungsergebnisse zur Modellgeltung auf Ebene der Items gibt Tabelle 58 wieder.

Tabelle 58: Item-Fit-Kreuzvalidierungsergebnisse Subtest „Matrizen“ ($N_{Gesamt} = 434$, $n_A = 207$, $n_B = 227$) (Fortsetzung der Tabelle auf folgender Seite)

Item	**MNSQ Infit A**	**MNSQ Infit B**	**MNSQ Infit Gesamt**	**ZSTD Infit A**	**ZSTD Infit B**	**ZSTD Infit Gesamt**	**MNSQ Outfit A**	**MNSQ Outfit B**	**MNSQ Outfit Gesamt**	**ZSTD Outfit A**	**ZSTD Outfit B**	**ZSTD Outfit Gesamt**
M81	1.03	0.97	1.01	0.2	-0.1	0.1	1.08	1.04	1.06	0.5	0.3	0.4
M82	1.22	1.15	1.18	2.9	2.0	3.5	1.56	1.76	1.65	4.3	5.4	6.8
M83	0.93	0.90	0.92	-0.4	-0.4	-0.6	0.89	0.92	0.85	-0.6	-0.2	-0.7
M84	1.10	1.05	1.09	1.7	0.6	1.6	1.29	1.11	1.20	2.0	0.9	2.1
M85	1.14	1.11	1.13	2.5	2.3	3.3	1.14	1.20	1.17	1.7	2.5	2.9
M86	0.98	1.09	1.01	-0.1	1.2	0.1	1.10	1.07	1.09	1.4	0.6	1.1
M87	0.97	1.03	0.98	-0.5	0.2	-0.1	0.89	0.91	0.90	-1.2	-0.3	-0.4
M88	0.95	0.98	0.96	-0.6	-0.2	-0.6	0.96	0.95	0.95	-0.2	-0.4	-0.5

Item	MNSQ Infit A	MNSQ Infit B	MNSQ Infit Gesamt	ZSTD Infit A	ZSTD Infit B	ZSTD Infit Gesamt	MNSQ Outfit A	MNSQ Outfit B	MNSQ Outfit Gesamt	ZSTD Outfit A	ZSTD Outfit B	ZSTD Outfit Gesamt
M89	0.94	0.98	0.96	-0.8	-0.4	-0.9	1.04	0.99	1.01	0.4	-0.1	0.2
M90	0.95	0.94	0.94	-0.5	-0.5	-0.7	0.95	0.86	0.91	-0.2	-0.8	-0.7
M91	0.99	1.01	1.00	-0.2	0.3	0.0	1.01	1.03	1.02	0.1	0.4	0.3
M92	1.10	1.14	1.12	1.6	2.3	2.7	1.15	1.21	1.18	1.6	2.1	2.6
M93	1.07	1.09	1.08	0.4	0.5	0.7	1.18	1.15	1.17	0.7	0.6	0.8
M94	0.99	1.04	0.98	-0.1	0.5	-0.3	0.81	0.88	0.97	-1.4	-1.2	-0.3
M95	0.99	0.94	0.96	-0.2	-1.3	-1.1	0.96	0.92	0.97	-0.3	-1.1	-0.6
M96	0.93	0.95	0.94	-0.7	-0.5	-0.9	0.82	0.86	0.83	-1.2	-0.7	-1.5
M97	0.89	0.81	0.85	-1.1	-1.9	-2.1	0.79	0.80	0.80	-1.3	-1.2	-1.8
M98	0.85	0.84	0.89	-1.6	-1.7	-1.7	0.79	0.69	0.85	-1.2	-2.1	-1.4
M99	0.95	0.94	0.90	-0.3	-0.3	-1.2	1.07	0.91	0.98	0.4	-0.4	-0.1
M100	0.83	0.89	0.86	-0.7	-0.3	-0.8	0.46	0.51	0.48	-1.7	-1.2	-2.2
r_{AB}	.86						.93					
$r_{GesamtA}$			.96						.95			
$r_{GesamtB}$			.92						.97			

Anmerkung. MNSQ Infit A: Meansquare-Infit-Statistiken in Zufallsstichprobe A; MNSQ Infit B: Meansquare-Infit-Statistiken in Zufallsstichprobe B; MNSQ Infit Gesamt: Meansquare-Infit-Statistiken in Gesamtstichprobe; ZSTD Infit A: z-standardisierte Meansquare-Infit-Statistiken in Zufallsstichprobe A; ZSTD Infit B: z-standardisierte Meansquare-Infit-Statistiken in Zufallsstichprobe B; ZSTD Infit Gesamt: z-standardisierte Meansquare-Infit-Statistiken in Gesamtstichprobe; MNSQ Outfit A: Meansquare-Outfit-Statistiken in Zufallsstichprobe A, MNSQ Outfit B: Meansquare-Outfit-Statistiken in Zufallsstichprobe B; MNSQ Outfit Gesamt: Meansquare-Outfit-Statistiken in Gesamtstichprobe; ZSTD Outfit A: z-standardisierte Meansquare-Outfit-Statistiken in Zufallsstichprobe A; ZSTD Outfit B: z-standardisierte Meansquare-Outfit-Statistiken in Zufallsstichprobe B; ZSTD Outfit Gesamt: z-standardisierte Meansquare-Outfit-Statistiken in Gesamtstichprobe; r_{AB}: Korrelation der Meansquare-Fit-Statistiken aus Zufallsstichprobe A und B; $r_{GesamtA}$: Korrelation der Meansquare-Fit-Statistiken aus Gesamtstichprobe mit Zufallsstichprobe A; $r_{GesamtB}$: Korrelation der Meansquare-Fit-Statistiken aus Gesamtstichprobe mit Zufallsstichprobe B.

Die Korrelationen der Meansquare-Fit-Statistiken zwischen den Stichproben fallen hoch bis sehr hoch aus. Die Korrelationen der Itemparameter zwischen den Stichproben fielen mit jwls. $r = .97$ sehr hoch aus. Die Itemparameterinvarianz in den hier verglichenen Stichproben ist daher gegeben.

Wie bei den Subtests „verbale Analogien“ und „Zahlenreihen“ stellt allerdings die Kreuzvalidierung des Speed-Effektes, der sich in der Gesamtstichprobenanalyse in einer Personenheterogenität niederschlug (s. Kap. 14.1.2.5) und der hieraus resultierenden Itemauswahl über eine Modellreformulierung, auch in diesem Subtest die entscheidende Kreuzvalidierungsanalyse dar. Tabelle 59 zeigt zunächst die Ergebnisse der Personenhomogenitätsanalysen auf Grundlage des mixed Rasch-Modells.

Tabelle 59: Kreuzvalidierungsergebnisse der mixed Rasch-Analysen Subtest „Matrizen“

Stichprobe	**Klassenanzahl**	**BIC**	**CAIC**	**Log L**	**-2(log(cL_{RM}) -log(cL_{2KL}))**
A	1	4219.08	4240.08	-2053.55	95.36**
	2	4239.95	4282.95	-2005.32	
B	1	4599.64	4620.64	-2242.86	125.44**
	2	4593.54	4636.54	-2180.14	
Gesamt-stichprobe	1	8733.71	8754.71	-4303.09	196.92**
	2	8670.4	8713.4	-4204.63	

Anmerkung: Log L: Logarithmierte Likelihood der Daten unter den jwl. Modellannahmen; BIC: Bayes Information Criterion; CAIC: Consistent Akaikes Information Criterion; cL_{RM}: Bedingte Likelihood des dichotomen Rasch-Modells; cL_{2KL}: Bedingte Likelihood des mixed Rasch-Modells in zwei Klassen; -2(log(cL_{RM})-log(cL_{2KL})): Bedingter Likelihood-Quotiententest;
**: $p < .01$

Übereinstimmend mit der Gesamtanalyse weisen BIC, CAIC und der Likelihood-Quotiententest die Zweiklassenlösung des mixed Rasch-Modells auch in den Zufallsstichproben als besser passend aus. Im Weiteren wird überprüft, ob sich die gefundene Personenheterogenität jeweils inhaltlich übereinstimmend mit der Gesamtstichprobenanalyse in einer bearbeitungsschnelleren und einer bearbeitungslangsameren Klasse niederschlägt. Tabelle 60 zeigt die für diese Analyse notwendigen Itemparameter in jeder der Klassen aus den Zufallsstichproben.

Tabelle 60: Itemparametervergleich mixed Rasch-Analysen Subtest „Matrizen“ (Fortsetzung der Tabelle auf folgender Seite)

Item	**Itemparameter Klasse 1A**	**Itemparameter Klasse 2A**	**Itemparameter Klasse 1B**	**Itemparameter Klasse 2B**
M81	-2.73	-1.40	-2.77	-1.07
M82	0.58	2.12	0.46	1.31
M83	-2.55	-1.82	-2.92	-2.21
M84	-1.46	0.60	-1.63	-0.42
M85	-0.60	0.41	-0.88	0.28
M86	0.69	0.71	0.47	1.64
M87	-2.95	-1.38	-3.08	-0.95
M88	-1.56	-0.36	-1.52	-0.92
M89	-1.24	-0.09	-0.82	-0.28
M90	-1.65	-1.97	-1.86	-1.85
M91	-0.27	1.06	-0.12	0.49
M92	0.49	0.44	0.16	1.12
M93	2.08	2.80	2.19	1.98
M94	0.94	1.48	0.80	1.02

Item	Itemparameter Klasse 1A	Itemparameter Klasse 2A	Itemparameter Klasse 1B	Itemparameter Klasse 2B
M95	-0.01	0.64	-0.02	-0.28
M96	1.11	1.14	1.56	0.65
M97	1.54	-0.62	1.86	-0.36
M98	1.45	-2.32	1.85	-0.96
M99	2.35	-1.70	2.05	-0.63
M100	3.81	0.26	4.20	1.44
Klassen-größe	91%	9%	85%	15%

Wie in den Analysen zur Gesamtstichprobe (s. Kap. 14.1.2.5), zeichnet sich die jeweils zweite Klasse gegenüber der ersten durch eine höhere Bearbeitungsgeschwindigkeit aus, was sich für diese Klasse jeweils in geringeren Itemschwierigkeiten ab Item M97 zeigt. Nicht repliziert werden kann hierbei der interessante Befund aus den Analysen der Gesamtstichprobe, wonach Item M83 und M90 von einem Wahrnehmungs-, Figural- oder auch Gestaltfaktor (Gallini, 1983; Mackintosh & Bennett, 2005; van der Ven & Ellis, 2000; Vigneau & Bors, 2005) beeinflusst werden, welcher gerade in der bearbeitungs- bzw. vermutlich sogar wahrnehmungsschnelleren zweiten Klasse zu deutlich geringeren Itemschwierigkeiten führte. Da sich dieser Effekt jedoch in der Gesamtstichprobenanalyse nicht auf die Itemauswahl auswirkte, ist seine Nicht-Replizierbarkeit vielmehr von theoretischem Interesse.
Analog zu den Subtests „verbale Analogien“ und „Zahlenreihen“ erfolgt wegen der Replikation des Speed-Effektes der Vergleich der Itemauswahl in den jeweiligen Stichproben über die Modellreformulierung nach Gleichung (47). Vorab wurde Item M82 wie schon in der Gesamtstichprobenanalyse wegen seiner fehlenden Trennschärfe durch die Uneindeutigkeit eines Distraktors aus der Analyse ausgeschlossen (vgl. Kap. 14.1.2.5). Die Ergebnisse der Kreuzvalidierungsanalysen nach Modellgleichung (47) zeigt Tabelle 61.

Tabelle 61: Item-Fit-Kreuzvalidierungsergebnisse Subtest „Matrizen“ nach Modellreformulierung ($N_{Gesamt} = 434$, $n_A = 207$, $n_B = 227$)

Item	MNSQ Infit A	MNSQ Infit B	MNSQ Infit Gesamt	ZSTD Infit A	ZSTD Infit B	ZSTD Infit Gesamt	MNSQ Outfit A	MNSQ Outfit B	MNSQ Outfit Gesamt	ZSTD Outfit A	ZSTD Outfit B	ZSTD Outfit Gesamt
M81	1.04	1.00	1.03	0.3	0.0	0.3	1.13	1.10	1.18	0.5	0.3	0.7
M83	1.02	1.01	0.99	0.2	0.1	0.0	0.89	0.93	0.95	-0.2	-0.2	-0.1
M84	1.08	1.06	1.09	1.0	0.7	1.7	1.12	1.12	1.14	0.5	0.5	0.6
M85	1.06	1.08	1.09	**1.5**	**2.0**	**3.1**	1.07	1.10	1.11	0.3	0.4	0.5
M86	0.94	0.93	0.97	-1.2	-0.8	-0.6	0.89	1.04	0.96	-0.3	0.2	-0.1
M87	1.04	1.01	1.00	0.3	0.1	0.0	0.94	0.91	1.00	-0.1	-0.2	0.1
M88	1.01	1.02	1.02	0.1	0.2	0.3	1.02	1.05	1.04	0.2	0.3	0.3
M89	0.99	1.00	1.01	-0.1	-0.1	0.3	0.98	1.00	1.02	0.0	0.1	0.2

Item	MNSQ Infit A	MNSQ Infit B	MNSQ Infit Gesamt	ZSTD Infit A	ZSTD Infit B	ZSTD Infit Gesamt	MNSQ Outfit A	MNSQ Outfit B	MNSQ Outfit Gesamt	ZSTD Outfit A	ZSTD Outfit B	ZSTD Outfit Gesamt
M90	1.00	0.99	1.01	0.0	0.0	0.2	0.99	0.98	1.05	0.1	0	0.3
M91	0.99	0.97	0.96	-0.2	-0.1	-1.8	0.99	1.01	0.96	0.1	0.1	-0.1
M92	1.04	1.04	1.06	0.7	0.8	1.9	1.04	1.05	1.07	0.2	0.2	0.4
M93	0.97	1.00	1.00	-0.1	0.1	0.0	1.06	1.07	1.05	0.3	0.3	0.3
M94	0.96	0.97	0.98	-0.6	-0.3	-0.5	0.93	0.96	0.96	-0.1	0.0	-0.1
M95	0.97	0.98	0.97	-0.7	-0.5	-1.2	0.96	0.98	0.97	0.0	0.0	0.0
M96	0.95	0.93	0.94	-0.6	-0.7	-0.9	0.91	0.89	0.90	-0.2	-0.3	-0.4
M97	0.93	0.90	0.90	-0.7	-1.1	-1.7	0.87	0.89	0.83	-0.4	-0.3	-0.6
M98	0.97	0.90	0.89	-0.4	-0.8	-1.9	0.97	0.92	0.84	0.0	-0.2	-0.6
M99	0.93	0.94	0.92	-0.3	-0.6	-1.0	0.88	0.95	0.84	-0.3	-0.1	-0.6
M100	0.95	0.94	0.97	-0.5	-0.5	-0.2	0.94	0.95	0.94	-0.1	-0.1	-0.2
r_{AB}	.89						.85					
$r_{GesamtA}$			.86						.88			
$r_{GesamtB}$			.95						.85			

Anmerkung. MNSQ Infit A: Meansquare-Infit-Statistiken in Zufallsstichprobe A; MNSQ Infit B: Meansquare-Infit-Statistiken in Zufallsstichprobe B; MNSQ Infit Gesamt: Meansquare-Infit-Statistiken in Gesamtstichprobe; ZSTD Infit A: z-standardisierte Meansquare-Infit-Statistiken in Zufallsstichprobe A; ZSTD Infit B: z-standardisierte Meansquare-Infit-Statistiken in Zufallsstichprobe B; ZSTD Infit Gesamt: z-standardisierte Meansquare-Infit-Statistiken in Gesamtstichprobe;
MNSQ Outfit A: Meansquare-Outfit-Statistiken in Zufallsstichprobe A, MNSQ Outfit B: Meansquare-Outfit-Statistiken in Zufallsstichprobe B; MNSQ Outfit Gesamt: Meansquare-Outfit-Statistiken in Gesamtstichprobe; ZSTD Outfit A: z-standardisierte Meansquare-Outfit-Statistiken in Zufallsstichprobe A; ZSTD Outfit B: z-standardisierte Meansquare-Outfit-Statistiken in Zufallsstichprobe B; ZSTD Outfit Gesamt: z-standardisierte Meansquare-Outfit-Statistiken in Gesamtstichprobe;
r_{AB}: Korrelation der Meansquare-Fit-Statistiken aus Zufallsstichprobe A und B; $r_{GesamtA}$: Korrelation der Meansquare-Fit-Statistiken aus Gesamtstichprobe mit Zufallsstichprobe A; $r_{GesamtB}$: Korrelation der Meansquare-Fit-Statistiken aus Gesamtstichprobe mit Zufallsstichprobe B

Die Ergebnisse, sowohl verglichen zwischen den Zufallsstichproben als auch jeweils mit der Gesamtstichprobe sind hoch bis sehr hoch korreliert und gelangen unter Zugrundelegung des Toleranzbereichs von 0.8 – 1.2 für die Meansquare-Fit-Statistiken zu identischen Itemselektionen. Einzig Item M85 identifizieren die Zufallsstichprobe B und die Gesamtstichprobe gegenüber der Zufallsstichprobe A als signifikant schlecht passend (in der Tabelle fett markiert), wobei die Größe der Abweichung nach der MNSQ-Infit-Statistik jeweils nicht groß ausfällt. Die vorgenommene Eliminierung dieses Items anhand der Gesamtstichprobe stellt somit ein strengeres Vorgehen der Itemselektion dar.

Die Korrelationen der Itemparameter zur Überprüfung von deren Invarianz zwischen den Stichproben fielen mit jeweils $r = .99$ sehr hoch aus.

15.6 Kreuzvalidierungsanalysen zum Subtest „SPARK“

Eine Übersicht über die Kreuzvalidierungsanalysen der Itemmodellpassung zeigt Tabelle 62.

Tabelle 62: Item-Fit-Kreuzvalidierungsergebnisse Subtest SPARK (N_{Gesamt} = 434, n_A = 207, n_B = 227)

Item	MNSQ Infit A	MNSQ Infit B	MNSQ Infit Gesamt	ZSTD Infit A	ZSTD Infit B	ZSTD Infit Gesamt	MNSQ Outfit A	MNSQ Outfit B	MNSQ Outfit Gesamt	ZSTD Outfit A	ZSTD Outfit B	ZSTD Outfit Gesamt
AK1	1.23	1.21	1.20	**2.5**	**2.2**	**3.1**	1.89	1.68	1.56	**7.3**	**5.3**	**7.1**
AK2	1.06	1.06	1.03	0.7	0.7	0.5	0.90	1.03	0.96	-0.6	0.3	-0.5
AK3	0.85	0.86	0.95	-1.6	-1.4	-0.7	0.63	0.71	0.86	**-4.4**	**-3.0**	**-2.1**
AK4	0.74	0.68	0.78	**-2.8**	**-3.1**	**-3.5**	0.51	0.51	0.60	**-6.1**	**-5.6**	**-7.0**
AK5	1.16	1.02	1.06	2.0	0.3	1.1	1.08	0.97	1.05	0.8	-0.3	0.7
r_{AB}	.95						.98					
$r_{GesamtA}$			.95						.98			
$r_{GesamtB}$			.98						.98			

Anmerkung. MNSQ Infit A: Meansquare-Infit-Statistiken in Zufallsstichprobe A; MNSQ Infit B: Meansquare-Infit-Statistiken in Zufallsstichprobe B; MNSQ Infit Gesamt: Meansquare-Infit-Statistiken in Gesamtstichprobe; ZSTD Infit A: z-standardisierte Meansquare-Infit-Statistiken in Zufallsstichprobe A; ZSTD Infit B: z-standardisierte Meansquare-Infit-Statistiken in Zufallsstichprobe B; ZSTD Infit Gesamt: z-standardisierte Meansquare-Infit-Statistiken in Gesamtstichprobe;
MNSQ Outfit A: Meansquare-Outfit-Statistiken in Zufallsstichprobe A, MNSQ Outfit B: Meansquare-Outfit-Statistiken in Zufallsstichprobe B; MNSQ Outfit Gesamt: Meansquare-Outfit-Statistiken in Gesamtstichprobe; ZSTD Outfit A: z-standardisierte Meansquare-Outfit-Statistiken in Zufallsstichprobe A; ZSTD Outfit B: z-standardisierte Meansquare-Outfit-Statistiken in Zufallsstichprobe B; ZSTD Outfit Gesamt: z-standardisierte Meansquare-Outfit-Statistiken in Gesamtstichprobe;
r_{AB}: Korrelation der Meansquare-Fit-Statistiken aus Zufallsstichprobe A und B; $r_{GesamtA}$: Korrelation der Meansquare-Fit-Statistiken aus Gesamtstichprobe mit Zufallsstichprobe A; $r_{GesamtB}$: Korrelation der Meansquare-Fit-Statistiken aus Gesamtstichprobe mit Zufallsstichprobe B

Die Stichproben zeigen eine sehr große Übereinstimmung der Itemmodellgeltung, ausgewiesen durch die durchgängig sehr hohen Korrelationen der Meansquare-Fit-Statistiken zwischen den Stichproben. Ebenso einheitlich fällt die Item-Modellgeltung anhand der z-standardisierten Werte aus (in der Tabelle fett markiert). Wie in der Gesamtanalyse (s. Kap. 14.1.2.6) fällt Item AK4 auch in den beiden Zufallsstichproben durch seine deutlich zu gute Modellpassung auf (vgl. insbesondere die Meansquare-Fit-Statistiken). In der Gesamtanalyse (s. Kap. 14.1.2.6) konnte ließ sich dies auf eine logische Abhängigkeit des Iteminhalts zurückführen, weil Item AK3 Informationen zur Lösung von AK4 enthielt. Die Itemantwort auf AK4 erwies sich daher als lokal nicht stochastisch unabhängig von der bereits auf AK3 gegebenen Antwort.

Die Konstanz der Itemparameter war mit Korrelationen von r_{AB} = .98 und jwls. r = .99 zwischen den Zufallsstichproben und der Gesamtstichprobe gegeben.

Die Personenhomogenitätsanalyse nach dem mixed Rasch-Modell fasst Tabelle 63 zusammen.

Tabelle 63: Kreuzvalidierungsergebnisse der mixed Rasch-Analysen Subtest SPARK

Stichprobe	**Klassenanzahl**	**BIC**	**CAIC**	**Log L**	**-2(log(cL_{RM}) -log(cL_{2KL}))**
A	1	1227.03	1233.03	-597.52	89.28**
	2	1175.08	1188.08	-552.88	
B	1	1342.34	1348.34	-654.89	83.36**
	2	1296.94	1309.94	-613.21	
Gesamtstichprobe	1	2543.67	2549.67	-1253.62	170.30**
	2	2415.89	2428.89	-1168.47	

Anmerkung: Log L: Logarithmierte Likelihood der Daten unter den jwl. Modellannahmen; BIC: Bayes Information Criterion; CAIC: Consistent Akaikes Information Criterion; cL_{RM}: Bedingte Likelihood des dichotomen Rasch-Modells; cL_{2KL}: Bedingte Likelihood des mixed Rasch-Modells in zwei Klassen; -2(log(cL_{RM})-log(cL_{2KL})): Bedingter Likelihood-Quotiententest;
**: $p < .01$.

In Übereinstimmung mit der Gesamtstichprobenanalyse ergibt sich sowohl nach den informationstheoretischen Maßen als auch nach dem Likelihood-Quotiententest in den Zufallsstichproben eine bessere Passung der Zwei- gegenüber der Einklassenlösung des mixed Rasch-Modells.

In der Gesamtstichprobenanalyse kamen die Unterschiede zwischen den beiden Klassen insbesondere durch die logische Abhängigkeit zwischen Item AK3 und AK4 zustande. Die leistungsstärkere Klasse 1 konnte einen deutlichen Nutzen hieraus ziehen: AK4 zeigte eine niedrigere Schwierigkeit als AK3, weil dessen Informationen für die Lösung von AK4 genutzt werden konnten. Bei der weniger leistungsfähigen Klasse 2 hingegen wirkte sich die Abhängigkeit genau in umgekehrter Richtung aus: AK4 war schwerer, da sich eine *falsche* Antwort auf AK3 wegen der logischen Abhängigkeit auch nachteilig auf die Lösung von AK4 auswirken *musste*.

Eine Kreuzvalidierung dieses Effektes sollte sich daher neben einem gleichsinnigen Profil der Itemparameter insbesondere in einer *geringeren* Schwierigkeit von AK4 gegenüber AK3 in Klasse 1 manifestieren, sich in Klasse 2 hingegen genau umgekehrt darstellen.

Tabelle 64 gibt einen Überblick über die Itemparameter in den jeweiligen Klassen der Zufallsstichproben.

Tabelle 64: Itemparametervergleich mixed Rasch-Analysen Subtest SPARK

Item	Item-parameter Klasse 1A	Item-parameter Klasse 2A	Item-parameter Klasse 1B	Item-parameter Klasse 2B	Itemparameter Klasse 1 Gesamt	Itemparameter Klasse 2 Gesamt
AK1	1.84	-0.99	1.99	-0.90	1.94	-1.06
AK2	0.48	-0.70	0.48	-0.88	0.49	-0.91
AK3	-1.35	0.01	-1.01	-0.29	-1.16	-0.25
AK4	-1.64	1.51	-2.18	1.09	-1.98	1.68
AK5	0.68	0.17	0.71	0.98	0.71	0.54
Klassen-größe	73%	27%	73%	27%	73%	27%

Das Profil der Itemparameter, insbesondere dasjenige zwischen AK3 und AK4, stimmt in allen Personenklassen der jeweiligen Stichproben überein. Die in der Gesamtstichprobe gefundene Personenheterogenität aufgrund logisch voneinander abhängiger Items lässt sich daher auch in den Zufallsstichproben nachweisen und als stabil bezeichnen.

15.7 Kreuzvalidierungsanalysen zum Subtest „Ideenflüssigkeit“

Die Ergebnisse des Kreuzvalidierungsvergleichs zur Modellgeltung auf Ebene der Items zeigt Tabelle 65.

Tabelle 65: Item-Fit-Kreuzvalidierungsergebnisse Subtest „Ideenflüssigkeit“ ($N_{Gesamt} = 434$, $n_A = 207$, $n_B = 227$)

Item	MNSQ Infit A	MNSQ Infit B	MNSQ Infit Gesamt	ZSTD Infit A	ZSTD Infit B	ZSTD Infit Gesamt	MNSQ Outfit A	MNSQ Outfit B	MNSQ Outfit Gesamt	ZSTD Outfit A	ZSTD Outfit B	ZSTD Outfit Gesamt
Zeichnungen Menge	1.04	1.11	1.09	0.5	1.3	1.6	1.07	1.18	1.15	0.9	2.4	2.8
Sätze Menge	1.14	1.19	1.17	2.0	2.8	3.4	1.14	1.18	1.16	1.9	2.7	3.2
Verwendungen Menge	0.75	0.71	0.74	**-3.6**	**-4.5**	**-5.6**	0.78	0.70	0.74	**-3.1**	**-4.7**	**-5.6**
r_{AB}	.99											
$r_{GesamtA}$			.99				.98		.98			
$r_{GesamtB}$			.99						.99			

Anmerkung. MNSQ Infit A: Meansquare-Infit-Statistiken in Zufallsstichprobe A; MNSQ Infit B: Meansquare-Infit-Statistiken in Zufallsstichprobe B; MNSQ Infit Gesamt: Meansquare-Infit-Statistiken in Gesamtstichprobe; ZSTD Infit A: z-standardisierte Meansquare-Infit-Statistiken in Zufallsstichprobe A; ZSTD Infit B: z-standardisierte Meansquare-Infit-Statistiken in Zufallsstichprobe B; ZSTD Infit Gesamt: z-standardisierte Meansquare-Infit-Statistiken in Gesamtstichprobe; MNSQ Outfit A: Meansquare-Outfit-Statistiken in Zufallsstichprobe A, MNSQ Outfit B: Meansquare-Outfit-Statistiken in Zufallsstichprobe B; MNSQ Outfit Gesamt: Meansquare-Outfit-Statistiken in Gesamtstichprobe; ZSTD Outfit A: z-standardisierte Meansquare-Outfit-Statistiken in Zufallsstichprobe A; ZSTD Outfit B: z-standardisierte Meansquare-Outfit-Statistiken in Zufallsstichprobe B; ZSTD Outfit Gesamt: z-standardisierte Meansquare-Outfit-Statistiken in Gesamtstichprobe; r_{AB}: Korrelation der Meansquare-Fit-Statistiken aus Zufallsstichprobe A und B; $r_{GesamtA}$: Korrelation der Meansquare-Fit-Statistiken aus Gesamtstichprobe mit Zufallsstichprobe A; $r_{GesamtB}$: Korrelation der Meansquare-Fit-Statistiken aus Gesamtstichprobe mit Zufallsstichprobe B

Die Korrelationen der Meansquare-Fit-Werte fallen jeweils sehr hoch aus und verweisen somit auf eine sehr gute Übereinstimmung der Kreuzvalidierungsergebnisse. Mit der Gesamtstichprobenanalyse in Einklang fällt die Aufgabe zur „Menge der Verwendungsmöglichkeiten für Schaumstoffpolster“ auch nach den Analysen in den Zufallsstichproben als deutlich zu trennscharf auf und liegt nach den Meansquare-Fit-Statistiken außerhalb des Bereichs einer guten Modellpassung von 0.8 – 1.2. Die Meansquare-Fit-Statistiken der übrigen Aufgaben liegen demgegenüber auch in den Zufallsstichproben jeweils innerhalb dieses Intervalls.
Die Itemparameter zeigten sich mit Korrelationen zwischen den jeweiligen Stichproben von jwls. $r = .99$ als invariant.

15.8 Kreuzvalidierungsanalysen zum Subtest „Ideenflexibilität“

Die Betrachtung der Item-Modellgeltung dieses Subtests nach Tabelle 66 ergibt insgesamt eine sehr gute Übereinstimmung zwischen den Stichproben.

Tabelle 66: Item-Fit-Kreuzvalidierungsergebnisse Subtest „Ideenflexibilität“ ($N_{Gesamt} = 434$, $n_A = 207$, $n_B = 227$)

Item	MNSQ Infit A	MNSQ Infit B	MNSQ Infit Gesamt	ZSTD Infit A	ZSTD Infit B	ZSTD Infit Gesamt	MNSQ Outfit A	MNSQ Outfit B	MNSQ Outfit Gesamt	ZSTD Outfit A	ZSTD Outfit B	ZSTD Outfit Gesamt
K.-M. Zeichnungen	1.04	1.03	1.05	0.5	0.3	0.9	1.07	1.04	1.06	0.9	0.5	1.2
K.-M. Verwendungen	0.89	0.95	0.93	-1.5	-0.6	-1.3	0.93	0.95	0.94	-0.9	-0.7	-1.0

Anmerkung. K.-M. Zeichnungen: Kategorienmenge Zeichnungen zu Teilvorlage; K.-M.Verwendungen: Kategorienmenge Verwendungen von Schaumstoffpolster;
MNSQ Infit A: Meansquare-Infit-Statistiken in Zufallsstichprobe A; MNSQ Infit B: Meansquare-Infit-Statistiken in Zufallsstichprobe B; MNSQ Infit Gesamt: Meansquare-Infit-Statistiken in Gesamtstichprobe; ZSTD Infit A: z-standardisierte Meansquare-Infit-Statistiken in Zufallsstichprobe A; ZSTD Infit B: z-standardisierte Meansquare-Infit-Statistiken in Zufallsstichprobe B; ZSTD Infit Gesamt: z-standardisierte Meansquare-Infit-Statistiken in Gesamtstichprobe;
MNSQ Outfit A: Meansquare-Outfit-Statistiken in Zufallsstichprobe A, MNSQ Outfit B: Meansquare-Outfit-Statistiken in Zufallsstichprobe B; MNSQ Outfit Gesamt: Meansquare-Outfit-Statistiken in Gesamtstichprobe; ZSTD Outfit A: z-standardisierte Meansquare-Outfit-Statistiken in Zufallsstichprobe A; ZSTD Outfit B: z-standardisierte Meansquare-Outfit-Statistiken in Zufallsstichprobe B; ZSTD Outfit Gesamt: z-standardisierte Meansquare-Outfit-Statistiken in Gesamtstichprobe;

Alle Meansquare-Fit-Werte der jeweiligen Stichproben liegen nahe an ihren Erwartungswerten. Auch die z-standardisierten Werte zeigen jeweils keine signifikante Modellabweichung an.

Die hohe Konvergenz der Ergebnisse ergibt sich darüber hinaus auch durch die sehr nahe beieinanderliegenden Itemparameterschätzungen wie in folgender Tabelle 67 wiedergegeben. (Die Berechnung einer Korrelation war hier wegen lediglich zwei Itemparametern nicht nötig). Auch die Invarianz der Itemparameter ist somit gegeben.

Tabelle 67: Itemparametervergleich Subtest „Ideenflexibilität“

Item	**Itemparameter Stichprobe A**	**Itemparameter Stichprobe B**	**Itemparameter Gesamt-stichprobe**
K.-M. Zeichnungen	-0.14	-0.10	-0.12
K.-M. Verwendungen	0.11	0.15	0.12

Anmerkung. K.-M. Zeichnungen: Kategorienmenge Zeichnungen zu Teilvorlage; K.-M.Verwendungen: Kategorienmenge Verwendungen von Schaumstoffpolster.

15.9 Kreuzvalidierungsanalysen zum Subtest „empiriebezogenes Denken“

Die Ergebnisse der Kreuzvalidierungsanalysen zur Item-Modellgülltigkeit zeigt Tabelle 68.

Tabelle 68: Item-Fit-Kreuzvalidierungsergebnisse Subtest „empiriebezogenes Denken“ (N_{Gesamt} = 434, n_A = 207, n_B = 227)

Item	**MNSQ Infit A**	**MNSQ Infit B**	**MNSQ Infit Gesamt**	**ZSTD Infit A**	**ZSTD Infit B**	**ZSTD Infit Gesamt**	**MNSQ Outfit A**	**MNSQ Outfit B**	**MNSQ Outfit Gesamt**	**ZSTD Outfit A**	**ZSTD Outfit B**	**ZSTD Outfit Gesamt**
Prinzip Gruppenvgl.	0.99	1.04	0.97	0.0	0.9	-0.5	1.01	1.02	0.91	1.0	1.1	-1.8
Störvariablen-Kontrolle	0.88	0.83	0.92	-1.6	-1.7	-1.5	0.85	0.87	0.92	-1.8	-1.7	-1.3
Verallgemeinerbarkeit	1.09	1.10	1.12	1.1	1.0	2.2	1.14	1.15	1.19	1.1	1.1	3.3
Externe Validität	0.98	0.86	0.91	-0.2	-1.4	-1.6	0.93	0.89	0.9	-0.3	-1.2	-1.4
r_{AB}	.86						.98					
$r_{GesamtA}$			.85						.83			
$r_{GesamtB}$			.87						.85			

Anmerkung. MNSQ Infit A: Meansquare-Infit-Statistiken in Zufallsstichprobe A; MNSQ Infit B: Meansquare-Infit-Statistiken in Zufallsstichprobe B; MNSQ Infit Gesamt: Meansquare-Infit-Statistiken in Gesamtstichprobe; ZSTD Infit A: z-standardisierte Meansquare-Infit-Statistiken in Zufallsstichprobe A; ZSTD Infit B: z-standardisierte Meansquare-Infit-Statistiken in Zufallsstichprobe B; ZSTD Infit Gesamt: z-standardisierte Meansquare-Infit-Statistiken in Gesamtstichprobe;
MNSQ Outfit A: Meansquare-Outfit-Statistiken in Zufallsstichprobe A, MNSQ Outfit B: Meansquare-Outfit-Statistiken in Zufallsstichprobe B; MNSQ Outfit Gesamt: Meansquare-Outfit-Statistiken in Gesamtstichprobe; ZSTD Outfit A: z-standardisierte Meansquare-Outfit-Statistiken in Zufallsstichprobe A; ZSTD Outfit B: z-standardisierte Meansquare-Outfit-Statistiken in Zufallsstichprobe B; ZSTD Outfit Gesamt: z-standardisierte Meansquare-Outfit-Statistiken in Gesamtstichprobe;
r_{AB}: Korrelation der Meansquare-Fit-Statistiken aus Zufallsstichprobe A und B; $r_{GesamtA}$: Korrelation der Meansquare-Fit-Statistiken aus Gesamtstichprobe mit Zufallsstichprobe A; $r_{GesamtB}$: Korrelation der Meansquare-Fit-Statistiken aus Gesamtstichprobe mit Zufallsstichprobe B

Die Ergebnisse sind hoch übereinstimmend Zum einen fallen die Korrelationen der Meansquare-Itemfit-Statistiken hoch bis sehr hoch aus und liegen jeweils im Toleranzintervall guter Modellpassung von 0.8 – 1.2.
Zum anderen ergaben sich auch für die Itemparameter sehr hohe Korrelationen von jwls. r = .99, was auf deren Invarianz über die hier betrachteten Stichproben hinweist.

15.10 Zusammenfassung der Kreuzvalidierungsanalysen zur Modellgültigkeit

Die jeweiligen Ergebnisse der Kreuzvalidierungen zur Itemauswahl, Personenhomogenität und Itemparameterinvarianz stimmten insgesamt hoch überein. Auf der Seite der Itemauswahl zeigten sich die Meansquare-Fit-Statistiken zwischen den Stichproben hoch bis sehr korreliert und gelangten zu weitestgehend identischen Itemselektionen. Dies betraf die Skalenanalysen nach dem dichotomen Rasch-Modell, diejenigen nach der Reformulierung mit einer Speed-Komponente wie auch das Mutifacetten-Rasch-Modell. Im Falle des Subtests „verbale Analogien“ und des Matrizentests führte zudem die Analyse nach der Gesamtstichprobe zu einer strengeren Itemauswahl als die Analysen anhand der beiden Zufallsstichproben. Auch die Personenheterogenität in Form einer bearbeitungsschnelleren und einer bearbeitungslangsameren Klasse fand sich in den Analysen der Zufallsstichproben konsistent zu denen an der Gesamtstichprobe wieder und unterstrich die Notwendigkeit der Modellreformulierung über die Spezifikation eines Speed-Parameters für diejenigen Subtests, die von der Testzeitbegrenzung stark betroffen waren. Auch die Klassenaufteilung aufgrund logisch voneinander abhängiger Items im Subtest SPARK fand sich in den Zufallsstichproben wieder.
Überdies erweisen sich die Itemparameter als sehr invariant über die Stichproben hinweg betrachtet. Hierbei muss jedoch beachtet werden, dass die hier vorgenommene *Zufalls*aufteilung nicht den strengsten Test auf Itemparameterinvarianz darstellt, wenngleich dieses Vorgehen für eine Kreuzreplikation des Rasch-Modells mit seiner expliziten Forderung nach Konstanz dieser Parameter unerlässlich war. *Systematische* Unterschiede der Itemparameter zwischen Personengruppen werden hierüber sehr unwahrscheinlich identifiziert. Daher stellt die vorgenommene Überprüfung der Personenhomogenität über das mixed Rasch-Modell auch in den Zufallsstichproben einen strengeren Test der Itemparameterinvarianz dar, weil hierdurch Personengruppen identifiziert werden, welche sich maximal hinsichtlich ihrer Itemparameter unterscheiden. Die dadurch in manchen Subtests wie schon in der Gesamtanalyse identifizierbare Personenheterogenität aufgrund von Speed-Effekten konnte auch hier durch

die Modellreformulierung nach Gleichung (47) „aufgelöst“ werden. Die über dieses Modell geschätzten Itemparameter waren über die Zufallsstichproben hinweg betrachtet wiederum sehr hoch positiv korreliert, sodass sich auch hierdurch ein unterstützender Befund der Stichprobenunabhängigkeit der Itemschwierigkeiten ergab.

Die Kreuzvalidierungsanalysen unterstützen durch die nachgewiesene hohe Konvergenz der Ergebnisse somit insgesamt betrachtet die in dieser Arbeit vorgenommene Itemselektion anhand der Gesamtstichprobe und legen somit auch ein Fundament für die Kriteriums-validierungsanalysen dieser Skalen.

16. Limitierungen der Ergebnisse zu den Skalenanalysen

Das Kapitel der Skalenkonstruktion nach den Methoden der Rasch-Modelle abschließend, sollen an dieser Stelle noch einige kritische Punkte zu den hier durchgeführten Analysen der Testoptimierungen benannt werden. Dies betrifft zunächst die Modellanpassung einiger Subtests über eine zusätzliche Speed-Komponente, im Weiteren generell die Itemselektion nach Fit-Statistiken und schließlich die Ergebnisse der Kreuzreplikationen.

Die Speed-Komponenten-Modellreformulierung nach Gleichung (47) erbrachte zwar in den von der Testzeitbegrenzung betroffenen Tests eine deutlich verbessere Modellpassung, allerdings kann sie, ganz im Sinne der Ausführungen in Kapitel 9, nicht letztes Ziel einer konstruktvaliden Abbildung durch einen Test sein. Die Passung eines Modells durch zusätzliche Parameter wird stets verbessert, zumal post hoc wie im vorliegenden Falle, wie etwa am Beispiel saturierter Modelle mit ebenso vielen Parametern wie unabhängigen Daten bekannt. Ob dies allerdings eine *Konstrukt*optimierung darstellt, kann aus folgendem Grund bezweifelt werden. Es ist das erklärte Ziel der hier konzipierten Tests, die Personenfähigkeiten möglichst eindimensional zu messen, um gerade in Bezug auf Kriteriumsleistungen zu validen Aussagen über Zusammenhänge eben dieser spezifizierten Fähigkeit zu gelangen. Es sollten daher in etwaigen Revisionen dieser Tests die hier dargestellten Analysen unter weniger stark zeitbegrenzten Bedingungen durchgeführt werden, ganz gleich, ob mit der hier für manche Subtests vorgenommenen Modellspezifikation mit einer zusätzlichen Speed-Komponente der Einfluss von Bearbeitungsgeschwindigkeit *statistisch* kontrolliert wurde.
Diese Unterscheidung zwischen Test- und Konstruktoptimierung trifft in analoger Weise auch auf die Technik der Itemselektion nach Item-Fit-Indizes zu. Wie bereits in Kapitel 9 dargelegt, ist die Gültigkeit des Rasch-Modells nach dem Entfernen von Items über Fit-Indizes kritisch

zu sehen, weil damit rein logisch das Ergebnis der Vergleiche zwischen zwei Personen nicht mehr unabhängig von der Itemauswahl sein kann. Dieses strenge Urteil ist natürlich dann nicht angemessen, wenn Items bspw. aufgrund von Rateanfälligkeit wegen ungenügender Distraktorkonstruktion aus dem Test entfernt werden müssen, also nicht aufgrund *struktureller* Merkmale wie Mehrdimensionalität innerhalb oder auch zwischen Items, sondern aufgrund von rein technischen Konstruktionsfehlern. Ansonsten wäre diesem strengen Kriterium aber zuzustimmen. Sollten sich diese Effekte nämlich selbst bei streng regelgeleitet konstruierten Items wiederfinden lassen, dann wäre die Generalisierbarkeit des hypothetisierten Konstruktes bzw. der Schluss auf das Itemuniversum sehr kritisch. Zwar wurde in dieser Arbeit weitestgehend versucht, die Items regelgeleitet zu konstruieren, Testrevisionen sollten aber diesen Weg noch stringenter gehen, um die geforderte Raschkonformität zu überprüfen. Auch positiv (im Sinne der Modellgeltung) ausfallende statistische Modelltests tragen nur so weit, wie die Befunde an verschiedenen Personen- aber ebenso Itemstichproben repliziert werden können. In diesem Sinne ist auch die Rasch-Konformität der hier untersuchten Tests nicht unzweifelbar, ganz gleich, ob diese durch Itemselektion und/oder Formulierung eines Rasch-Modells mit Kontrolle der Speed-Komponente verbessert wurde. Die angestellten Analysen markieren also lediglich einen Punkt, von dem in weiteren Untersuchungen ausgegangen werden kann.

Schließlich bedürfen auch die hier Ergebnisse der Kreuzreplikationen zu den Skalenanalysen einschränkender Anmerkungen. Wie Guttman (1977, S. 86) in seiner Kritik am Signifikanztest betont, ist „...the essence of science ... replication“, da es sich bei der konventionellen statistischen Inferenz genau genommen um ein Verfahren handelt, welches auf die Relation zwischen Population und Stichprobe (Untermenge) abzielt, aber nichts über die Relation zwischen zwei Stichproben (Untermengen) aus derselben Population aussagt. Ganz in diesem Sinne ist also das hier gewählte Vorgehen einer Kreuzreplikation der Itemparameter und der Itemfitindizes zusätzlich zu ihrer inferenzstatistischen Absicherung über die Gesamtstichprobe sinnvoll. Im Falle der vorliegenden Skalenanalysen ist allerdings einschränkend anzumerken, dass, im Gegensatz zur gängigen Praxis der Hypothesentestung, die Annahmen über modellkonforme bzw. nicht modellkonforme Items anhand der Analysen aus der Gesamtstichprobe gebildet wurden und nicht an einem *unabhängigen* Datensatz überprüft wurden. Vielmehr wurden die Hypothesen *erst post hoc aus den Analysen der Gesamtstichprobe aufgestellt* und an zwei *abhängigen* Zufallsstichproben aus der Gesamtstichprobe nachträglich überprüft. Wie Stelzl (2005, S. 137) hierzu ausführt, ist die Wahrscheinlichkeit, in einer der beiden derart gebildeten Kreuzreplikationsstichproben bei Gültigkeit der Nullhypothese zwar gleich Alpha. Hingegen ist die bedingte Wahrscheinlichkeit, in *beiden* Zufallsstichproben ein

signifikantes Ergebnis zu erzielen, wenn in den Gesamtanalysen bereits ein Alphafehler begangen wurde, komplizierter zu berechnen. Im vorliegenden Fall hätte man hierbei die Verteilung der Itemfitindizes in beiden Zufallsstichproben bei gegebenen Itemfitindizes in der Gesamtstichprobe beachten müssen, was jedoch mit einem unverhältnismäßig hohem Aufwand bei gleichzeitig geringem Erkenntnisgewinn verbunden gewesen wäre, denn Stelzl (2005) merkt generell zur Technik einer Kreuzvalidierung über einen Zufallsaufteilung der Gesamtdaten in zwei Stichproben an:

> Was in den Gesamtdaten ‚signifikant' war, hat gute Aussichten, auch bei Teilung des Datenmaterials in zwei Hälften ‚signifikant' zu sein – und zwar auch dann, wenn es sich bei der Signifikanz um einen (durch langes Suchen und Hypothesen im nachhinein herbeigeführten) Alpha-Fehler handelt. (S. 138)

Da allerdings im Falle der Rasch-Konformitätsüberprüfung von Items die Hypothese der Modellgültigkeit in Form der Nullhypothese formuliert ist, würde dies bedeuten, dass womöglich gut zu den Modellannahmen passende Items fälschlich selegiert würden. In den hier durchgeführten Kreuzvalidierungsanalysen ist diese Gefahr in sofern vermindert, weil der Vergleich der Itemfitindizes in den jeweiligen Stichproben in erster Linie anhand der Effektstärkemaße und nicht alleine über die auf Signifikanzaussagen abzielenden z-standardisierten Werte vorgenommen wurde. Analog gilt dies auch für die Überprüfung der Personenhomogenitätsannahme über die informationstheoretischem Maße BIC und CAIC, da auch sie Effektstärkemaße für den Vergleich der Einklassen- gegenüber der Zweiklassenlösung des mixed Rasch-Modells darstellen.
Gleichwohl muss der generellen Kritik von Stelzl (2005) am Vorgehen, an den Gesamtdaten aufgestellt Hypothesen über die Modellgültigkeit im Nachhinein an abhängigen Daten zu kreuzvalidieren, zugestimmt werden. Es stellt ein methodisch schwächeres Design dar als eine Kreuzreplikation an gänzlich neuen Daten, auch wenn eine Replikation anhand einer Zufallsaufteilung in der Literatur gängige Praxis ist und auch in Lehrbüchern empfohlen wird (s. z. B. Bortz, 1999, S. 562). Jedoch hätte auch auf diese Art der Kreuzreplikation nicht verzichtet werden können, denn, wie Cliff (1983) in Bezug auf dieses Vorgehen treffend anmerkt:

> …it does allow one to test his model, rather than leaving the investigator, and the consumers of his research, in the position of trying to make use of results which they know are unstable to an unknown degree, or, worse yet, trying to do so when they do not know that the results are unstable. (S. 124)

In diesem Sinne verstehen sich die hier durchgeführten Kreuzreplikationen einerseits als Approximationen für Erhebungen an neuen Daten, andererseits als unerlässliche Bedingungen für eine Verallgemeinerbarkeit der hier gemachten Modellgültigkeitsaussagen des Rasch-Modells und schließen somit das Kapitel der Modellgeltung ab.
Auf Basis der aus den beschriebenen Skalenanalysen resultierenden Skalen werden im Folgenden Validitätsanalysen unter verschiedenen Fragestellungen und mit verschiedenen Stichproben berichtet.

17. Validitätsanalysen

Die in den folgenden Kapiteln berichteten Analysen basieren mit Ausnahme von SPARK durchweg auf den Personenparametern der hier jeweils spezifizierten Rasch-Modelle. Um jeweils zu einem Gesamtmaß verbaler und numerischer Intelligenz zu gelangen, wurden die jeweiligen Subtests durch eine Test-Equating-Prozedur nach Moulton (2004, S. 6f.) zu einer verbalen und einer numerischen Intelligenzskala kombiniert. Für die von Speed-Effekten betroffenen Subtests wurden für diese Prozedur die Personenparameter nach der Modell-Formulierung gemäß Gleichung (47) verwendet, für die nicht hiervon betroffenen diejenigen des dichotomen Rasch-Modells nach Gleichung (2). In gleicher Weise (allerdings mit der Formulierung nach dem Multifacetten-Rasch-Modell) wurden die Kreativitätsfacetten so zu einem Gesamtmaß für Kreativität zusammengefasst. Das Maß der fluiden Intelligenz (Matrizen-Test) waren die nach Gleichung (47) geschätzten Personenparameter, dasjenige für empiriebezogenes Denken die nach dem Multifacetten-Rasch-Modell geschätzten. Für SPARK wurden die Summenrohwerte verwendet. Somit resultierten sechs Testskalen: verbale, numerische und fluide Intelligenz sowie empiriebezogenes Denken, Kreativität und SPARK.

Die folgenden Kapitel berichten die Ergebnisse verschiedener Fragestellungen zu Validitäten der Testverfahren und der Abiturleistungen. Die Darstellung geht hierbei zunächst auf die fächerdiskriminante Validität ein. Darauf folgend wird der zentrale Teil der Validitätsfrage-

stellung, die retrospektive, prädiktive und inkrementelle Validität der Leistungstests dargestellt, sowie abschließend die Validität der Fragebögen unter einer Normal- gegenüber und einer Faking-good-Instruktion untersucht.

17.1 Fächerdiskriminante Validität

Ziel dieser Analyse ist die Beantwortung der Frage, ob die auf theoretischer Basis speziell für das Psychologiestudium abgeleiteten Testverfahren auch Unterscheidungen gegenüber Studierenden des ersten Semesters aus anderen Fächern im Sinne eines typischen Leistungsprofils von Psychologiestudierenden treffen kann. Dies ist insbesondere dann von Interesse, wenn man, wie im vorliegenden Fall, eine bereits durch die Numerus-clausus-Regelung hochselegierte Stichprobe vorliegen hat und fachspezifisch konstruierte Verfahren auch zwischen Bewerbern aus Fächern mit derartiger Zulassungsbeschränkung differenzieren sollen. Die Differenzierung durch die Testverfahren sollte sich also weitgehend unabhängig vom Einfluss der Abiturleistungen zeigen.

Für diese Analyse wurden Daten von Studierenden des ersten Semesters im Wintersemester 2004/05 aus den Fächern mit den meisten Testteilnehmern aus der Gesamtstichprobe herangezogen. Dies waren, neben Psychologie (n = 81), Jura (n = 91), Volkswirtschaftslehre (VWL), (n = 40) Geographie (n = 39) und Soziologie (n = 24).

Eine univariate Varianzanalyse nach dem Allgemeinen Linearen Modell mit der Zugehörigkeit der Probanden zum Studienfach als unabhängigen Faktor und der Abiturdurchschnittsnote als abhängige Variable ergab auch statistisch bedeutsame Unterschiede zwischen den betrachteten Fächern in der mittleren Abiturnote. Die Ergebnisse der Varianzanalyse zeigt Tabelle 69.

Tabelle 69: Ergebnisse der einfaktoriellen, univariaten Varianzanalyse über die Abiturdurchschnittsnote in den Fächern Psychologie (n = 81), Jura (n = 91), VWL (n = 40), Geographie (n = 39) und Soziologie (n = 24), jwls. 1.Semester

Quelle	**Quadratsumme vom Typ III**	**df**	**Mittel der Quadrate**	***F***	***p***	η^2
Korrigiertes Modell	49.97	4	12.49	57.69	.00	.47
Konstanter Term	842.61	1	842.61	3891.17	.00	.93
Studienfach	49.97	4	12.49	57.69	.00	.47
Fehler	55.86	258	0.21			
Gesamt	962.65	263				
Korrigierte Gesamtvariation	105.84	262				

Es zeigt sich ein signifikanter Haupteffekt für den Faktor Studienfach, der 47% der Varianz in den Abiturdurchschnittsnoten aufklärt. Im Anschluss durchgeführte Post-hoc-Tests (nach Dunnets-T3) ergaben jeweils signifikant niedrigere mittlere Abiturnoten der Studienteilnehmer beider Numerus-clausus-Fächer Psychologie und Jura ($M_{\text{Psychologie}} = 1.52$, $SD_{\text{Psychologie}} = 0.50$; $M_{\text{Jura}} = 1.42$, $SD_{\text{Jura}} = 0.37$) gegenüber allen übrigen ($M_{\text{VWL}} = 2.47$, $SD_{\text{VWL}} = 0.57$; $M_{\text{Geographie}} = 2.23$, $SD_{\text{Geographie}} = 0.46$; $M_{\text{Soziologie}} = 2.37$, $SD_{\text{Soziologie}} = 0.42$) mit jeweils $p < .01$. Keine signifikanten Unterschiede ergaben sich zwischen Psychologie und Jura und zwischen den Studienfächern ohne NC-Regelung (jwls. $p > .05$).
Diese Ergebnisse legen es nahe, den möglichen Einfluss der Abiturleistung auf die Testergebnisse zu kontrollieren. Etwaige Mittelwertsunterschiede in den Testleistungen zwischen Erstsemestern der Psychologie gegenüber solchen aus Studienfächern *ohne* Abiturnotenvorselektion wären durch den Einfluss besserer schulischer Leistungsfähigkeit womöglich konfundiert. Eine nahe liegende statistische Bereinigung der Testergebnisse um den Einfluss der Abiturdurchschnittsnote über eine Kovarianzanalyse ist in diesem Falle allerdings nach Bortz (1999, S. 357f.) problematisch, da die Kontrollvariable mit dem Faktor Studienfach korreliert ist. Hierdurch würde nicht alleine die Fehler- sondern auch die Treatmentvarianz reduziert werden und der eventuell noch vorhandene Effekt schwer interpretierbar sein. Allerdings ist im vorliegenden Fall die Fragestellung genau die, ob unter statistischer Kontrolle der Abiturnote möglicherweise vorhanden Unterschiede zwischen den Testteilnehmern verschiedener Fächer verschwinden und somit die Abiturnote die Hauptursache für die Mittelwertsunterschiede wäre. Die angestrebte studienfachbezogene Differenzierungsfähigkeit der Testverfahren wäre damit widerlegt. Die Ergebnisse einer Kovarianzanalyse ist daher ebenso Teil der folgenden Analysen.

17.1.1 Ergebnisse der Fachvergleiche in den Testleistungen

Zur Untersuchung der Fragestellung, ob die Verfahren zwischen Psychologiestudierenden aus dem ersten Semester und solchen anderer Fächer differenzieren, wurde eine multivariate Varianzanalyse nach dem Allgemeinen Linearen Modell durchgeführt. Die z-standardisierten Personenparameter in verbaler, numerischer und fluider Intelligenz (Matrizentest), dem empiriebezogenen Denken, Kreativität und die z-standardisierten Rohwerte aus SPARK gingen hierbei als abhängige Variablen ein, die Studienfachzugehörigkeit als unabhängige Variable.

Um nur die spezifisch interessierenden Mittelwertsunterschiede zwischen Studierenden der Psychologie mit jeweils denjenigen anderer Fächer zu untersuchen, wurden einfache geplante Kontraste berechnet.

Bortz (1999, S. 575f.) empfiehlt als globale multivariate Teststatistik für kleine Stichproben Pillais Spurkriterium als konservativere multivariate Prüfgröße gegenüber Wilks Lambda. Aufgrund der im vorliegenden Fall stark unterschiedlichen Stichprobengrößen werden daher beide Teststatistiken berichtet. Tabelle 70 zeigt die Ergebnisse der multivariaten Analyse.

Tabelle 70: Ergebnisse des multivariaten Tests zu Mittelwertsunterschieden in den Testleistungen von Studienteilnehmern aus verschiedenen Studienfächern

Teststatistik	**Wert**	***F***	**Hypothese *df***	**Fehler *df***	***p***
Pillai-Spur	.24	2.87	24.00	1072.00	.00
Wilks-Lambda	.77	2.95	24.00	925.68	.00

Beide Teststatistiken zeigen einen signifikanten Einfluss des Faktors Studienfach auf die Testleistungen. Nach Wilks Lambda werden hierbei 23% der Unterschiede durch die Studienfachzugehörigkeit aufgeklärt (Wilks Lambda ist hierbei als eine inverse Statistik des aufgeklärten Varianzanteils zu verstehen). Abbildung 14 gibt einen Überblick über die Einzelergebnisse.

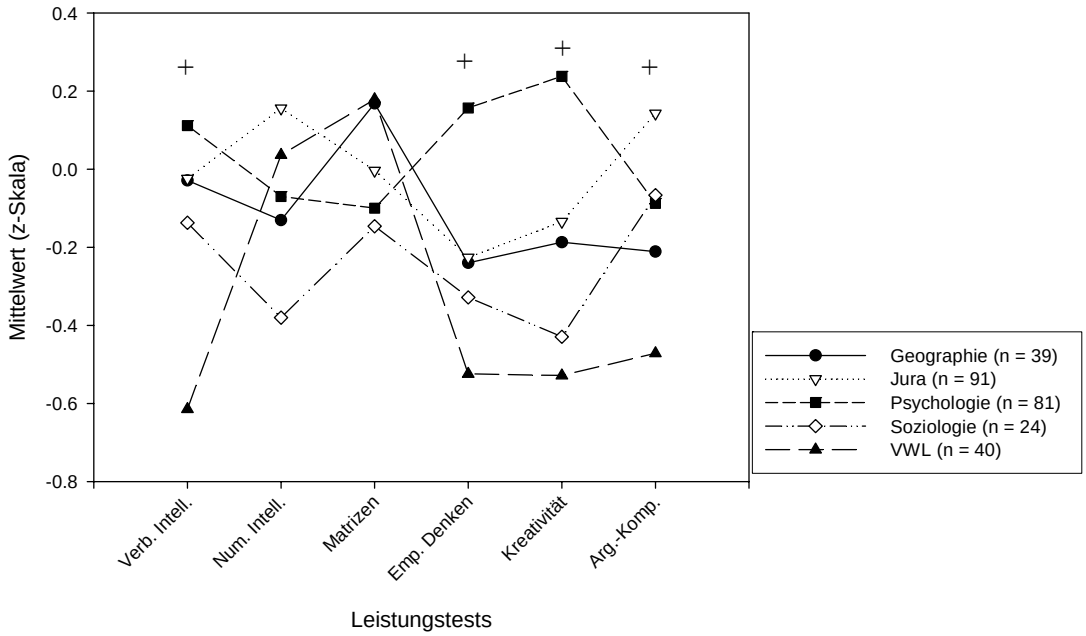


Abbildung 14: Mittelwerte der Testleistungen von Erstsemestern verschiedener Fächer
Anmerkung. Verb. Intell.: Skala verbale Intelligenz; Num. Intell.: Skala numerische Intelligenz; Emp. Denken: Skala empiriebezogenes Denken; Arg.-Komp.: Skala Passive Argumentationskompetenz (SPARK). +: signifikante einfache Kontraste zwischen Psychologie und mindestens einem weiteren Fach (nähere Erläuterungen s. Text). Psychologie (n = 81), Jura (n = 91), VWL (n = 40), Geographie (n = 39) und Soziologie (n = 24), jwls. 1.Semester

Es ergeben sich in vier von sechs Skalen bedeutsame Mittelwertsdifferenzen zwischen Studierenden des ersten Semesters in Psychologie und denjenigen der anderen Fächer, wobei die Psychologiestudierenden mit Ausnahme der Skala SPARK im Durchschnitt höhere Durchschnittsleistungen aufweisen. Auffallend niedrig, außer für SPARK, fällt insgesamt die mittlere Leistung der Soziologie-Studierenden aus. Auch diejenige der Studierenden von VWL, mit Ausnahme der Skalen zur numerischen und fluiden Intelligenz, sind vergleichsweise sehr niedrig. Ein sehr ausgeprägtes Leistungsprofil zeigen der Studierenden der Geographie, welches insbesondere in den spezifischeren Tests (empiriebezogenes Denken, Kreativität und SPARK) gegenüber den Intelligenzskalen deutlich niedriger ausfällt. Insgesamt betrachtet sind die Unterschiede zwischen den Fächern in den Skalen mit spezifischen Leistungsinhalten (empiriebezogenes Denken, Kreativität und SPARK) und sprachlicher Intelligenz ausgeprägter als in numerischer und fluider Intelligenz.
Einen detaillierten Überblick über die Art der Mittelwertsunterschiede anhand einfacher geplanter Kontraste mit Angabe der Effektstärken der Studierenden der Psychologie gegenüber denen anderer Fächer gibt Tabelle 71.

Tabelle 71: Ergebnisse der einfachen multivariaten Kontraste: Richtung des Mittelwertsunterschiedes und Effektstärken nach Cohen. Psychologie (n = 81), Jura (n = 91), VWL (n = 40), Geographie (n = 39) und Soziologie (n = 24), jwls. 1.Semester.

	Verb. Intell.	**Numer. Intell.**	**Matrizen**	**Emp. Denken**	**Kreativ.**	**SPARK**
Psych. vs. Soziologie	n.s.	n.s.	n.s.	Psych. > Soz. d = .83 p < .05	Psych. > Soz. d = .73 p < .01	n.s.
Psych. vs. Jura	n.s.	n.s.	n.s.	Psych. > Jura d = .45 p < .01	Psych. > Jura d = .38 p < .05	n.s.
Psych. vs. Geographie	n.s.	n.s.	n.s.	Psych. > Geogr. d = .08 p < .05	Psych > Geogr. d = .49 p < .05	n.s.
Psych. vs. VWL	Psych. > VWL d = .84 p < .01	n.s.	n.s.	Psych. > VWL d = .93 p < .001	Psych. > VWL d = .96 p < .01	Psych. > VWL d = .49 p < .05

Anmerkung. Verb. Intell.: Skala verbale Intelligenz; Num. Intell.: Skala numerische Intelligenz; Emp. Denken: Skala empiriebezogenes Denken; SPARK: Skala Passive Argumentationskompetenz. Psych.: Psychologie; VWL: Volkswirtschaftslehre

Auf Ebene der einfachen multivariaten geplanten Kontraste ergibt sich eine deutliche Überlegenheit der Psychologiestudierenden besonders in den Skalen zu spezifischen Leistungen („Empiriebezogenes Denken“ und Kreativität). Die Effektstärken liegen, mit Ausnahme des Vergleichs zwischen Psychologie und Geographie bezüglich empiriebezogenen Denkens, im mittleren bis hohen Bereich. Keine signifikanten Unterschiede zeigen sich hingegen in den Skalen zur numerischen und fluiden Intelligenz. Allerdings lag bei diesen auch die post-hoc berechnete Teststärke mit 1-β = .51 (numerische Intelligenz) bzw. 1-β = .32 (fluide Intelligenz) unter der üblicherweise geforderten von mindestens .80 (s. hierzu etwa Bortz & Döring, 1995, S. 567). Bei den übrigen Skalen lag diese durchwegs darüber bzw. nicht darunter (1-$\beta_{\text{verbale Intelligenz}}$ = .85, 1-$\beta_{\text{empiriebezogenes Denken}}$ = .91, 1-$\beta_{\text{Kreativität}}$ = .94, 1-β_{SPARK} = .80). Für verlässlichere Aussagen zu Unterschieden der verglichenen Studienfächer in diesen Skalen müssten daher in Replikationsuntersuchungen die Stichprobengrößen erhöht werden. Eine Ausnahme bei den Mittelwertsunterschieden kognitiver Leistungsfähigkeit bildet allerdings die Skala zur verbalen Intelligenz, in der die Psychologie-Studierenden denjenigen in VWL mit hoher Effektstärke überlegen sind. Insbesondere verglichen mit diesem Fach ergeben sich

zudem in den spezifischen Skalen mittlere bis überwiegend hohe Effektstärken zugunsten der Psychologie-Studierenden.
Gegenüber dem Numerus-clausus-Fach Jura differenziert die Testbatterie lediglich in den Skalen zum empiriebezogenen Denken und zur Kreativität. Allerdings trifft dieser Befund mit Ausnahme von VWL auch auf die Vergleiche mit anderen Studienfächern zu. Wahrscheinlich schlägt sich hier ein Selbstselektionseffekt nieder, erweisen sich doch die Psychologie-studierenden gerade in diesen Skalen zu spezifischen Fähigkeiten *allen* kontrastierten Fächern als überlegen. Gleichwohl liefern die hier gefundenen Effekte lediglich erste Hinweise zur Art der Differenzierungsfähigkeit der Testverfahren und müssten in weiteren Erhebungen mit mehr Studienfächern und jwls. größeren Stichprobenumfängen untersucht werden.

Der Einfluss der Abiturnote auf die beobachteten Mittelwertsdifferenzen stellt allerdings die eigentlich kritische Überprüfung der fächerbezogenen Diskriminationsfähigkeit der Testverfahren dar. Die Ergebnisse der hierzu angewandten Kovarianzanalyse zeigen die folgenden Tabellen und Erläuterungen.

Tabelle 72: Ergebnisse des multivariaten Tests zu Mittelwertsunterschieden in den Testleistungen von Studienteilnehmern aus verschiedenen Studienfächern unter statistischer Kontrolle der Abiturnote

Effekt	**Teststatistik**	**Wert**	***F***	**Hypothese *df***	**Fehler *df***	***p***
Abiturnote	Pillai-Spur	.05	2.16	6.00	252.00	.04
	Wilks-Lambda	.95	2.16	6.00	252.00	.04
Studienfach	Pillai-Spur	.23	2.65	24.00	1020.00	.00
	Wilks-Lambda	.78	2.71	24.00	880.33	.00

Beide multivariaten Teststatistiken weisen auf einen signifikanten Einfluss der Abitur-durchschnittsnote als Kovariate auf die Testergebnisse hin. Allerdings ist der Überlappungs-bereich allgemeiner schulischer Leistungsfähigkeit und den Testergebnissen nicht groß. Nach Wilks Lambda als inverser Statistik des aufgeklärten Varianzanteils liegt dieser nur bei 5%. Demgegenüber können nach Wilks Lambda auch nach Auspartialisierung der Abitur-durchschnittsnote noch 22% der Varianz in den Testergebnissen aus der Studienfachzuge-hörigkeit aufgeklärt werden.

Wie in der Multivariaten Varianzanalyse dienten auch im Folgenden geplante einfache multivariate Kontraste als Maß fachspezifischer Differenzierungsfähigkeit. Tabelle 73 zeigt die Ergebnisse.

Tabelle 73: Ergebnisse der einfachen multivariaten Kontraste nach Auspartialisierung der Abiturnote: Richtung des Mittelwertsunterschiedes und Effektstärken nach Cohen. Psychologie (n = 81), Jura (n = 91), VWL (n = 40), Geographie (n = 39) und Soziologie (n = 24), jwls. 1.Semester

	Verb. Intell.	**Numer. Intell.**	**Matrizen**	**Emp. Denken**	**Kreativ.**	**SPARK**
Psych. vs. Soziologie	n.s.	n.s.	n.s.	n.s.	Psych. > Soz. d = 1.26 p < .05 (d = .73)	n.s.
Psych. vs. Jura	n.s.	n.s.	n.s.	Psych. > Jura d = .45 p < .05 (d = .45)	n.s. (d = .38)	n.s.
Psych. vs. Geographie	n.s.	n.s.	Psych < Geogr. d = .29 p < .05	n.s. (d = .08)	n.s. (d = .49)	n.s.
Psych. vs. VWL	Psych. > VWL d = .87 p < .05 (d = .84)	Psych. < VWL d = .14 p < .05	Psych. < VWL d = .32 p < .01	Psych. > VWL d = 1.06 p < .001 (d = .93)	Psych. > VWL d = .93 p < .05 (d = .96)	n.s. (d = .49)

Anmerkung. Effektstärkemaße der einfachen geplanten multivariaten Kontraste der MANOVA nach Tabelle 71 zum Vergleich in Klammern. Psych.: Psychologie; Verb. Intell.: verbale Intelligenz; Numer. Intell.: numerische Intelligenz; Emp. Denken: empiriebezogenes Denken; Kreativ.: Kreativität.

Zunächst fällt auf, dass bei weiterhin bestehenden Mittelwertsunterschieden die Effektstärken erhalten bleiben oder größer ausfallen. Die angestrebte Differenzierungsfähigkeit der Testverfahren über verschiedene Fächer hinweg ist im Vergleich zu den Ergebnissen nach Tabelle 71 für die Hälfte der signifikanten Unterschiede also immer noch vorhanden und betrifft insbesondere die Tests zum „empiriebezogenen Denken“ und der Kreativität. Sie verschwinden hingegen dort, wo *vor* Auspartialisierung der Abiturdurchschnittsnote nur geringe Effekte zu beobachten waren. Neu hinzu tritt das bessere Abschneiden der Geographie- und VWL-Studierenden in fluider Intelligenz (Matrizentest) als auch die leichte Überlegenheit der VWL-Studierenden in numerischer Intelligenz. Die Richtung dieses

Unterschiedes ist allerdings inhaltlich schwer zu erklären. Problematisch hierfür ist insbesondere die Annahme der multivariaten Kovarianzanalyse, dass die Kovariate messfehlerfrei gemessen wurde. Werden nicht perfekt reliable Kovariaten eliminiert, bleibt auch nach Auspartialisierung ein undefinierbarer Rest an Fehlervarianz in den systematischen Effekten bestehen. Wie Stelzl (2005, S. 277f.) hierzu anmerkt: „Die Variable, die auspartialisiert werden soll ... kann nicht frei von Messfehlern und irrelevanten Komponenten genau in der Zusammensetzung gemessen werden, wie sie sich auf die andere(n) Variable(n) ... auswirkt. Als Folge davon gelingt die Auspartialisierung nur unvollständig". Allerdings geht man im kovarianzanalytischen Design nun davon aus, dass sich die Personen nicht mehr in der Abiturnote unterscheiden und die verbleibenden Unterschiede auf Intelligenz rückführbar sind. Misst die Abiturdurchschnittsnote allerdings ebenso Intelligenz, wird auch der Zusammenhang zwischen Gruppe und Intelligenz auspartialisiert, aber nur unvollständig. Im Durchschnitt korrelierten die Intelligenzmaße in der Gesamtgruppe der hier analysierten Probanden mit der Abiturdurchschnittsnote zu r = -.38. Hieraus entstehen drei Erklärungsansätze:
Die Effekte können in der Tat (1) systematisch sein, wobei die Ursachen für diese Systematik aus den Daten nicht rekonstruierbar sind. Sie können (2) dadurch bedingt sein, dass auch ein *spezifischer Anteil* an Intelligenz mit der Abiturnote erfasst und auspartialisiert wird. Der verbleibende Effekt bestünde sodann aus dem Intelligenzanteil, welcher sich aus Intelligenz minus demjenigen Anteil zusammensetzt, der durch die Abiturnote vorhergesagt wird. Was eben diesen Anteil inhaltlich ausmacht, bleibt jedoch unklar. Schließlich kann der Effekt auch (3) aufgrund der nicht vollständigen Auspartialisierung ein Zufallseffekt aufgrund eines Alpha-Fehlers sein, wobei dies zudem von der Reliabilität bzw. Fehlervarianzanteil der abhängigen Variablen abhängt.
In jedem Fall bleibt die genaue Ursache der unerwartet neu zu beobachtenden Mittelwertsdifferenzen im Unklaren, sodass an dieser Stelle lediglich Erklärungsansätze möglich sind.

17.1.2 Zusammenfassung der Analysen fächerdiskriminanter Validität

Für die hier durchgeführten Analysen seitens der Diskriminierungsfähigkeit der Testverfahren ergeben sich nur hinsichtlich spezifischer Testkomponenten des „empiriebezogenen Denkens", der Kreativität und teilweise der passiven Argumentationskompetenz Mittelwertsdifferenzen in der „erwünschten" Richtung: hier zeigen sich die Psychologiestudierenden des ersten Semesters denen anderer Fächer überlegen.

Die Analyse der Effekte der Abiturdurchschnittsnote auf die Testleistungen mit geplanten einfachen multivariaten Kontrasten in einer Kovarianzanalyse ergab, dass die nach der multivariaten Varianzanalyse bestehenden geringen Mittelwertsdifferenzen insignifikant ausfielen, wohingegen große Effektstärken weiterhin erhalten blieben und signifikant waren. Die somit teilweise erhaltene Differenzierungsfähigkeit der Testverfahren betraf insbesondere die Skalen zu spezifischen Leistungen („empiriebezogenes Denken“ und Kreativität). Neu auftretende signifikante Mittelwertsdifferenzen, die insbesondere für eine Überlegenheit der VWL-Studierenden gegenüber denen der Psychologie besonders in fluider und numerischer Intelligenz sprachen, ließen sich aufgrund methodischer Probleme nur schwer interpretieren.

Die Befunde aus beiden Analysen integriert betrachtet lässt sich feststellen, dass die fächerdiskriminante Validität der Testverfahren sich nur in spezifischen Leistungen zeigt.
Streng genommen differenziert das Gesamtverfahren daher lediglich in den Skalen zum „empiriebezogenen Denken“ und der Kreativität in der hypothetisierten Richtung. Im Rückbezug auf die Ergebnisse der Anforderungsanalyse (s. Kap. 7.2) ist dieser Befund nicht ohne eine gewisse Ironie, da die Kreativitätsdimension die einzige ist, welche von den anforderungsanalytisch abgeleiteten Personenmerkmalen differenziert. Die Dimension des „empiriebezogenen Denkens“ ging nicht direkt aus der Anforderungsanalyse hervor, sondern aus Überlegungen des Autors als fachspezifischere Eignungskomponente des Gesamtverfahrens. Wie bereits in Kap. 7.3 kritisch angemerkt, zeigt sich nun auf empirischer Ebene der Mangel an Fachspezifität der abgeleiteten Anforderungsdimensionen.
Kritisch an den Ergebnissen anzumerken ist außerdem, dass gerade in den anderen Fächern als Psychologie noch mit Testteilnahmeeffekten gerechnet werden muss, was die Repräsentativität dieser Stichproben fraglich erscheinen lässt und die Verallgemeinerbarkeit der Ergebnisse einschränkt. Den Aussagen über die Differenzierungsfähigkeit kommt somit vorerst der Status einer Arbeitshypothese zu.

17.2 Retrospektive und prädiktive Validitätsanalysen

Für die Analysen dieses Kapitels werden zum einen die Leistungen in den zurückliegenden Vordiplomprüfungen von Teststeilnehmern aus dem Hauptstudium in Psychologie korrelationsstatistisch zu den Testleistungen, der Abiturdurchschnittsnote und den Abitureinzelnoten in Deutsch und Mathematik analysiert. Zum anderen werden die Testleistungen und Abiturnoten der Psychologie-Studierenden mit Studienbeginn im Wintersemester 2004/05 zu Klausurleistungen im Fach Methodenlehre und in einer Orientierungsprüfung korrelationsstatistisch untersucht. Weitere Kriteriumswerte lagen für letztere Substichprobe beim Verfassen dieser Arbeit noch nicht vor. Vorab wichtig zu betonen ist, dass die in diesem Zusammenhang abgeleiteten Inferenzschlüsse aus statistischen Verfahren nicht auf eine generelle Studierendenpopulation im Fach Psychologie bezogen werden können. Dies ist alleine schon aus dem bislang bestehenden Zulassungsmodus aus Numerus clausus, Wartezeit und Härtefällen nicht möglich. Die Zielpopulation der Inferenzschlüsse sind daher vielmehr die bislang nach dieser Regelung vorselegierten Psychologiestudierenden am Heidelberger Institut.

17.2.1 Retrospektive Validitätsanalysen

Zunächst wird in diesem Kapitel ein Überblick über wesentlichen Statistiken der Abitur- und Vordiplomnoten der Substichprobe aus dem Hauptstudium gegeben, welche für die nachfolgenden Validitätsuntersuchungen von Belang sind. Wichtig hierbei zu ergänzen ist, dass an dieser Erhebung auch 19 Studierende teilnahmen, die ihr Vordiplom in Psychologie nicht am Heidelberger Institut abgelegt hatten. Von diesen Probanden konnten daher keine Einzelnoten der Vordiplomprüfungen abgefragt werden. Die Analysen zur Güte der Kriteriumswerte basieren daher auf Daten von $n = 55$ Probanden, welche ihr Vordiplom am Heidelberger Institut abgelegt hatten. Die Analysen der Abiturnoten beziehen sich hingegen auf die komplette Hauptstudiumsstichprobe bzw. im Falle der Einzelnoten auf die Personen, welche sie berichtet hatten.

17.2.1.1 Analysen zur Güte der Abiturnoten in der Hauptstudiumsstichprobe

Einen Überblick über Mittelwerte und Standardabweichungen der Abiturnoten zeigt Tabelle 74.

Tabelle 74: Deskriptivstatistiken der Abiturnoten in der Hauptstudiumsstichprobe

Noten	***M***	***SD***	**Minimum**	**Maximum**	**Schiefe**	***N***
Abiturdurchschnitt	1.69	0.54	1.00	3.70	1.36	74
Mathematiknote	2.05	1.34	1.00	5.00	1.40	63
Deutschnote	1.84	0.75	1.00	4.00	0.76	61
Englischnote	1.79	0.92	1.00	6.00	1.86	56

Durch die Numerus-clausus-Regelung bedingt, sind die jeweiligen Notenverteilungen positiv schief verteilt und weisen geringe Streuungen auf, wodurch bereits mit Minderungen der Noteninterkorrelationen und Kriteriumskorrelationen zu rechnen ist.
Einen Überblick über die Abiturnoteninterkorrelationen in dieser Stichprobe gibt *Tabelle 75*.

Tabelle 75: Interkorrelationen der Abiturnoten in der Hauptstudiumsstichprobe

Noten	**Abitur-durchschnitt**	**Deutschnote**	**Mathematik-note**
Deutschnote	.46** (n = 61)		
Mathematiknote	.51** (n = 63)	-.008 (n = 59)	
Englischnote	.78** (n = 56)	.37** (n = 54)	.35** (n = 55)

Anmerkung. Einseitige Signifikanztestung ohne Korrektur für multiples Testen
*: $p < .05$
**: $p < .01$

Der höchste Zusammenhang ergibt sich zwischen der Englischnote und dem Abiturdurchschnitt, der geringste zwischen der Mathematik- und Deutschnote, wobei zu beachten ist, dass die Zusammenhänge der Einzelnoten mit der Abiturdurchschnittsnote durch die Eigenkorrelationsanteile überschätzt sind.
Cronbachs Alpha über die drei Einzelnoten als grobe Reliabilitätsschätzung der Abitur-durchschnittsnote liegt bei $\alpha = .61$ ($n = 53$). Dieser Wert stellt insofern eine untere Grenze der Reliabilität der Abiturdurchschnittsnote dar, da in diese nicht alleine die Abiturs- und

Vorjahresleistungen dieser drei Hauptfächer einfließen, sondern ebenso solche von Nebenfächern. Auch aus diesem Grund soll in den Validitätsanalysen daher das Hauptaugenmerk schulischer Leistungen auf der im Vergleich zu Einzelnoten reliableren Abiturdurchschnittsnote liegen.

17.2.1.2 Analysen zur Güte der Kriteriumswerte

Tabelle 76 gibt zunächst einen Überblick über die durchschnittlichen Testleistungen und deren Standardabweichungen in der Substichprobe aus dem Hauptstudium. Die Einzelnoten beziehen sich hierbei auf alle Probanden, die ihr Vordiplom am Heidelberger Institut abgelegt hatten, die durchschnittliche Vordiplomnote hingegen bezieht sich auf die vollständige Hauptstudiumsstichprobe.

Tabelle 76: Deskriptivstatistiken
der Vordiplomnoten in der Hauptstudiumsstichprobe

Fach	***M***	***SD***	**Schiefe**	***N***
Allgemeine Psychologie I	1.72	0.80	0.92	55
Allgemeine Psychologie II	1.60	0.85	2.26	55
Biopsychologie	1.94	0.83	1.63	55
Differenzielle Psychologie	2.05	1.06	1.31	55
Methodenlehre	1.92	1.09	1.56	55
Entwicklungspsychologie	2.02	1.05	.90	55
Sozialpsychologie	1.77	0.77	1.76	55
Vordiplom-Durchschnittsnote	1.85	0.70	1.55	74

Die Verteilungen der Einzelnoten und diejenige der Durchschnittsnote sind mit niedrigen Mittelwerten, Standardabweichungen und Werten der Schiefe größer Null deutlich positiv schief verteilt, weisen also einen Bodeneffekt auf. Bereits hierdurch ist mit deutlichen Minderungen der Kriteriumskorrelationen zu rechnen.
Um im Weiteren die Güte der Kriteriumswerte zu analysieren, wurden die einzelnen Vordiplomnoten interkorreliert. Wie bereits oben erläutert, nahmen auch Studierende an der Testung teil, die ihr Vordiplom nicht am Heidelberger Institut abgelegt hatten und von denen keine Einzelfachleistungen vorlagen. Die nachfolgende Tabelle 77 gibt daher die Interkorrelationen der Vordiplomnoten von 55 Heidelberger Probanden wieder.

Tabelle 77: Interkorrelationen der Vordiplomnoten (n = 55) in der Hauptstudiumsstichprobe

Vordiplomfächer	**Allg. Psych. I**	**Allg. Psych. II**	**Biopsych.**	**Diff. Psych**	**Methoden-lehre**	**Entwickl.-psych.**
Allg. Psych. I						
Allg. Psych. II	.52**					
Biopsych.	.22*	.29*				
Diff. Psych.	.33**	.42**	.51**			
Methodenlehre	.25*	.28*	.26*	.39**		
Entwickl.-Psych.	.28*	.48**	.28*	.44**	.28*	
Sozial-psych.	.24*	.19	.31*	.41**	.37**	.32**

Anmerkung. Einseitige Signifikanztestung (ohne Korrektur für multiples Testen).

**: p <.01.

Allg. Psych. I: Allgemeine Psychologie I; Allg. Psych. II: Allgemeine Psychologie I; Biopsych.: Biopsychologie; Diff. Psych.: Differenzielle Psychologie; Entwickl.-psych.: Entwicklungspsychologie; Sozial.-psych.: Sozialpsychologie

Nach den Effektstärkemaßen nach Cohen (s. Bortz & Döring, 1995, S. 568) streuen die Interkorrelationen vom mittleren bis in den hohen Bereich. Am niedrigsten fällt hierbei die Korrelation zwischen Sozialpsychologie und Allgemeiner Psychologie II aus, am höchsten diejenige von Differenzieller Psychologie und Biopsychologie. Der Mittelwert der Interkorrelationen nach Fishers z-Transformation liegt mit r = 34. im moderaten Bereich. Im Folgenden interessierte, inwieweit die Leistungen in den Einzelprüfungen mit der Gesamtleistung interkorrelierten. Hierzu wurden die Part-whole-korrigierten Trennschärfen der Einzelprüfungsleistungen berechnet. Das Ergebnis zeigt die folgende Tabelle 78.

Tabelle 78: Part-whole-korrigierte Trennschärfen der Teilprüfungen im Vordiplom (n = 55) gegenüber der Gesamtleistung (= Summe aller Fachnoten)

Fachnotenbezeichnung	**Part-whole-korrigierte Trennschärfe**
Allgemeine Psychologie I	.52
Allgemeine Psychologie II	.71
Biopsychologie	.63
Differenzielle Psychologie	.71
Methodenlehre	.58
Entwicklungspsychologie	.60
Sozialpsychologie	.64

Alle Trennschärfen liegen mit Werten größer $r_{it, -i}$ = .50 im hohen Bereich. Den größten Beitrag zur Gesamtleistung zeigen hierbei die Teilleistungen in Differenzieller und Allgemeiner Psychologie II, den relativ niedrigsten zeigt diejenige in Allgemeiner Psychologie I. Um die Messgenauigkeit der Vordiplomprüfungen zu untersuchen, wurde Cronbachs Alpha über die Teilprüfungen berechnet. Hierbei ergab sich ein Alpha von .85. Die interne Konsistenz der Vordiplomprüfungen kann in dieser Stichprobe daher als gut bezeichnet werden. In einem weiteren Schritt wurde die Eindimensionalität im Sinne der Faktorenanalyse der Einzelprüfungsleistungen analysiert. Eine oblique Hauptachsenanalyse mit Promax-Rotation (Delta-Wert = 4) extrahierte nach dem Kaiser-Guttman-Kriterium lediglich einen Faktor mit einer Varianzaufklärung von 48%. Die Faktorladungsmatrix gibt Tabelle 79 wieder.

Tabelle 79: Faktorladungsmatrix der Vordiplomeinzelprüfungsleistungen (n = 55)

Fachnotenbezeichnung	**Faktorladung**
Differenzielle Psychologie	.79
Allgemeine Psychologie II	.77
Biopsychologie	.70
Sozialpsychologie	.69
Entwicklungs-psychologie	.66
Methodenlehre	.64
Allgemeine Psychologie I	.57

Die Faktorladungen stehen in guter Übereinstimmung mit den Part-whole-korrigierten Trennschärfen nach Tabelle 78. Sechs der sieben Ladungen (mit Ausnahme von Allgemeiner Psychologie I) sind größer .60, wodurch nach Guadagnoli und Velicer (1988, zit. nach Bortz, 1999, S. 534) der Faktor interpretierbar ist.

Insgesamt betrachtet ist anhand dieser Analysen die Güte der Vordiplomnoten als gut zu bezeichnen. Cronbachs Alpha weist die Vordiplomnoten als reliables Kriterium aus, und die Leistungen im Vordiplom können nach der Faktorenanalyse als eindimensional bezeichnet werden.

17.2.1.3 Beziehungen der Testverfahren mit Abiturnoten in der Hauptstudiumsstichprobe

Um im Weiteren die Beziehungen der Verfahren mit Abiturnoten zu analysieren, wurden die Abiturleistungen der Substichprobe aus dem Hauptstudium mit den Testleistungen korreliert. Zusätzlich zu den Subtestergebnissen wurde ein Summenwert der Intelligenzskalen als Gesamtmaß der Intelligenztestleistungen gebildet. Nicht in die Analyse aufgenommen wurde die Skala zum empiriebezogenen Denken. Es ist bei Hauptstudiumsstudierenden nämlich davon auszugehen, dass diese Skala vielmehr das bereits gelernte methodische *Wissen* abfragt als ein *Denken* in empirisch-erfahrungswissenschaftlichen Kontexten. Gleichwohl lässt sich über die Analyse der Mittelwertsunterschiede auf dieser Skala von Studierenden aus dem Hauptstudium gegenüber solchen aus dem ersten Semesters untersuchen, ob die Skala überhaupt in der erwartbaren Richtung differenziert: Hauptstudiumsstudierende sollten hier im Durchschnitt bessere Leistungen aufweisen. Ein t-Test für Stichproben mit ungleichen Varianzen über die Personenparameter ergab in diesem Sinne auch eine deutliche Überlegenheit der Studierenden des Hauptstudiums ($M = 0.51$, $SD = 1.66$) gegenüber denjenigen des ersten Semesters ($M = -0.08$, $SD = 1.15$), $t\,(128)$, $p < .05$, $d = 0.45$. Die Validität der Skala im weiter gefassten Sinne einer Extremgruppendifferenzierung ist damit belegt. Tabelle 80 gibt einen Überblick über die Korrelationen der Abitur- und Testleistungen der Substichprobe aus dem Hauptstudium. (Da manche Probanden die Mathematik-, Deutsch- und Englischnote nicht angaben, kam es hierbei zu einer Verringerung der Stichprobengröße).

Tabelle 80: Korrelationen der Abiturleistungen mit den Testergebnissen in der Hauptstudiumsstichprobe ($N_{Abiturdurchschnitt}$ = 74, $N_{Mathematiknote}$ = 63, $N_{Englischnote}$ = 56)

Abiturnote	**Verbale Intelligenz**	**Numerische Intelligenz**	**Matrizen**	**Kreativ.**	**SPARK**	**Intelligenz-gesamtscore**
Abiturdurchschnitt	-.08	-.21*	-.10	-.07	-.40**	-.24*
Mathematiknote	-.19	-.26*	-.01	.03	-.12	-.18
Deutschnote	-.04	-.08	.02	.00	-.16	-.09
Englischnote	-.18	-.08	.04	-.15	-.42*	-.15

Anmerkung. Einseitige Signifikanztestung (ohne Korrektur für multiples Testen).
*: $p < .05$
**: $p < .01$

Die Korrelationen fallen insgesamt sehr niedrig aus und erreichen selten signifikante Höhen. Die höchste Korrelation besteht zwischen der Abiturdurchschnittsnote und SPARK, gefolgt von derjenigen mit dem Intelligenzgesamtscore sowie zwischen der Mathematiknote und numerischer Intelligenz. Eine Ursache für die niedrigen Korrelationen kann man in der eingeschränkten Messwertvarianz der durch die Numerus-clausus-Regelung vorselegierten Stichprobe finden. In der Gesamtstichprobe fielen demgemäß die Interkorrelationen auch höher aus (s. Anhang N). Wie Tabelle 81 zeigt, sind die Standardabweichungen sowohl der Abiturnoten und der Vordiplomnote als auch der Testleistungen jeweils gering.

Tabelle 81: Deskriptive Statistiken der Abitur- und z-standardisierten Testleistungen in der Hauptstudiumsstichprobe

Abiturnoten und Tests	***M***	***SD***	***N***
Vordiplomnote	1.78	0.62	74
Abiturdurchschnitt	1.69	0.54	74
Mathematiknote	2.05	1.34	63
Deutschnote	1.84	0.75	61
Englischnote	1.79	0.92	56
verbale Intelligenz	0.39	0.82	74
numerische Intelligenz	0.18	0.99	74
Matrizen	-0.03	1.13	74
Intelligenzgesamtscore	0.31	0.96	74
SPARK	0.35	0.99	74
Kreativität	0.28	0.82	74

Die jeweils vergleichsweise niedrige Personenseparationsreliabilität der Skalen stellt eine weitere Ursache für die niedrigen Abiturnoten-Testwert-Korrelationen dar. Es ist jedoch darüber hinaus wahrscheinlich, dass diese Substichprobe von allen getesteten Probanden die größte Testerfahrung im Sinne einer Testwiseness aufweist. Im Vorgriff auf die Analysen an einer Erstsemesterstichprobe in Psychologie (s. Kap. 17.2.4) fallen nämlich die dort berichteten konvergenten Kriteriumskorrelationen zwischen Testergebnissen und Abiturnoten höher aus. Die Erfahrung zum einen aus dem Grundstudium durch das Ableisten von Versuchspersonenstunden in psychologischen Untersuchungen, zum anderen die aus dem Hauptstudium gewonnene mit zahlreichen Lehrveranstaltungen und nicht zuletzt auch das Wissen aus Praktika, macht einen derartigen Einfluss wahrscheinlich. An dieser Stelle wird überdies bereits deutlich, dass sich die Effekte von Testwiseness jenseits von Mittelwertsdifferenzen wie sie in Kapitel 17.2.3.1 analysiert werden, manifestieren können, insbesondere dann, wenn die Konstruktvalidität bzw. wie hier die konvergente Validität davon betroffen sein kann.

17.2.2 Kriteriumsvaliditäten und Regressionsanalysen

Um einen ersten Überblick über die Beziehungen der Abiturnoten und der Testleistungen zur *Vordiplomdurchschnittsnote* zu erhalten, wurden entsprechende bivariate Korrelationen berechnet. (Für eine Übersicht der Abiturnoten- und Testkorrelationen mit allen Vordiplomnoten s. Anhang O). Zur Verringerung der Wahrscheinlichkeit einer Fehlsignifikanz wegen der Alphafehlerinflation bei multipler Signifikanztestung diente die sequenzielle Testprozedur nach Holm (1979). Die Ergebnisse zeigt Tabelle 82.

Tabelle 82: Bivariate Korrelationen der Schulnoten
und Testleistungen mit der Vordiplomnote (Fortsetzung der Tabelle auf folgender Seite)

Leistungs-variable	**Korrelation mit Vordiplomnote**	***N***
Abitur-durchschnitt	.54**	74
Mathematiknote	.24	62
Deutschnote	.39**	60
Englischnote	.39**	55
Verbale Intelligenz	-.17	74
Numerische Intelligenz	-.26*	74
Matrizen	-.40**	74

Leistungs-variable	**Korrelation mit Vordiplomnote**	***N***
SPARK	-.31*	74
Kreativität	-.10	74
Intelligenz-gesamtscore	-.39**	74

Anmerkung. Sequenzielle Signifikanztestung nach Holm (1979).
*: $p < .05$
*: $p < .01$

Übereinstimmend mit anderen Studien (Gold & Souvignier, 2005; Höppel & Moser, 1993; Köller & Baumert, 2002; Schuler, Funke & Baron-Boldt, 1990) zeigt die Abiturdurchschnittsnote die beste *prädiktive* Validität. Sie liegt mit $r = .54$ sogar noch über der von Schuler et al. (1990) metaanalytisch ermittelten mittleren Validität von $r = .45$ für akademische Prüfungen. Allerdings muss die vergleichsweise geringe Stichprobengröße bei der Korrelationsschätzung bedacht werden, eine Über- oder Unterschätzung der Populationskorrelation ist daher sehr wahrscheinlich. Der Einbezug von 17 Probanden in die Stichprobe, die sich zum Zeitpunkt der Erhebung im dritten Semester befanden und die beim Verfassen dieser Arbeit das Vordiplom abgelegt hatten (mit einer daraus resultierenden Stichprobengröße von $n = 91$), ergab eine etwas niedrigere prädiktive Validität der Abiturdurchschnittsnote von $r = .50$ ($p < .01$). Generell ist daher bei der Interpretation von allen hier berichteten Validitäten stets der durch die relativ geringe Stichprobengröße höhere Stichprobenfehler zu beachten. Daher fällt auch das weiterhin berechnete Konfidenzintervall zur Sicherheistwahrscheinlichkeit von 95% für die Kriteriumskorrelation der Abiturdurchschnittsnote bei $n = 91$ mit einem Bereich von .33 bis .64 breit aus.

Gleich valide zeigen sich die Deutsch- und Englischnote und der Matrizentest, wobei hier zu unterscheiden ist, dass die Validität dieser Einzelnoten prädiktiv, die des Matrizentests hingegen retrospektiv ist. Da der prädiktiven Validität gegenüber der retrospektiven ein höherer Stellenwert eingeräumt werden muss, ist die Aussagekraft der Einzelnoten hierbei verlässlicher. Das Ergebnis unterstützt zudem Forschungsbefunde, wonach Einzelnoten aus den Abiturleistungen keine bessere Prädiktion als die Abiturdurchschnittsnote leisten (Lissmann, 1977; Steyer, Yousfi & Würfel, 2005). Die retrospektive Validität des Intelligenzgesamtscores liegt mit $r = -.39$ über dem von Trost und Bickel (1979) in einem Literaturüberblick berichteten Median der Validität von Intelligenztests mit (umgepolten) Studiennoten von $r = .22$. Der Befund, dass die Mathematiknote im Vergleich zum Test der numerischen Intelligenz keine Signifikanz erzielte, ist durch die hier geringere

Stichprobengröße zu erklären. Die post-hoc berechnete Teststärke lag für die Mathematiknote bei 1 - β = .61, für den Test zur numerischen Intelligenz hingegen bei 1 - β = .74.

Die Betrachtung der einzelnen bivariaten Kriteriumskorrelationen birgt jedoch die Gefahr einer „Überinterpretation“ der einzelnen Koeffizienten, weil bestehende Kovarianzanteile innerhalb der Abiturnoten wie auch innerhalb der Testergebnisse unberücksichtigt bleiben. So bestehen beispielsweise wegen des stark sprachlich ausgerichteten schulischen Unterrichtes gerade innerhalb der Abiturnoten neben fachspezifischen Varianzanteilen weiterhin sprachliche Kovarianzanteile, welche die Kriteriumskorrelationen in unbekanntem Ausmaß beeinflussen. Und auch seitens der Testergebnisse ergibt sich wegen der Gleichheit des Antwortformates eine ähnliche Kovarianzquelle. Die eigentlich interessierende Interpretation fach- bzw. skalenspezifischer Kriteriumsvaliditäten ist dadurch insgesamt deutlich erschwert. Eine gleiche Einschränkung betrifft auch die Seite der Kriterien, da die Vordiplomprüfungen überwiegend mündlich abgehalten werden; rein sprachliche Kovarianzanteile erschweren also auch hier die Erklärung fachspezifischer Leistungsvarianz. Eine annähernde Lösung dieses Problems bestehender Methodenkovarianz bietet eine *simultane* Analyse aus einer Kombination aller Prädiktoren sowie einer aus allen Kriterienmaßen über eine kanonische Korrelationsanalyse. Zum einen liegt ihr Vorteil darin, dass auch bei der Berechnung der Kriteriumsvalidität auch die Prädiktorinterkorrelationen beachtet werden. Zum anderen bietet sie die Möglichkeit, diejenige *Kombination* aus Abitur- und Testleistungsmaßen zu identifizieren, welche maximal mit einer aus den Kriterienmaßen assoziiert ist, wobei auch mehrere solcher Paare der sog. kanonischen Variablen resultieren können. Die Beurteilung der Prädiktoren kann somit, ganz im Gegensatz zur Betrachtung der bivariaten Korrelationen, *als Ganzes* vorgenommen werden.

Zur Durchführung der nachfolgend berichteten kanonischen Korrelationsanalyse dienten die Abiturdurchschnittsnote, die Abitureinzelnoten und die Testleistungen als Prädiktorsatz, die Vordiplomnoten als Kriteriumssatz. Der Intelligenzgesamtwert wurde nicht in den Prädiktorsatz aufgenommen, da er sich als Summenwert direkt aus den Ergebnissen der Intelligenzsubtests ergibt und somit zur Multikollinearität und Suppressoreffekten geführt hätte (Tabachnick & Fidell, 2001, S. 181). Die Gefahr der Multikollinearität besteht ebenso für die Aufnahme der Abiturdurchschnittsnote zu den Abitureinzelfachnoten. Zwar gehen in die Abiturdurchschnittsnote nicht nur die Leistungen aus den drei miterhobenen Einzelfachnoten ein, sodass sich hier keine vollständige lineare Abhängigkeit ergibt, allerdings fiel das multiple *R* bereits zwischen den Fachnoten und dem Abiturduchschnitt in dieser Stichprobe mit *R* = .76

hoch aus, sodass hier eine hohe lineare Abhängigkeit vorhanden war. Daher wurden für die Aufnahme aller Schulleistungen in die kanonische Korrelationsanalyse die *Residuen* aus den jeweiligen Schätzungen der Einzelfachnoten aus der Abiturdurchschnittsnote als Prädiktoren verwendet.

In der nachfolgend durchgeführten kanonischen Korrelationsanalyse erwies sich lediglich die erste kanonische Korrelation mit $r_c = .96$ als signifikant ($\chi^2_{(63,\, n = 55)} = 172$, $p < .001$). (Zweite kanonische Korrelation: $r_c = .70$ ($\chi^2_{(48,\, n = 55)} = 62$, $p > .05$)). Der Stewart-Love Index verwies mit einem Wert von .83 auf einen sehr hohen Anteil gemeinsamer Varianz zwischen dem Prädiktor- und Kriteriumsvariablensatz. Zur inhaltlichen Interpretation der kanonischen Variable können nach Empfehlungen von Tabachnick und Fidell (2001, S. 185) Ladungen größer oder gleich .30 herangezogen werden. Levine (1977, S. 18-19) weist allerdings darauf hin, dass beim Vorliegen von Multikollinearität die Interpretation der Ladungskoeffizienten zu falschen Interpretationen führen kann. Suppressoreffekte können in diesem Falle dazu führen, dass ein Ladungskoeffizient ein anderes Vorzeichen als der Strukturkoeffizient aufweist. Strukturkoeffizienten sind in diesem Fall besser zur Interpretation geeignet, weil sie das relative Gewicht einer Variable für die jeweilige kanonische Variable darstellen, das um den Einfluss der übrigen Variablen bereinigt wurde.

Zur Interpretation und zum Identifizieren möglicher Suppressoreffekte werden im Folgenden sowohl Ladungen als auch Strukturkoeffizienten berichtet und miteinander verglichen.

Tabelle 83 zeigt die Ergebnisse der kanonischen Korrelation.

Tabelle 83: Prädiktor- und Kriteriumsstrukturkoeffizienten und -ladungen auf der kanonischen Variablen in der Hauptstudiumsstichprobe ($N = 55$) (Fortsetzung der Tabelle auf folgender Seite)

Variablensatz	**Variable**	**Ladung**	**Struktukoeffizient**
Prädiktoren	Abiturdurchschnittsnote	.99	.99
	Mathematiknote	-.30	-.08
	Deutschnote	.05	-.05
	Englischnote	-.10	.05
	Verbale Intelligenz	.29	-.01
	Numerische Intelligenz	.13	.06
	Matrizen	-.11	-.01
	SPARK	.08	-.09
	Kreativität	.38	-.001

Variablensatz	Variable	Ladung	Struktukoeffizient
Kriterien	Allgemeine Psychologie I	.94	.32
	Allgemeine Psychologie II	.94	.08
	Biopsychologie	.96	.56
	Differenzielle Psychologie	.91	-.36
	Methodenlehre	.91	.10
	Entwicklungspsychologie	.90	.05
	Sozialpsychologie	.94	.29

Aufseiten der Prädiktoren dominiert die Abiturdurchschnittsnote mit einer extrem hohen Ladung. Die Betrachtung der übrigen Ladungskoeffizienten gibt Aufschluss darüber, weshalb sie derart hoch ausfällt. Zunächst lässt sich feststellen, dass entgegen der Befunde zu den bivariaten Korrelationen nach Tabelle 82 besonders die Kreativitätsskala und diejenige zur verbalen Intelligenz nun retrospektive Validitätsbeiträge aufweisen und zwar mit umgekehrtem Vorzeichen, verglichen mit ihren Strukturkoeffizienten. Nach Thompson (1995) weisen Variablen mit niedrigen Stukturkoeffizienten, aber mit vergleichsweise hohen positiven Ladungen darauf hin, dass es sich um Suppressorvariablen handelt. Im vorliegenden Fall ist es sehr wahrscheinlich, dass beide Skalen mit ihren großen Anteilen verbaler Fähigkeiten als Suppressor für die Abiturdurchschnittsnote fungieren und somit deren prädiktive Kraft erhöhen. Die Vorzeichenumkehrung ist hierbei ein Effekt der sog. „Net-Suppression“ (Cohen, Cohen, West & Aiken, 2002), bei der ein Prädiktor X_1 (Abiturnote) eine höhere Korrelation mit dem Kriterium Y (Vordiplomnote) aufweist als ein Prädiktor X_2 (Kreativitätsskala), welcher stärker mit X_1 korreliert. In der Folge unterdrückt X_2 Fehlervarianz in X_1, leistet aber für Y nur eine geringe Vorhersage. Im vorliegenden Fall ist die „Net-Suppression“ wahrscheinlich, da in dieser Stichprobe die Abiturdurchschnittsnote und die Kreativitätsskala zu $r = .40$ ($p < .05$) korrelierten, die Kriteriumskorrelation der Kreativitätsskala hingegen nach Tabelle 82 lediglich $r = .10$ ($p > .05$) betrug. Wegen der offensichtlichen Suppressoreffekte ist daher die Interpretation der Prädiktorbeiträge anhand der Strukturkoeffizienten im Sinne Levines (1977, S. 18-19) verlässlicher.

Wie ebenfalls ersichtlich fallen nun die prädiktiven Beiträge der um den gemeinsamen Kovarianzanteil mit der Abiturnote bereinigten Abitureinzelfachnoten im Vergleich zu ihren bivariaten Korrelationen (s.Tabelle 82) drastisch geringer aus. In den bivariaten Korrelationen flossen die gemeinsamen Varianzanteile mit der Abiturdurchschnittsnote, etwa sprachliche,

vollkommen mit ein und erschwerten die Interpretation *fachspezifischer* Prädiktionsanteile erheblich. Die Befunde der kanonischen Korrelation hingegen verweisen einmal mehr darauf, dass Einzelfachnoten keine bessere Vorhersagegenauigkeit aufweisen als die Abiturdurchschnittsnote (vgl. auch Kap.4.1.2). Allerdings ist auch bei den Einzelfachnoten für Englisch und Deutsch wiederum eine Vorzeichenumkehrung zu beobachten. Dies kann hypothetisch damit erklärt werden, dass, wie bereits im Kap. 17.1.1 erläutert, die vor der kanonischen Korrelationsanalyse vorgenommene Auspartialisierung der Anteile der Abiturdurchschnittsnote in diesen Fachnoten wegen der nicht perfekten Reliabilität der Variablen nur unvollständig gelingt (Stelzl, 2005, S. 277). Als Resultat beinhalten die Einzelfachnoten-Residuen weiterhin Anteile an kriteriumsirrelvanter Varianz der Abiturdurchschnittsnote und wirken als Suppressorvariablen.
Die übrigen bisher nicht diskutierten Ladungen der Tests zur numerischen und fluiden Intelligenz sowie von SPARK ergeben keinerlei prädiktiven Nutzen. Offenbar wird bereits durch die Abiturdurchschnittsnote der Großteil kriteriumsrelevanter Leistungsvarianz gebunden.
Insgesamt betrachtet kann daher die Prädiktorseite der kanonischen Variablen wegen der hohen Ladung der Abiturdurchschnittsnote als „Allgemeine schulische Leistungsfähigkeit“ bezeichnet werden.

Auf der Kriterienseite weisen alle Vordiplomnoten sehr hohe Ladungen auf, was die Interpretation einer allgemeinen akademisch-psychologischen Leistungsfähigkeit nahelegt. Auffallend auch hier sind die großen Differenzen zwischen den Ladungen und Strukturkoeffizienten, besonders bei Allgemeiner Psychologie II, Methodenlehre und Entwicklungspsychologie. Offensichtlich kommt es auch hier zu ausgeprägten Suppressoreffekten. Im Falle der Prüfungsleistung im Fach Differenzielle Psychologie kommt es sogar zu einer Vorzeichenumkehrung. Die Prüfungsleistung in Differenzieller Psychologie agiert hierbei offenbar als „Net-Suppressor“ für eine oder mehrere mündliche Prüfungen in Bezug auf die Vordiplomnoten. Dies würde allerdings voraussetzen, dass gerade in die Prüfungsleistung der Differentiellen Psychologie besonders hohe Verbalanteile im Vergleich zu anderen mündlichen Prüfungen einfließen. Und in der Tat korreliert die Deutschnote in dieser Stichprobe mit $r = .51$ ($p < .01$) am höchsten mit der Note in Differenzieller Psychologie (s. Anhang O).

Die kanonische Korrelation führt insgesamt zu den folgenden Schlussfolgerungen. Aufseiten der Prädiktoren dominiert die Abiturdurchschnittsnote als Maß allgemeiner schulischer Befähigung und bestimmt somit den Zusammenhang der Prädiktor- mit der Kriterienseite

maßgeblich. Die Abitureinzelfachnoten ergeben entgegen den Befunden zu bivariaten Korrelationen in der simultanen Analyse nur geringe Zusammenhänge mit den Kriterienmaßen. Gleiches gilt für die Leistungstests. Zudem zeigen sich Suppressoreffekte durch verbal akzentuierte Leistungstests, sodass die Ladung der Abiturdurchschnittsnote eine überschätzt wird. Auch die kanonische Korrelation weist somit die Abiturdurchschnittsnote als besten Einzelprädiktor aus.
Auf der Ebene der Kriterien bilden die Prüfungsresultate eine gemeinsame Dimension allgemeiner psychologisch-akademischer Leistungsfähigkeit ab. Allerdings zeigt die kanonische Korrelation im Gegensatz zur Analyse der bivariaten Korrelationen, dass auch hier Fehlervarianzeinflüsse vorhanden sind, welche durch Suppressoreffekte auspartialisiert werden und daher zu Überschätzungen der Ladungskoeffizienten führen.

Die kanonische Korrelation beantwortet allerdings nicht die zentrale Frage, ob durch einen oder mehrere Subtests gegenüber der Abiturdurchschnittsnote eine inkrementelle Validität in Bezug auf die Vordiplomdurchschnittsnote zu erzielen ist. Dieser Frage wurde daher mit einer multiplen Regression nachgegangen. Hierbei gingen die Abiturdurchschnittsnote und die Personenparameter in verbaler, numerischer und fluider Intelligenz (Subtest Matrizen) und empiriebezogenem Denken sowie die Rohwerte passiver Argumentationskompetenz (SPARK) als Prädiktoren der Vordiplomnote ein. Tabelle 84 gibt einen Überblick über den Gesamtzusammenhang dieser Prädiktoren und den durch sie aufgeklärten Varianzanteil.

Tabelle 84: Modellzusammenfassung des Regressionsmodells Abiturdurchschnittsnote und Testleistungen ($N = 74$).

R	***R*²**	**Korrigiertes *R*²**	**Standardfehler des Schätzers**	***F***	***df***	***p***
.65	.42	.36	.49	8.03	6	.00

Insgesamt können durch die Schätzgleichung 42% der Varianz der Vordiplomnoten erklärt werden ($p < .01$). Allerdings wird der multiple Determinationskoeffizient in der Population bei Schätzung aus Stichprobendaten insbesondere bei geringen Fallzahlen mit steigender Anzahl von Prädiktoren überschätzt (Scheiblechner, 2002). Daher ist zusätzlich das an der vergleichsweise hohen Anzahl an Prädiktoren relativierte schrumpfungskorrigierte R^2 ausgewiesen („Korrigiertes R^2“). Hierdurch resultiert ein korrigierter aufgeklärter

Varianzanteil von 36%. In einem weiteren Schritt wurde geprüft, welche Einzelprädiktoren in signifikanter Weise einen Beitrag zur Vorhersage leisten. Die Ergebnisse hierzu gibt Tabelle 85 wieder.

Tabelle 85: Regressionskoeffizienten des Regressionsmodells Abiturdurchschnittsnote und Testleistungen ($N = 74$).

	Unstandardisierte Koeffizienten		**Standardisierte Koeffizienten**			**95%-Konfidenzintervall für *B***	
Prädiktor	*B*	Standardfehler	Beta	***T***	***p***	Untergrenze	Obergrenze
(Konstante)	.97	.24		3.92	.00	.48	1.47
Abiturdurchschnitt	.52	.13	.41	3.91	.00	.25	.79
Verbale Intelligenz	-.01	.10	-.01	-.15	.87	-.22	.19
Numerische Intelligenz	-.03	.06	-.05	-.53	.59	-.15	.09
Matrizen	-.16	.05	-.30	-2.96	.00	-.27	-.05
SPARK	-.06	.03	-.17	-1.68	.09	-.14	.01
Kreativität	-.02	.12	-.01	-.18	.85	-.26	.21

Wie bereits bei der Betrachtung der bivariaten Korrelationen (s. Tabelle 82) erweist sich die Abiturdurchschnittsnote mit einem standardisierten Beta-Gewicht von .41 ($p <. 01$) als relativ bester Prädiktor. Das breite Konfidenzintervall für *B* weist jedoch auf eine ungenaue Schätzung des Populationswertes hin.
Von den übrigen Prädiktoren zeigt einzig der Marizentest zur Erfassung fluider Intelligenz mit einem standardisierten Beta-Gewicht von -.30 ($p < .01$) einen weiteren Schätzbeitrag. Die statistische Unsicherheit bei der Schätzung des Populationskennwertes fällt allerdings auch hier groß aus, wie man am Konfidenzintervalls für das unstandardisierte Gewicht erkennt.

In einer weiteren Analyse wurde über hierarchische Regressionen die inkrementelle Validität des Matrizentests bestimmt. In einer ersten Regressionsanalyse ging zunächst nur die Abitur-durchschnittsnote als Prädiktor ein, in einer weiteren wurden die Personenparameter des Matrizentests zusätzlich aufgenommen. Die Ergebnisse der hierarchischen Regressions-analysen zeigt Tabelle 86.

Tabelle 86: Modellzusammenfassung der hierarchischen Regressionen in der Hauptstudiumsstichprobe zur Schätzung der Vordiplomnote ($N = 74$).

Modell	**Beta**	***R***	***R²***	**Korrigiertes *R²***	**Standardfehler**	**Änderungsstatistiken**				
						ΔR²	*ΔF*	*df1*	*df2*	*Δp*
Abiturnote	.54	.54	.29	.28	.53	.29	29.89	1	71	.00
Abiturnote + Matrizentestleistung	.48 -.31	.62	.39	.37	.49	.09	11.38	1	70	.00

Wie schon in Tabelle 82 ausgewiesen, zeigt die Abiturnote alleine eine prädiktive Validität von $r = .54$ bzw. klärt 29% (unkorrigiertes R^2) bzw. 28% (korrigiertes R^2) der Varianz in den Vordiplomnoten auf. Durch die Hinzunahme der Leistungen im Matrizentest können inkrementell 9% (korrigiertes R^2) an zusätzlicher Varianzaufklärung in den Vordiplomnoten gewonnen werden.

17.2.2.1 Kreuzvalidierungsanalyse der Regressionsgleichungen

Die in den vorigen Analysen ermittelte multiple Korrelation ist nur bedingt dazu geeignet, den Populationsparameter zu schätzen, da sie die Beta-Gewichte so bestimmt, dass die Korrelation in der Stichprobe maximiert wird. Die hierüber bestimmte Regressionsgleichung kann daher zu stark an die Gegebenheiten der Stichprobe angepasst sein und lässt sich gegebenenfalls nicht auf die Population generalisieren (Tabachnick & Fidell, 2001, S. 135). Für eine Überprüfung der Stabilität der Regressionsvorhersage ist daher eine Kreuzvalidierung nötig. Im vorliegenden Fall wurde daher die Stichprobe per Zufall in etwa zwei gleich große Stichproben unterteilt ($n_A = 36$, $n_B = 38$) und die in der Gesamtstichprobe ermittelte Regressionsgleichung, bestehend aus den Prädiktoren Abiturdurchschnittsnote und Matrizentest, dazu verwendet, die Vordiplomnote in beiden Zufallsstichproben vorherzusagen. Eine große Diskrepanz von R^2 zwischen den kleineren Zufallsstichproben und der Gesamtstichprobe würde auf eine stichprobenbedingte Überanpassung der Ergebnisse verweisen.

Die Ergebnisse dieser Analyse zeigt Tabelle 87.

Tabelle 87: Ergebnisse der Kreuzvalidierungsanalysen für das Regressionsmodell Abiturdurchschnittsnote und Matrizentest zur Schätzung der Vordiplomnote

Regressionsmodell	**R^2 Gesamtstichprobe (N = 74)**	**R^2 Stichprobe A (n = 36)**	**R^2 Stichprobe B (n = 38)**	**95%-Konfidenzintervall für R^2 anhand der Gesamtstichprobe**	
				Untergrenze	Obergrenze
Abiturnote + Matrizentest	.39**	.50**	.20**	.20	.55

Der multiple aufgeklärte Varianzanteil fällt im Vergleich zur Gesamtstichprobe stark unterschiedlich aus. Für Stichprobe A resultiert eine höhere Varianzaufklärung durch die an der Gesamtstichprobe ermittelte Regressionsgleichung, für Stichprobe B hingegen eine deutlich geringere. Diese Ergebnisse korrespondieren sehr gut mit den Konfidenzintervallschätzungen für R^2 nach Cox und Hinckley (1974, S. 213) anhand der Gesamtstichprobenergebnisse zur Sicherheitswahrscheinlichkeit von 95%. Die Parameterschätzungen der Zufallsstichproben liegen nahe an den äußeren Grenzen dieses Konfidenzintervalls. Die Verallgemeinerung der Ergebnisse aus der Gesamtstichprobe ist daher nur bedingt möglich, da die Schätzung instabil bzw. stark an die Stichprobengegebenheiten angepasst ausfällt.

17.2.3 Zusammenfassung der Analysen zu retrospektiven Validitäten

Die bivariaten Korrelationen zeigten übereinstimmend mit bisherigen Studien insbesondere für die Abiturdurchschnittsnote eine gute prädiktive Validität, die leicht über der metaanalytisch von Schuler et al. (1990) ermittelten liegt. Die Testverfahren standen, außer im Falle verbaler Intelligenz und der Kreativität, in substanziellen retrospektiven Zusammenhängen mit der Vordiplomdurchschnittsnote. Eine kanonische Korrelationsanalyse zeigte eine sehr starke Dominanz der Abiturdurchschnittsnote im Prädiktoren-Kriterienzusammenhang.
In Regressionsanalysen zur Bestimmung inkrementeller Validität erzielte lediglich der Test zur Messung der fluiden Intelligenz eine inkrementelle Validität von 9% zusätzlicher Varianzaufklärung. Über ihn scheint eine Verbesserung der Schätzung der Vordiplomdurchschnittsleistung möglich zu sein. Die relativ geringe Präzision der Parameterschätzung angesichts breiter Konfidenzintervalle sowohl der unstandardisierten *B*-Gewichte als auch von R^2 und

anhand der Befunde aus der Kreuzvaliderungen beschränkt aber die Aussagekraft der Befunde stark.

Problematisch ist zudem der statistisch simultane Vergleich einer dem Studienerfolg zeitlich vorgeschalteten Variable (Abiturnote) mit einer zu dessen *retrospektiver* Schätzung (Testergebnisse). Gerade vor dem Hintergrund einer größeren Vertrautheit bzw. Testwiseness von Probanden aus dem Hauptstudium mit psychologischen Leistungstests können daraus resultierende Effekte im Hinblick auf Verzerrungen inkrementeller Validitäten der Testverfahren in Richtung ihrer Über- als auch Unterschätzung nicht ausgeschlossen werden. Diese lassen sich nämlich (wie im anschließenden Kapitel dargestellt) nur partiell über Mittelwertsdifferenzen aufklären, etwa, indem die Testergebnisse der Studierenden aus dem Hauptstudium mit denen des ersten Semesters verglichen werden. Es ist durchaus möglich, dass sich die Mittelwerte *nicht* unterscheiden, obgleich in den verglichenen Stichproben unterschiedliche Lösungsstrategien etwa durch Testerfahrung eingesetzt wurden. Demgemäß würden mehrere Wege (Lösungsstrategien) zum Erfolg (gleiche Mittelwerte und Varianzen) führen. Hieran sehr problematisch wäre, dass in der Gruppe der Hauptstudiumsstudierenden neben der eigentlich interessierenden Variablen auch die Kenntnis von Lösungsstrategien gemessen wurde. In diesem Zusammenhang sind die sehr niedrigen konvergenten Beziehungen der Testverfahren zu den Abiturleistungen der Studienteilnehmer aus dem Hauptstudium Psychologie zu nennen, welche möglicherweise ein Hinweis auf Testwiseness--Effekte in dieser Stichprobe darstellen (vgl. hierzu Tabelle 80). Dass dies Auswirkungen auf die Kriteriumskorrelationen haben kann, liegt auf der Hand, wenn auch einschränkend hinzugefügt sei, dass hierzu „abiturschwächere“ Personen besonders von der Erfahrung mit Testverfahren profitieren sollten. Im Rahmen der allgemeinen statistischen Probleme von Veränderungsmessungen ließe sich dies dadurch erklären, dass leistungsschwächere Personen stärkere Lernzuwachs-Beträge zeigen als leistungsstärkere, was sich in einer negativen Korrelation zwischen Ausgangs- und Differenzwert zeigen müsste (s. hierzu etwa Rost, 2004, S. 276f.). Die psychologische Begründung hierfür ist, dass Probanden mit niedrigeren Ausgangswerten „größere Chancen“ (Rost, 2004, S. 277) haben, etwas dazu zu lernen.

In jedem Fall stellen die niedrigen Test-Abiturleistungsinterkorrelationen eine empirische Anomalie dar, welche die inhaltliche Interpretierbarkeit der Validität der Testverfahren in der Hauptstudiumsstichprobe einschränkt. Es bleibt somit fraglich, zu welchen Anteilen die Testvaliditäten durch die Leistungsvariable oder die konstruktirrelevanten Testkenntnisse beeinflusst sind.

17.2.3.1 Nebenanalyse: Untersuchung von Testwiseness-Effekten

Unter Testwiseness wird die Fähigkeit eines Prüflings verstanden, unabhängig von dem in den Testaufgaben geforderten Wissen durch Anwendung formaler Entscheidungskriterien auf die Struktur der Aufgaben und/oder durch Überlegungen zu den Umständen der Testdurchführung einen hohen Testwert zu erhalten. Die folgenden Analysen dienten dem Zweck, möglicherweise vorhandene Effekte der Testwiseness in der Hauptstudiumsstichprobe auf die Testergebnisse aufzudecken, da sie zu Unterschätzungen in den berichteten Kriteriumsvaliditäten der Testverfahren geführt haben könnten. Darüber hinaus dient die Analyse dem Zweck, die Äquivalenz der Stichprobe aus dem Hauptstudium und derjenigen des ersten Semesters hinsichtlich verschiedener statistischer Parameter zu überprüfen, um zu generalisierbaren Aussagen der Validitätsanalysen zu gelangen.

Um einen möglichen Effekt der Testwiseness abschätzen zu können, müssen die Testwerte der Hauptstudiumsstichprobe mit denen des ersten Semesters in Psychologie verglichen werden. Ein Einfluss von Testerfahrung könnte sich in höheren Mittelwerten der testerfahreneren Hauptstudiumsstichprobe niederschlagen oder bzw. auch in verminderten Testwertvarianzen gegenüber der Erstsemesterstichprobe. Insbesondere letzteres könnte eine Ursache für unterschätzte Kriteriumsvaliditäten der Tests sein. Allerdings entzieht sich die genaue Untersuchung des Effektes von Testerfahrung auf Testwertvarianzen einem Untersuchungsdesign wie dem vorliegenden mit vorgefundenen Gruppen und ohne eine aktive Manipulation von Testantwortstrategien etwa über Testtrainings. Gerade in einem solchen Design ist der Beitrag von Lernerfahrung neben zahlreichen anderen Faktoren auf interdindividuelle Unterschiede nicht abschätzbar. Ebenso kann dem Einfluss von Testwiseness auf Ebene der Itemantworten der zu vergleichenden Gruppen, wie es im Rahmen von Rasch--Modellen über differenzielle Itemfunktionen („Differential Item Functioning“, s. u.a. Holland & Wainer, 1993) geschieht, wegen zu geringer Stichprobengrößen hier nicht nachgegangen werden. Daher fokussiert die folgende Analyse lediglich auf den Vergleich der Varianzen und Mittelwerte zwischen diesen Gruppen als grobe Abschätzung von Testwiseness.

Zunächst wurden die Stichproben hinsichtlich Unterschieden in ihren durchschnittlichen Abiturleistungen verglichen, weil auch hier ggf. vorhandene Unterschiede einen weiteren Einflussfaktor auf die Testergebnisse darstellen können.
Für zwei Probanden aus dem ersten Semester fehlte die Abiturdurchschnittsnote, weshalb sich hier der Stichprobenumfang von $n = 81$ auf $n = 79$ geringfügig reduzierte. Ein t-Test für Stich-

proben mit homogenen Varianzen ergab für die Probanden aus dem ersten Semester eine bessere Abiturdurchschnittsnote ($M_{\text{Abitur Erstsemester}} = 1.52$, $SD_{\text{Abitur Erstsemester}} = 0.50$) gegenüber denjenigen aus dem Hauptstudium ($M_{\text{Abitur Hauptstudium}} = 1.69$, $SD_{\text{Abitur Hauptstudium}} = 0.54$), $t(151) = -2.05$, $p < .05$, wobei die Größe der Differenz mit $d = 0.32$ lediglich gering ausfiel. Die Varianzen der Abiturdurchschnittsnote fielen nach dem Levene-Test nicht unterschiedlich aus ($F_{(1, 151)} = 0.64$, $p > .05$). Hinsichtlich der durchschnittlichen Abiturleistung und deren Varianz lassen sich die Gruppen daher als weitgehend vergleichbar bezeichnen. Effekte durch unterschiedliche schulische Leistungsfähigkeit auf die Testleistungen sind damit weitgehend ausgeschlossen.

Zur Analyse der Testwiseness auf Mittelwertsebene wurde eine multivariate Varianzanalyse nach dem Allgemeinen Linearen Modell durchgeführt mit dem Faktor Gruppenzugehörigkeit (Erstsemester vs. Hauptstudium) und den z-standardisierten Personenparametern der Testleistungen bzw. Rohwerten (bei SPARK) als abhängige Variablen. Der Test zum empiriebezogenen Denken ging nicht in die Analyse ein, da sein Messbereich bei der Hauptstudiumsstichprobe nicht mehr vergleichbar mit demjenigen im ersten Semester sein kann. Bedenkt man die größere Schulung der in ihm enthaltenen Inhalte durch Veranstaltungen im Fach Methodenlehre, so misst dieser Test im Hauptstudium weitaus mehr das methodisch-empiriebezogene *Wissen*, denn ein empiriebezogenes *Denken*. (Zudem ist der Nachweis besserer durchschnittlicher Testleistung der Hauptstudiumsstichprobe bereits in Kapitel 17.2.1.3 erbracht worden). Eine Übersicht über Deskriptivstatistiken und die Ergebnisse des Levene-Tests zur Überprüfung der Varianzhomogenität gibt Tabelle 88.

Tabelle 88: Deskriptivstatistiken und Levene-Test auf Homogenität der Testwertvarianzen der z-standardisierten Testleistungen von Studierenden aus dem ersten Semester ($N = 81$) und dem Hauptstudium ($N = 74$) in Psychologie

Skala	**Stichprobe**	***M***	***SD***	**Levene Statistik**	***df*1**	***df*2**	***p***
Verbale Intelligenz	Erstsemester	0.11	0.95	0.76	1	153	.38
	Hauptstudium	0.39	0.82				
Numerische Intelligenz	Erstsemester	-0.06	1.00	0.00	1	153	.96
	Hauptstudium	0.18	0.99				
Matrizen	Erstsemester	-0.10	0.87	2.03	1	153	.16
	Hauptstudium	-0.03	1.13				
SPARK	Erstsemester	-0.09	0.97	0.04	1	153	.85
	Hauptstudium	0.34	0.98				
Kreativität	Erstsemester	0.23	0.83	0.00	1	153	.99
	Hauptstudium	0.28	0.82				

Rein deskriptiv betrachtet liegen die Mittelwerte der Hauptstudiumsstichprobe durchweg über denjenigen der Erstsemesterstichprobe. Am größten fällt die Differenz für den Subtest SPARK aus, mit einer Effektstärke im mittleren Bereich von $d = 0.44$, am geringsten für die Kreativitätsskala. Die Varianzhomogenitätsannahme muss nur für den Matrizentest verworfen werden, da das Alpha-Niveau zur Verringerung des unbekannten Beta-Fehlers auf 20% gesetzt wurde. Hier ergibt sich für die Hauptstudiumsstichprobe eine größere Testwertvarianz. Mögliche Unterschiede in den Kriteriumskorrelationen dieses Tests in beiden Stichproben aufgrund inhomogener Varianzen sind demnach nicht auszuschließen. Die inferrenzstatistische Untersuchung der Mittelwertsdifferenzen erfolgte mit einer Multivariaten Varianzanalyse nach dem Allgemeinen Linearen Modell. Die Ergebnisse der multivariaten Teststatistik zeigt Tabelle 89.

Tabelle 89: Ergebnisse der multivariaten Teststatistik zum Vergleich mittlerer Testleistungen der Erstsemester- und Hauptstudiumsstichprobe in Psychologie

Teststatistik	**Wert**	***F***	**Hypothese *df***	**Fehler *df***	***p***
Wilks-Lambda	.92	2.39	5.00	149.00	.04

Wilks Lambda wird signifikant, wobei die Gruppenzugehörigkeit 8% der Varianz in den Skalen aufklärt. Die Varianzanalyse zeigte weiterhin nur für Subtest SPARK eine signifikant bessere Durchschnittsleistung der Hauptstudiumsstichprobe (s. Anhang Q). Post-hoc könnte man hierfür eine Erklärung in der längeren Auseinandersetzung der Hauptstudiumsstichprobe mit wissenschaftlicher Literatur finden, wodurch ein Vorteil bei der Beurteilung von Argumenten erwachsen sein kann, zumal der Inhalt der dargestellten Diskussion in SPARK sehr stark auf psychologische Fragestellungen („Ist Intelligenz angeboren?“) ausgerichtet war. Im Falle von SPARK lässt sich hier also nur schwer von einem Vorteil bei der Testbearbeitung im Sinne einer besseren Bearbeitungs*strategie* als Definition von Testwiseness sprechen, sondern von einer verbesserten Textrezeption durch das Hochschulstudium, also vielmehr einer Verbesserung der Fähigkeit an sich.

In Bezug auf die Analyse von Testwiseness-Effekten für die übrigen Skalen lag die post-hoc berechnete Teststärke deutlich unter der unteren Grenze von 80% (s. Anhang Q), sodass hier nicht mit der nötigen Sicherheit von nicht unterschiedlichen Mittelwerten ausgegangen werden kann. Als vorläufige Arbeitshypothese kann jedoch aus diesen Analysen angenommen werden, dass sich bis auf den Subtest SPARK auf Mittelwertsebene keine signifikanten Unterschiede

zwischen Personen des ersten Semesters und solchen aus dem Hauptstudium bestehen, sodass die Vergleichbarkeit der Stichproben auf Ebene der Mittelwerte und Varianzen der übrigen Tests angenommen werden kann.

17.2.4 Prädiktive Validitätsanalysen

Für die Bestimmung der prädiktiven Validität der Abiturnoten und der Testleistungen von Psychologie-Studierenden mit Studienbeginn im Wintersemester 2004/05 standen zum Zeitpunkt dieser Arbeit zum einen Klausurleistungen in einer Orientierungsprüfung (OP-Klausur) zur Verfügung, welche die Probanden zum Ende des zweiten Fachsemesters abgelegt hatten. Zum anderen lagen die Leistungen von zwei Klausuren aus dem Fach Methodenlehre vor, welche in der Mitte und zum Ende des ersten Fachsemesters geschrieben worden waren. Im Folgenden wird zunächst, analog zur Darstellung der Ergebnisse der Hauptstudiumsstichprobe, die Güte der Abiturnoten sowie der Kriteriumsleistungen dargestellt, um im Weiteren die Ergebnisse zur Vorhersage der Leistungen in der OP-Klausur zu berichten. Die Ergebnisdarstellung der Validitätsanalysen im Hinblick auf die OP-Klausur soll an erster Stelle stehen, da ihr ein höherer Kriteriumsrang eingeräumt werden muss als den Methodenklausuren, entscheidet das Bestehen der OP-Klausur über das Fortsetzen des Studiums.

17.2.4.1 Analysen zur Güte der Abiturnoten in der Erstsemesterstichprobe

Einen Überblick über die deskriptiven Kennwerte der Abiturleistungen der Stichprobe von Studienbeginnern zum Wintersemester 2004/05 zeigt Tabelle 90.

Tabelle 90: Deskriptivstatistiken der Abiturnoten in der Erstsemesterstichprobe

Abiturnoten	***M***	***SD***	**Minumum**	**Maximum**	**Schiefe**	***N***
Abiturdurchschnitt	1.52	0.50	1.00	3.30	1.76	79
Mathematik	1.73	0.94	1.00	5.00	1.44	77
Deutsch	1.51	0.57	1.00	3.00	0.58	76
Englisch	1.61	0.79	1.00	4.00	1.00	74

Typisch für die überwiegende Selektion der Probanden anhand der Abiturdurchschnittsnote ist die durchweg positive Schiefe der Abiturnoten und deren geringe Streuung. Gerade die bei der Abiturdurchschnittsnote numerisch größte positive Schiefe und geringste Standardabweichung lässt verringerte Kriteriumszusammenhänge erwarten als beim Vorliegen einer repräsentativen Streuung.
Tabelle 91 gibt eine Übersicht über die Interkorrelationen der Abiturleistungen.

Tabelle 91: Interkorrelationen der Abiturnoten in der Erstsemesterstichprobe

Abiturteilfächer	Abitur-durchschnitt	Mathematiknote	Deutschnote
Mathematiknote	.70** (n = 76)	-	-
Deutschnote	.50** (n = 75)	.15 (n = 75)	-
Englischnote	.71** (n = .74)	.59** (n = 73)	.39** (n = 73)

Anmerkung. Einseitige Signifikanztestung ohne Korrektur für multiples Testen
*: $p < .05$
**: $p < .01$

Die Korrelationen fallen mit Ausnahme des Zusammenhangs zwischen der Deutsch und der Mathematiknote signifikant positiv aus. Wie bereits in der Hauptstudiumsstichprobe (vgl. Tabelle 75) ergibt sich der numerisch höchste Zusammenhang zwischen der Englisch- und der Abiturdurchschnittsnote. Insgesamt zeigen sich mit der Abiturdurchschnittsnote die jeweils höchsten Werte. Die tatsächlichen Korrelationen sind hiebei indess wegen der Eigenkorrelationsanteile der Einzelnoten in der Gesamtnote niedriger anzusetzen.
Cronbachs Alpha für die drei Einzelfachnoten als grobe Reliabilitätsschätzung der Abiturdurchschnittsnote liegt bei $\alpha = .64$ (n = 72). Die Abiturdurchschnittsnote weist somit eine mindestens ausreichende Reliabilität auf.

17.2.4.2 Analysen zur Güte der Kriteriumswerte

Für diese Analysen standen die Daten von n = 66 Psychologie-Studierende mit Studienbeginn im Wintersemester 2004/05 zur Verfügung. Die Reduktion von ursprünglich 81 Probanden kam dadurch zustande, dass zwei Personen ihr Studium abgebrochen hatten und 13 Personen ihre Immatrikulationsnummer bei der Erhebung zu dieser Arbeit fehlerhaft eingetragen hatten,

sodass eine Zuordnung der Klausurergebnisse mit den Testergebnissen nicht mehr möglich war. Die deskriptiven Statistiken der reduzierten Stichprobe zeigt Tabelle 92.

Tabelle 92: Deskriptivstatistiken der Orientierungsprüfungsklausur ($N = 66$)

Klausurteil	***M***	***SD***	**Minimum**	**Maximum**	**Schiefe**
Note Orientierungsprüfung	1.40	0.69	1.00	4.00	-1.52
Punktwert Orientierungsprüfung	32.14	3.64	19.75	36.00	2.05
Note Wahrnehmung	1.42	0.60	1.00	3.30	1.38
Note Lernen	1.28	0.63	1.00	4.00	2.90
Note Gedächtnis	1.71	1.10	1.00	5.00	1.59
Note Denken	1.50	0.91	1.00	5.00	2.43
Note Emotion	1.38	0.71	1.00	4.00	2.53
Note Motivation	1.73	1.05	1.00	5.00	1.82

Sowohl für die Verteilung der Gesamtnote bzw. die des Gesamtpunktwertes als auch aller einzelnen Klausurteile resultiert eine extreme Schiefe. Auch die Standardabweichungen fallen jeweils sehr niedrig aus. Beide Befunde lassen bereits deutliche Validitätsminderungen erwarten. Um die Güte der OP-Klausuren als Kriterium weiter zu analysieren, wurden die Noten aus den Teilfachgebieten miteinander korreliert. Die Ergebnisse hierzu zeigt Tabelle 93.

Tabelle 93: Interkorrelationen der Orientierungsprüfungsklausurnoten und deren Korrelationen mit dem Gesamtpunktwert ($N = 67$)

OP-Klausurteilnote	**Wahrnehmung**	**Lernen**	**Gedächtnis**	**Denken**	**Emotion**	**Motivation**
Wahrnehmung	-					
Lernen	.37**	-				
Gedächtnis	.58**	.37**	-			
Denken	.33**	.35**	.65**	-		
Emotion	.39**	.36**	.56**	.80**	-	
Motivation	.45**	.44**	.59**	.36**	.38**	-
Gesamtpunktwert	-.56**	-.45**	-.69**	-.53**	-.58**	-.61**

Anmerkung. Einseitige Signifikanztestung ohne Korrektur für multiples Testen
*: $p < .05$
**: $p < .01$

Die Interkorrelationen fallen in mittlerer Höhe und durchweg statistisch signifikant aus. Die höchste Korrelation ergibt sich zwischen den Klausurleistungen in Emotions- und Denkpsychologie mit $r = .80$, die niedrigste zwischen denjenigen in Denk- und Wahrnehmungspsychologie. In einem weiteren Schritt wurde jwls. die part-whole-korrigierte Trennschärfe der Einzelnoten berechnet. Die Ergebnisse zeigt Tabelle 94:

Tabelle 94: Part-whole-korrigierte Trennschärfen der OP-Klausurteilnoten ($N = 67$)

OP-Klausurteilnote	**Part-whole-korrigierte Trennschärfe**
Wahrnehmung	.56
Lernen	.48
Gedächtnis	.75
Denken	.66
Emotion	.67
Motivation	.58

Die Diskriminationsfähigkeiten der Einzelnoten liegen durchweg im mittleren bis hohen Bereich. Am stärksten zwischen leistungsschwächeren und leistungsstärkeren Studierenden trennt hier die Note in Gedächtnispsychologie, am schwächsten diejenige in Lernpsychologie. Die über Cronbachs Alpha ermittelte interne Konsistenz der OP-Klausurnoten kann mit $\alpha = .86$ als gut bezeichnet werden. Um die Eindimensionalität im Sinne der Faktorenanalyse zu überprüfen, wurde eine oblique Hauptachsenanalyse (Delta-Wert = 4) durchgeführt. Nach dem Kaiser-Guttman-Kriterium wurde ein Faktor mit einer aufgeklärten Gesamtvarianz von 48% extrahiert. Tabelle 95 zeigt die Faktorladungen.

Tabelle 95: Faktorladungsmatrix der OP-Klausurnoten ($N = 67$)

Teilnoten	**Faktorladung**
Gedächtnis	.84
Denken	.76
Emotion	.75
Motivation	.62
Wahrnehmung	.59
Lernen	.51

Die Faktorladungen stehen in guter Übereinstimmung mit den Part-whole-korrigierten Trennschärfen nach Tabelle 94. Da vier Ladungen größer .60 vorliegen, ist der Faktor nach Guadagnoli und Velicer (1988, zit. nach Bortz, 1999, S. 534) interpretierbar, und die Klausurleistungen können nach den Ergebnissen der Faktorenanalyse als eindimensional angesehen werden. Die Verwendung des OP-Punktwertes als Gesamtmaß zur Bestimmung der kriterienbezogenen Validiät ist somit gerechtfertigt.

17.2.4.3 Beziehungen der Testverfahren mit Abiturnoten in der Erstsemesterstichprobe

Die korrelativen Beziehungen zwischen Schulnoten und Testleistungen finden sich in Tabelle 96.

Tabelle 96: Korrelationen der Abiturnoten mit den Testleistungen ($N_{\text{Abiturdurchschnitt}} = 79$, $N_{\text{Mathematiknote}} = 77$, $N_{\text{Deutschnote}} = 76$, $N_{\text{Englischnote}} = 74$) in der Erstsemesterstichprobe

Abiturnoten	**Verbale Intelligenz**	**Numerische Intelligenz**	**Matrizen**	**Intelligenz-gesamtscore**	**Kreativ.**	**SPARK**	**Emp. Denken**
Abiturdurchschnitt	-.34**	-.22*	-.26**	-.38**	.09	-.04	-.24*
Mathematiknote	-.42**	-.28**	-.35**	-.48**	.08	-.10	-.18
Deutschnote	-.17	.05	-.09	-.07	-.05	.09	-.06
Englischnote	-.34**	-.10	-.31**	-.32**	-.06	-.12	-.19

Anmerkung. Einseitige Signifikanztestung ohne Korrektur für multiples Testen
*: $p < .05$
**: $p < .01$
Kreativ.: Kreativität
Emp. Denken: Empiriebezogenes Denken

Die Korrelation der Abiturdurchschnittsnote mit dem Intelligenztestgesamtwert liegt mit $r = -.38$ etwa in der Höhe, wie sie auch in der Literatur zwischen schulischen und Intelligenzleistungen zu finden ist (vgl. Amelang & Bartussek, 1997, S. 246; Amelang & Zielinski, 1997, S. 209). Insgesamt fallen die Zusammenhänge der Intelligenztests mit der Mathematiknote mit $r = -.48$ am höchsten aus, sogar höher als diejenige zwischen der Abiturdurchschnittsnote und dem Intelligenzgesamtrohwert mit $r = -.34$. Die Deutschnote zeigt hingegen keinerlei signifikante Zusammenhänge mit den Testleistungen. Selbst mit dem stark auf verbale Fähigkeiten abzielende Subtest SPARK korreliert die Deutschnote nicht bedeutsam. Als eine mögliche Erklärung hierfür lässt sich eine sehr geringe interne Konsistenz der Deutschnote vermuten. Zum einen schlägt sich hierbei wahrscheinlicher als bspw. bei Mathematik-

leistungen die geringere Auswertungsobjektivität etwa von Deutschaufsätzen nieder. Zum anderen zeigen darüber hinaus erste Forschungsbefunde zu interessenbezogenen Lernaktivitäten von Schülern, dass sich diese für den Deutschunterricht sehr heterogen darstellen und etwa vom Lernen formaler und grammatikalischer Formen bis hin zur Beschäftigung mit wiederum interindividuell sehr unterschiedlicher Literatur reichen (Birgit Spinath, persönliche Mitteilung vom 2.7.2006). Statistisch betrachtet setzt sich somit die Varianz der Deutschnote aus sehr unterschiedlichen Quellen zusammen, wodurch es über diese Heterogenisierung zu geringeren Korrelationen mit den übrigen Schulleistungen kommt.
Die durchweg höheren Korrelationen der Abiturleistungen mit der Mathematiknote sind zum Teil auf ihre größere Varianz im Vergleich zu den anderen Noten rückführbar, wie aus dem Vergleich der Standardabweichungen aus Tabelle 97 zu erkennen.

Tabelle 97: Überblick über Deskriptivstatistiken der Abiturnoten
und z-standardisierten Testleistungen der Stichprobe des ersten Semesters

Abiturnoten und Testskalen	***M***	***SD***	***N***
Abiturdurchschnitt	1.52	0.50	79
Mathematiknote	1.73	0.94	77
Deutschnote	1.51	0.57	76
Englischnote	1.61	0.79	74
Verbale Intelligenz	0.11	0.95	81
Numerische Intelligenz	-0.06	1.00	81
Matrizen	-0.10	0.87	81
Intelligenzgesamtscore	-0.02	1.01	81
SPARK	-0.08	0.96	81
Kreativität	0.23	0.83	81
Empiriebezogenes Denken	0.15	0.72	81

Eine weitere Erklärung für die höhere Korrelation der Mathematik-Note mit den Testleistungen nach Tabelle 96 kann man in der relativ hohen Anforderung des Fachs Mathematik an die abstrakt-kognitiven Fähigkeiten sehen, während im Fach Deutsch zu einem großen Teil auch Lernleistungen einfließen.

Von den Tests zu spezifischen Fähigkeiten korreliert einzig derjenige zum empiriebezogenen Denken in nennenswerter Weise mit der Abiturdurchschnittsnote. Interessanterweise ergeben sich im Vergleich zur Stichprobe aus dem Hauptstudium generell höhere Korrelationen der Abiturnoten mit den Testleistungen (vgl. Tabelle 80), was allerdings nicht an inhomogenen Testwert- oder Abiturnotenvarianzen der beiden Stichproben liegen kann, da die Varianzhomogenitätsannahme lediglich für den Matrizentest verworfen werden musste (s. hierzu Tabelle 88). Hinsichtlich der Abiturdurchschnittsnote unterschieden sich die beiden Stichproben zwar signifikant, wie in Kapitel 17.2.3.1 berichtet, allerdings lediglich mit geringer Effektstärke von $d = 0.32$. Einen Grund für diese differenzielle konvergente Validität kann man in einer unterschiedlichen Stichprobenzusammensetzung vermuten, weil die Teilnehmer aus dem Hauptstudium lediglich eine Stichprobe *freiwilliger* und somit eine selbstselegierte Gruppe darstellte, wohingegen diejenigen der Analysen dieses Kapitels alle Psychologie--Studierende mit Studienbeginn im Wintersemester 2004/05 umfasste. Selbstselektionsbedingte Minderungen der Korrelationen sind daher zumindest nicht auszuschließen (s. hierzu u.a. Derous & Born, 2005; Marcus & Schütz, 2005). Zudem ist, wie in Kapitel 17.2.3.1 ausgeführt, aufgrund von fehlenden Retestreliabilitäten der hier konstruierten Verfahren noch unklar, wie „lernresistent“ die Tests sind. Im Falle niedriger Retestreliabilitäten muss daher mit starker Verzerrung *jeglicher* Validität etwa durch Testwiseness gerechnet werden, insbesondere der Bestimmung retrospektive Validität.

Die Aussagekraft der Analysen für die Kohorte des ersten Semesters sind also zum einen aufgrund *prädiktiver* Analysen und der in Tabelle 96 mit der Literatur besser übereinstimmenden berichteten Abiturnoten-Intelligenztestkorrelationen verlässlicher. Zum anderen sind sie angesichts einer Kompletterhebung an diesem Semesterjahrgang nicht von Selbstselektionseffekten der Teilnahmebereitschaft betroffen und deshalb repräsentativer. Für die Evaluation des Gesamtverfahrens haben die nachfolgend berichteten Ergebnisse anhand dieser Stichprobe daher eine größere Bedeutung.

17.2.5 Kriteriumsvaliditäten und Regressionsanalysen zur Orientierungsprüfungsklausur

Um einen ersten Überblick über die Kriteriumsbeziehungen der Abitur- und Testleistungen zum Gesamtwert der OP-Klausur zu erhalten, wurden jwls. deren bivariate Korrelationen berechnet. Wegen der ohnehin geringen Varianz der OP-Klausurgesamtleistung (s. Tabelle 92) wurde der Gesamtpunktwert wegen seiner größeren Streuung als Kriterium herangezogen und

nicht die (noch) weniger differenzierende Gesamtnote. Da wegen multipler Signifikanztestungen die Gefahr einer Alphafehlerkumulierung bestand, wurde das Signifikanzniveau nach Holm (1979) adjustiert. Die Ergebnisse der bivariaten Korrelationen der Schulnoten und Testleistungen mit dem OP-Rohwert zeigt Tabelle 98. (Für eine Übersicht der Abiturnoten- und Test-Validitäten mit allen OP-Einzelnoten, s. Anhang R).

Tabelle 98: Kriteriumskorrelationen und jeweilige Stichprobengrößen der Abitur- und Testleistungen in der Erstsemesterstichprobe

Abiturnoten und Testskalen	**Rohwert OP**	***N***
Abiturdurchschnitt	-.32**	66
Mathematiknote	-.16	65
Deutschnote	-.35**	64
Englischnote	-.47**	63
Verbale Intelligenz	.29**	66
Numerische Intelligenz	.15	66
Matrizen	.08	66
Intelligenz-gesamtscore	.21*	66
SPARK	.01	66
Kreativität	.23*	66
Empiriebezogenes Denken	.19	66

Anmerkung. Einseitige multiple Signifikanztestung nach Holm (1979)
*: $p < .05$
**: $p < .01$

Den deskriptiv betrachtet höchsten prädiktiven Beitrag liefert die Englischnote, gefolgt von der Deutschnote und dem Abiturdurchschnitt. Der hohe prädiktive Beitrag der Englischnote ist angesichts bisheriger Befunde zur besseren Prädiktion von Studienerfolg über die Abiturdurchschnittsnote als über Fachnoten (z. B. Schuler et al. 1990, Hell et al. 2005) nur als stichprobenfehlerbedingt erklärbar.
Hinsichtlich der Testleistungen zeigen die Skala verbaler Intelligenz, der Intelligenzgesamtscore und die Kreativitätsskala statistisch bedeutsame Zusammenhänge mit den Kriteriumsleistungen.

Da die bivariaten Korrelationen wegen beispielsweise gemeinsamer sprachlicher Kovarianzanteile keine verlässlichen Schätzungen darstellen, wurde, wie bereits für die Hauptstudiumsstichprobe, eine kanonische Korrelationsanalyse durchgeführt. Die Abiturdurchschnittsnote, die Abitureinzelnoten und die Testleistungen dienten als Prädiktor-, die Vordiplom*noten* als Kriteriumssatz. Der Intelligenzgesamtrohwert wurde wegen seiner direkten linearen Abhängigkeit nicht in die Analyse aufgenommen. Da auch die Abiturdurchschnittsnote eine hohe Abhängigkeit mit den erhobenen Einzelnoten aufwies ($R = .85$), gingen die Residuen der aus der Abiturdurchschnittsnote geschätzten Einzelnoten in die Analyse ein.

Nur die erste kanonische Korrelation mit $r_c = .75$ war signifikant ($\chi^2_{(60,\, n = 62)} = 103$, $p < .001$). (Zweite kanonische Korrelation: $r_c = .62$ ($\chi^2_{(45,\, n = 55)} = 59$, $p > .05$)). Der Stewart-Love Index verwies mit einem Wert von .35 auf einen moderat hohen gemeinsamen Varianzanteil des Prädiktor- und Kriteriumvariablensatzes. Die oben beschriebene geringe Varianz der Schul- und Klausurnoten beschränkt auch hier den Gesamtzusammenhang.

Tabelle 99 gibt einen Überblick über die Ladungen und Strukturkoeffizienten auf der kanonischen Variablen.

Tabelle 99: Prädiktor- und Kriteriumsstrukturkoeffizienten und -ladungen auf der kanonischen Variablen in der Erstsemesterstichprobe ($N = 64$) (Fortsetzung der Tabelle auf folgender Seite)

Variablensatz	**Variablen**	**Ladung**	**Strukturkoeffizient**
Prädiktoren	Abiturdurchschnittsnote	.34	.39
	Mathematiknote	-.09	-.01
	Deutschnote	.46	.61
	Englischnote	.73	.60
	Verbale Intelligenz	-.12	.23
	Numerische Intelligenz	.02	-.10
	Matrizen	-.22	.-18
	SPARK	-.17	-.14
	Kreativität	-.35	-.09
	Empiriebezogenes Denken	-.29	-.24

Variablensatz	Variablen	Ladung	Strukturkoeffizient
Kriterien	Wahrnehmung	-.35	.08
	Lernen	-.12	.46
	Gedächtnis	-.63	-.07
	Denken	-.78	-.19
	Emotion	-.88	-.73
	Motivation	-.60	-.42

Mit Ausnahme der Mathematikleistung dominieren die schulischen Leistungsmaße die Prädiktorseite. Wie schon in Tabelle 98 der bivariaten Korrelationen weisen auch die um den Kovarianzanteil der Abiturdurchschnittsnote bereinigte Englisch- und Deutschnote gegenüber der Abiturdurchschnittsnote jwls. höhere Gewichte auf. Wie dort diskutiert, handelt es sich hierbei wahrscheinlich um ein stichprobenfehlerbedingtes Ergebnis, welches eine Überschätzung des Zusammenhanges darstellt. Für die Testresultate fällt auf, dass die Skala zur verbalen Intelligenz entgegen den Befunden aus Tabelle 98 keinen prädiktiven Wert mehr aufweist. In diesem Zusammenhang auffallend ist das unterschiedliche Vorzeichen des Ladungs- und Strukturkoeffizienten dieser Skala, was nach Thomson (1995) darauf hinweist, dass diese Skala als Suppressor für eine oder mehrere Prädiktoren fungiert. Das Ladungsgewicht des Kreativitätstests, welches eigentlich in interpretierbarer Höhe ausfällt, darf beim Vorliegen von Suppressoreffekten nicht interpretiert werden (Levine, 1977, S. 18-19), sondern der Strukturkoeffizient. Nach diesem ergibt sich unter Auspartialisierung des Einflusses aller übrigen Prädiktoren für diesen Test kein Vorhersagebeitrag.
Insgesamt betrachtet wird somit die Prädiktorseite deutlich von schulischen Leistungsmaßen bestimmt; von den Testskalen weist unter Beachtung des Problems von Suppressoreffekten keine Skala einen interpretierbaren Vorhersagebeitrag auf.

Die Interpretation der Kriterienseite anhand der Ladungskoeffizienten ist schwierig, da alle Noten erwartungswidrig *negativ* laden. Offenbar kommt es hier zu besonders ausgeprägten Multikollinearitäten und Suppressionseffekten, denn, wie Garson (2005) ausführt:
„To the extent that the variables within the dependent sets of variables are highly intercorrelated, the canonical coefficients will be unstable. The coefficients for some variables may be misleadingly low or even negative because variance has already been explained by other variables.” Die Vorzeichen sind demnach *inhaltlich* nicht interpretierbar. Im vorliegenden Fall

betrifft die dies jedoch ebenso einige der Strukturkoeffizienten, was die inhaltliche Gesamtinterpretation der Kriterienseite erheblich erschwert. Festzustellen ist, dass die Noten in Wahrnehmung, Lernen und Gedächtnis als Suppressoren agieren, wie an den Vorzeichenumkehrungen bei „Wahrnehmung“ und „Lernen“ und anhand der großen Differenz zwischen der Ladung und dem Strukturkoeffizienten für „Gedächtnis“ zu sehen. In Übereinstimmung mit den faktorenanalytischen Befunden nach Tabelle 95 korrespondieren die Noten in „Wahrnehmung“ und „Lernen“ am geringsten mit der Kriteriumsdimension, betrachtet man alleine die Höhe der Koeffizienten. Dies ist nun durch ihre Suppressorfunktion statistisch erklärbar, wobei die inhaltlichen Ursachen der Suppression allerdings post-hoc nicht identifizierbar sind. Demgegenüber ist die hohe Ladung von „Gedächtnis“ bei gleichzeitiger Suppressorfunktion über die zuvor beschriebene „Net-Suppression“ (Cohen, Cohen, West & Aiken, 2002) aufklärbar.
Am prägnantesten wird demnach die Kriterienseite durch die Noten in den Klausurteilbereichen „Denken“, „Emotion“ und „Motivation“ repräsentiert, zieht man die Höhe ihrer Ladungen und die Ergebnisse der Faktorenanalyse (vgl. Tabelle 95) zur Interpretation heran.

Zusammenfassend betrachtet wird die Prädiktorseite von den schulischen Leistungsmaßen bestimmt; die Testleistungen zeigen hier keinen substanziellen Beitrag. Im Gegensatz zur Betrachtung der bivariaten Korrelationen zeigt der Test zur verbalen Intelligenz keinen Prädiktionsbeitrag, sondern erweist sich bei *simultaner* Analyse der Variablen als Suppressorvariable. In Anbetracht vorhandener Suppressoreffekte ergibt sich weiterhin auch kein prädiktiver Beitrag des Kreativitätstests, welcher noch bei den bivariaten Korrelationen zu beobachten war. Die Kriterienseite ist von zahlreichen Suppressionseffekten gekennzeichnet, welche die Interpretation erheblich erschweren. Insbesondere die Klausurnoten in den Teilfächern „Wahrnehmung“ und „Lernen“ zeigen starke Suppressoreffekte, die übrigen Leistungen markieren noch am deutlichsten auch in Übereinstimmung mit faktorenanalytischen Befunden die Kriterienseite. Die Befunde verweisen somit einmal mehr auf schulische Leistungen als vergleichsweise beste Prädiktoren akademischer Leistungsmaße.

Da die kanonische Korrelationsanalyse keinen Hinweis auf die inkrementelle Validität von Testverfahren in Bezug auf den OP-Gesamtrohwert liefert, dienten die Abiturdurchschnittsnote und die Testverfahren als Prädiktoren in der im Folgenden berichteten Regressionsanalyse. Die Ergebnisse dieses Gesamtmodells zur Vorhersage der Rohwerte in der OP-Klausur zeigt Tabelle 100.

Tabelle 100: Modellzusammenfassung des Regressionsmodells Abiturdurchschnittsnote mit Testleistungen zur Vorhersage der OP-Klausurpunkte ($N = 66$)

R	***R²***	**Korrigiertes *R²***	**Standardfehler des Schätzers**	***F***	***df***	***p***
.45	.20	.10	3.45	2.07	7	.06

Für das *Gesamt*modell mit allen Prädiktoren in die Regressionsgleichung ergibt sich keine signifikante Varianzaufklärung. Betrachtet man die Parameterschätzungen auf Einzelprädiktorebene in der nachfolgenden Tabelle 101, so ergibt sich lediglich für die Abiturdurchschnittsnote ein signifikanter Beitrag zur prädiktiven Validität.

Tabelle 101: Regressionskoeffizienten des Regressionsmodells Abiturdurchschnittsnote und Testleistungen zur Vorhersage der OP-Rohwerte ($N = 66$)

Modell	**Unstandardisierte Koeffizienten**		**Standardisierte Koeffizienten**			**95%-Konfidenzintervall für *B***	
	B	Standardfehler	*Beta*	*T*	*p*	Untergrenze	Obergrenze
(Konstante)	35.10	1.61		21.73	.00	31.87	38.34
Abiturdurchschnitt	-2.23	1.03	-.28	-2.15	.03	-4.31	-0.15
Verbale Intelligenz	1.23	0.85	.23	1.44	.15	-0.47	2.94
Numerische Intelligenz	-0.07	0.48	-.02	-0.14	.88	-1.03	0.89
Matrizen	-0.57	0.58	-.14	-0.98	.32	-1.74	0.59
Empiriebezogenes Denken	0.20	0.42	.06	0.49	.62	-0.63	1.05
SPARK	-0.23	0.29	-.10	-0.80	.42	-0.83	0.35
Kreativität	1.32	1.00	.17	1.31	.19	-0.69	3.34

Mit der Abiturdurchschnittsnote als einzigem signifikantem Prädiktor können demnach 9% der Varianz im OP-Rohwert aufgeklärt werden (s. das Quadrat der Korrelation nach Tabelle 98). Die Varianzaufklärung ist damit insgesamt niedrig, was hier auch durch die geringen Prädiktor- und Kriteriumsstreuungen und durch die deutliche Schiefe der jeweiligen Verteilung bedingt ist. Auch in dieser Stichprobe erweist sich die Abiturdurchschnittsnote somit als vergleichsweise validester Prädiktor von Studienleistung.

17.2.5.1 Prädiktive Validitäten für Methodenlehreklausuren

Wie bereits zu Eingang dieses Kapitels genannt, lagen für weitere prädiktive Kriteriumsanalysen die Ergebnisse von zwei Klausuren aus dem Fach Methodenlehre vor, von denen die eine etwa zur Mitte des ersten Fachsemesters geschrieben worden war, die andere zu dessen Ende. An der inhaltlichen Zusammenstellung der ersten Klausur waren neben Hr. Prof. Werner noch die Fachtutoren beteiligt, weshalb sie im Folgenden mit TestTutVL abgekürzt werden soll (für „Test Tutorium-Vorlesung"). Sie beinhaltete neben Fragen zur deskriptiven Statistik auch Aufgaben zu Wahrheitstafeln der formalen Logik. Die zweite Klausur (TestVL) behandelte alleine die Inhalte der Methodenlehre-Vorlesung mit Fragen zur Inferenzstatistik, dem Allgemeinen Linearen Modell und zur Versuchsplanung. Der inhaltliche Überschneidungsbereich der Klausuren ist daher vergleichsweise gering. Vor den Analysen muss angemerkt werden, dass nicht alle Psychologie-Studierenden mit Studienbeginn im Wintersemester 2004/05 an den Klausuren teilgenommen hatten, sodass die Stichprobengröße für TestTutVL bei $N = 57$ und für TestVL bei $N = 51$ lag.

17.2.5.1.1 Güte der Kriteriumswerte

Zunächst gibt Tabelle 102 einen Überblick über deskriptive Kennwerte der Klausurpunkteverteilungen.

Tabelle 102: Deskriptive Statistiken der Methodenlehreklausuren

Klausur	***M***	***SD***	**Minimum**	**Maximum**	**Schiefe**	***N***
TestTutVL	6.36	1.32	2.75	8.00	-0.79	57
TestVL	2.19	0.76	0.00	4.00	0.16	51

Der hohe Mittelwert von TestTutVL und die leicht negative Schiefe deuten auf einen Deckeneffekt der Punkteverteilung hin. Auch ist die Streuung der Skala relativ niedrig. Die Verteilung von TestVL weist ebenfalls eine niedrige Standardabweichung auf, die Schiefe hingegen ist lediglich schwach positiv und unauffällig.
Die Korrelation zwischen beiden Klausuren fiel mit $r = .35$ ($p < .01$) moderat aus. Der mittelhohe Zusammenhang ist zum einen auf die niedrige Varianz in beiden Skalen zurückzuführen, zum anderen vermutlich auch auf die inhaltliche Heterogenität, wie eingangs dieses Kapitels

beschrieben. Da dem Autor die Originalantworten der Studierenden in den einzelnen Klausuraufgaben nicht vorlagen, konnte keine getrennte Reliabilitätsschätzung durchgeführt werden. Die Berechnung von Cronbachs Alpha für beide Klausuren ergab einen vor dem Hintergrund der niedrigen Klausuren-Interkorrelation erwartungsgemäß niedrigen Wert von $\alpha = .49$. Bereits die niedrige Klausureninterkorrelation verbietet eine Analyse aggregierter Klausurleistungen, weshalb die folgenden Kriteriumsanalysen getrennt für jede Methodenklausur dargestellt werden.

17.2.5.2 Kriteriumsvaliditäten und Regressionsanalysen zu den Methodenlehreklausuren

Eine erste Übersicht über die prädiktiven Validitäten der Abiturleistungen und der Testverfahren liefert Tabelle 103:

Tabelle 103: Kriteriumskorrelationen und jwl. Stichprobengrößen der Abitur- und Testleistungen für die Rohwerte von TestTutVL und TestVL aus der Erstsemesterstichprobe

Abiturnote/Skala	TestTutVL	TestVL
Abiturdurchschnitt	-.38** (n = 57)	-.47** (n = 51)
Mathematik	-.18 (n = 56)	-.30 (n = 51)
Deutsch	-.21 (n = 56)	-.30 (n = 50)
Verbale Intelligenz	.19 (n = 57)	.52** (n = 51)
Numerische Intelligenz	.21 (n = 57)	.11 (n = 51)
Matrizen	.30 (n = 57)	.30 (n = 51)
Intelligenzgesamtscore	.29 (n = 57)	.38** (n = 51)
SPARK	-.03 (n = 57)	.15 (n = 51)
Kreativität	.34** (n = 57)	.36* (n = 51)
Empiriebezogenes Denken	.23 (n = 57)	.44** (n = 51)

Anmerkung. Einseitige multiple Signifikanztestung nach Holm (1979)
*: $p < .05$
**: $p < .01$

Für TestTutVL ergeben sich nur für die Abiturdurchschnittsnote und die Kreativitätsskala signifikante Korrelationen. Ein deutlich verändertes Bild zeigt sich bei TestVL. Hier erzielt neben der Abiturdurchschnittsnote der Test zur verbalen Intelligenz einen hohen prädiktiven Beitrag. Allerdings erweisen sich die Kriteriumskorrelationen der hierzu umgepolten Abitur-

durchschnittsnote und dem verbalen Intelligenztest nach einem Korrelationsdifferenztest für korrelierte Variablen als nicht signifikant verschieden (z_{Diff} = -0.37, p > .05). Beide zeigen demnach einen gleich hohe Validitäten in Bezug auf TestVL. Neben dem Intelligenz-gesamtscore wird nun auch die Beziehung des Tests zum empiriebezogenen Denken signifikant. Letzterer Befund mag damit zu erklären sein, dass in der TestVL neben statistischen Inhalten auch Fragen zur Versuchsplanung enthalten waren, was zentraler Gegenstand dieses Tests ist.

Im Gesamten betrachtet liegen die Kriteriumsvaliditäten der Abiturleistungen und der Tests in Bezug auf TestVL jeweils höher als für TestTutVL und es ergeben sich mehr signifikante Korrelationen. Die Abiturdurchschnittsnote weist für beide Klausuren prädiktive Validität auf. Die bivariaten Analysen liefern jedoch noch keine Rückschlüsse auf die inkrementelle Validität der Testverfahren gegenüber der Abiturdurchschnittsnote. Zur Untersuchung dieser Fragestellung dienten wiederum multiple Regressionen mit der Abiturdurchschnittsnote und den Testverfahren als Prädiktoren zur Vorhersage der jeweiligen Klausurleistungen.
Die Ergebnisse einer multiplen Regression mit der Abiturdurchschnittsnote und den Testleistungen als Prädiktoren für die Rohwerte von TestTutVL zeigt Tabelle 104.

Tabelle 104: Modellzusammenfassung des Regressionsmodells Abiturdurchschnittsnote mit Testleistungen zur Vorhersage der Methodenklausurleistungen TestTutVL (N = 57)

R	***R²***	**Korrigiertes *R²***	**Standardfehler des Schätzers**	***df***	***F***	***p***
.56	.31	.21	1.17	7	3.22	.00

Durch das Modell können insgesamt 21% der Varianz (korrigiertes R^2) in den Rohwerten der Klausur TestTutVL signifikant aufgeklärt werden. Im Weiteren wurde geprüft, welche Einzel-prädiktoren in signifikanter Weise einen Beitrag zur Vorhersage zu leisten. Die Ergebnisse gibt Tabelle 105 wieder.

Tabelle 105: Regressionskoeffizienten des Regressionsmodells Abiturdurchschnittsnote und Testleistungen zur Vorhersage der Rohwerte von TestTutVL ($N = 57$)

Modell	**Unstandardisierte Koeffizienten**		**Standardisierte Koeffizienten**			**95%-Konfidenzintervall für *B***	
	B	Standardfehler	Beta	*T*	*p*	Untergrenze	Obergrenze
(Konstante)	7.59	0.61		12.35	.00	6.36	8.83
Abiturdurchschnitt	-0.98	0.40	-.33	-2.46	.01	-1.79	-0.18
Verbale Intelligenz	-0.32	0.30	-.17	-1.05	.29	-0.94	0.29
Numerische Intelligenz	0.14	0.17	.10	0.82	.41	-0.20	0.49
Matrizen	0.39	0.21	.27	1.89	.06	-0.02	0.82
SPARK	-0.14	0.10	-.16	-1.33	.18	-0.35	0.07
Empiriebezogenes Denken	0.04	0.17	.03	0.24	.80	-0.30	0.39
Kreativität	0.93	0.41	.28	2.22	.03	0.09	1.77

Die Abiturdurchschnittsnote erweist sich mit einem Beta-Gewicht von -.33 ($p < .05$) als der beste Prädiktor. Die Populationsschätzung ist allerdings angesichts der Breite des Konfidenzintervalls nur grob möglich. Einen zusätzlichen Prädiktionsbeitrag zeigt allein die Kreativitätsskala mit einem Beta-Gewicht von .28 ($p < .05$). Die Genauigkeit der Populationsschätzung fällt nach dem Konfidenzintervall auch hier gering aus.
Der Größe der inkrementellen Validität durch die Kreativitätsskala wurde in einer hierarchischen Regression nachgegangen, deren Ergebnisse Tabelle 106 wiedergibt.

Tabelle 106: Regressionskoeffizienten der hierarchischen Regressionen zur Bestimmung inkrementeller Validität für TestTutVL ($N = 57$)

Modell	**Beta**	***R***	***R*²**	**Korrigiertes *R*²**	**Standardfehler des Schätzers**	**Änderungsstatistiken**				
						ΔR²	*ΔF*	*df1*	*df2*	*Δp*
Abiturnote	-.38	.38	.14	.13	1.23	.14	9.43	1	55	.00
Abiturnote + Kreativität	-.33 .29	.48	.23	.20	1.18	.08	5.90	1	54	.01

Über die Abiturnote alleine können 13% (korrigiertes R^2) aufgeklärt werden. Durch die Hinzunahme des Kreativitätstests wird somit eine signifikante inkrementelle Varianzaufklärung von 7% (Differenz der korrigierten R^2) erreicht.

Die folgenden Analysen geben die Ergebnisse hinsichtlich der Vorhersage der Klausurpunktwerte von TestVL wieder. Zunächst zeigt Tabelle 108 die Ergebnisse des Gesamtmodells.

Tabelle 107: Modellzusammenfassung des Regressionsmodells Abiturdurchschnittsnote mit Leistungen in Methodenklausur TestVL ($N = 51$)

R	***R²***	**Korrigiertes *R²***	**Standardfehler des Schätzers**	***df***	***F***	***p***
.69	.48	.39	0.59	7	5.74	.00

Die Gesamtvarianzaufklärung fällt im Vergleich der Ergebnisse zur Klausur TestTutVL mit 39% (korrigiertes R^2) deutlich höher aus. Welche Prädiktoren hierbei einen bedeutsamen Beitrag leisten, gibt Tabelle 108 wieder.

Tabelle 108: Regressionskoeffizienten des Regressionsmodells Abiturdurchschnittsnote und Testleistungen zur Vorhersage der Rohwerte von TestVL ($N = 51$)

Modell	**Nicht standardisierte Koeffizienten**		**Standardisierte Koeffizienten**			**95%-Konfidenzintervall für *B***	
	B	Standardfehler	Beta	*T*	*p*	Untergrenze	Obergrenze
(Konstante)	2.77	0.33		8.19	.00	2.08	3.45
Abiturdurchschnitt	-0.49	0.22	-.27	-2.22	.03	-0.95	-0.04
Verbale Intelligenz	0.45	0.16	.41	2.70	.01	0.11	0.79
Numerische Intelligenz	-0.12	0.09	-.16	-1.29	.20	-0.31	0.06
Matrizen	0.01	0.11	.01	0.08	.93	-0.21	0.23
SPARK	-0.01	0.06	-.03	-0.26	.79	-0.14	0.11
Empiriebezogenes Denken	0.17	0.09	.23	1.87	.06	-0.01	0.35
Kreativität	0.27	0.22	.15	1.23	.22	-0.17	0.72

Erstaunlicherweise weisen die Beta-Gewichte den Test zur verbalen Intelligenz als einen stärkeren Prädiktor als die Abiturdurchschnittsnote aus. Die Regressionsgewichte aller übrigen Prädiktoren sind hingegen insignifikant. Indes fällt auch in dieser Analyse die Populationsschätzung anhand der Konfidenzintervalle aller Prädiktorgewichte statistisch unsicher aus.

Die Ergebnisse zum Inkrement der Skala verbaler Intelligenz über eine hierarchische Regression zeigt Tabelle 109.

Tabelle 109: Regressionskoeffizienten der hierarchischen Regressionen zur Bestimmung inkrementeller Validität für TestVL ($N = 51$)

Modell	**Beta**	***R***	***R²***	**Korrigiertes *R²***	**Standardfehler des Schätzers**	**Änderungsstatistiken**				
						ΔR^2	ΔF	*df1*	*df2*	Δp
Abitur	-.47	.47	.22	.21	0.68	.22	14.49	1	49	.00
Abitur + verbale Intelligenz	-.34 .41	.61	.37	.35	0.61	.15	11.56	1	48	.00

Für diese Klausur belegt eines der hier eingesetzten Testverfahren das größte Inkrement überhaupt: Klärt die Abiturnote alleine 21% der Klausurrohwertevarianz auf, lässt sich durch Aufnahme der Skala zur verbalen Intelligenz 14% (Differenz der korrigierten R^2) zusätzliche Varianzaufklärung gewinnen.

17.2.5.3 Kreuzvalidierungsanalyse der Regressionsgleichungen

Auf die Gefahr einer Überanpassung regressionsanalytischer Ergebnisse an die untersuchte Stichprobe wurde bereits in Kapitel 17.2.2.1 verwiesen. Analog zu der dort gewählten Strategie einer Kreuzvalidierung der Regressionsgleichung anhand zweier Zufallsstichproben wurden daher auch in dieser Stichprobe zwei Zufallsstichproben von annähernd gleichem Umfang für jede der dargestellten Kriteriumsanalysen gebildet und die anhand der Gesamtstichprobe jeweils ermittelten Regressionsgleichungen zur inkrementellen Validität kreuzvalidiert. Tabelle 110 gibt einen Überblick über die Ergebnisse dieser Analysen.

Tabelle 110: Ergebnisse der Kreuzvalidierungsanalysen für Regressionsgleichungen zur inkrementellen Validität in der Erstsemesterstichprobe

Kriterium	**Regressions-modell**	**R^2 Gesamtstich-probe**	**R^2 Stichprobe A**	**R^2 Stichprobe B**	**95%-Konfidenzintervall für R^2 anhand der Gesamtstichprobe**	
					Untergrenze	Obergrenze
OP-Klausur	Abiturnote	.10* (n = 66)	.14* (n = 36)	.09 (n = 30)	.006	.26
TestTutVL	Abiturnote + Kreativität	.23* (n = 57)	.32** (n = 29)	.17* (n = 28)	.04	.41
TestVL	Abiturnote + verbale Intelligenz	.37** (n = 51)	.31** (n = 26)	.40** (n = 25)	.14	.55

Die jeweiligen multiplen Determinationskoeffizienten für das Regressionsmodell mit der Abiturnote als einzig signifikanten Prädiktor differieren rein numerisch nicht wesentlich, der Gesamtstichprobenbefund ist daher stabil und vergleichsweise leichter generalisierbar. Die Insignifikanz von R^2 in Zufallsstichprobe B ist auf die gegenüber Zufallsstichprobe A deutlich geringere post-hoc berechnete Teststärke rückführbar ($1\text{-}\beta_{\text{Stichprobe A}} = .64$, $1\text{-}\beta_{\text{Stichprobe B}} = .35$). Gleichwohl fällt das Konfidenzintervall für R^2 anhand der Gesamtstichprobe breit aus. Die Schätzung des wahren Determinationskoeffizienten durch die Abiturnote ist daher nur grob möglich.

Größere Unterschiede in den multiplen Determinationskoeffizienten ergeben sich für das Regressionsmodell mit TestTutVL als Kriterium und der Kombination aus Abiturnote und Kreativitäts-Testergebnis. Die Ergebnisse der Zufallsstichproben differieren deutlich, der Effekt erweist sich daher als stark stichprobenabhängig. Auch das breite Konfidenzintervall für R^2 deutet darauf hin. Die Generalisierbarkeit der aus diesem Regressionsmodell abgeleiteten Aussagen über die inkrementelle Validität ist somit stark begrenzt.

Das Regressionsmodell für TestVL zeigt sich demgegenüber als stabiler und die Schätzung anhand der Gesamtstichprobe gelingt verlässlicher, da sich die Koeffizienten deskriptiv betrachtet nicht drastisch unterscheiden. Dessen ungeachtet verweist das breite Vertrauens-intervall auch in dieser Analyse auf die Ungenauigkeit bei der Populationsschätzung von R^2 anhand der Gesamtstichprobe.

17.2.6 Zusammenfassung der Analysen zur prädiktiven Validität

Die Analysen unterstützen insgesamt betrachtet bisherige Befunde, welche die Abiturdurchschnittsnote als den besten Prädiktor in Bezug auf Studiennoten ausweisen (s. u.a. Schuler et al. 1990, Hell et al. 2005). In den einzelnen Klausurleistungen klärte sie bereits in dieser nach Abiturleistung vorselegierten Stichprobe 9% (OP-Klausur), 13% (TestTutVL) und 21% (TestVL) der Varianz in den jeweiligen Leistungsunterschieden auf und erwies sich in allen Regressionsanalysen als signifikanter Prädiktor. Im Zusammenhang der prädiktiven Validität muss man auch die Korrelation der Abiturdurchschnittsnote hinsichtlich der Vordiplomnote nennen. Sie lag, unter Einschluss aller Probanden, die zum Zeitpunkt dieser Arbeit das Vordiplom abgelegt hatten ($N = 91$), bei $r = .50$. Die daraus resultierende Varianzaufklärung von 25% in einer überwiegend nach Abiturdurchschnittsnote vorselegierten Stichprobe ist beachtlich.

Die Leistungstests korrelierten meist signifikant in erwarteter Richtung mit den Klausurleistungen. In der simultanen Analyse aller erhobenen schulischen und leistungstestsbezogenen Prädiktoren in einer kanonischen Korrelationsanalyse bezüglich der OP-Einzelnoten dominierten allerdings die schulischen Leistungsmaße die Korrelation zwischen der Prädiktor- und Kriterienseite.
Für die OP-Klausur war die Abiturdurchschnittsnote der einzig signifikante Prädiktor in einer multiplen Regressionsanalyse. In einer von zwei Methodenklausuren (TestVL) war durch die Skala zur verbalen Intelligenz hingegen ein Inkrement von 14% möglich, wobei ihr Beta--Gewicht hier sogar noch höher als das der Abiturdurchschnittsnote ausfiel. In der Methodenklausur TestTutVL erzielte die Kreativitätsskala eine inkrementelle Validität von 7% zusätzlicher Varianzaufklärung.

Im Weiteren wurde der Stabilität der Parameterschätzungen aus den Regressionsanalysen zur inkrementellen Validität der Testverfahren in Kreuzvalidierungsanalysen nachgegangen.
In einer einfachen Regression war die Abiturdurchschnittsnote ein stabiler Prädiktor für die OP-Klausurergebnisse. Auch die Regression zur Vorhersage der Leistung in der Methodenklausur TestVL, bestehend aus Abiturdurchschnittsnote und der Skala zur verbalen Intelligenz, war in den Kreuzvalidierungen robust. Hingegen differenzierten die Ergebnisse der Kreuzvalidierungen einer Regressionsgleichung mit den Prädiktoren Abiturdurchschnittsnote und dem Kreativitätstest in Bezug auf die Methodenklausurergebnisse TestTutVL stark, was an

der Generalisierbarkeit der Befunde zur inkrementellen Validität durch den Kreativitätstest stark zweifeln lässt.
Generell wiesen die Populationsschätzungen der durch die jeweiligen Prädiktoren aufgeklärten Varianzanteile in den Klausurergebnissen statistische Unsicherheiten auf, wie die breiten Konfidenzintervalle für R^2 anhand der Gesamtstichprobe jeweils auswiesen.

Die Befunde zur inkrementellen Validität der hier verwendeten Testverfahren sind insbesondere nach den Ergebnissen der Kreuzvalidierungsanalysen uneinheitlich. Anhand der Analysen der Gesamtstichprobe erbrachten nur einzelne und jeweils unterschiedliche Subtests überhaupt ein Inkrement über die Abiturdurchschnittsnote hinaus. Verallgemeinerbare Aussagen darüber, zumindest welcher Subtest in verschiedenen Klausuren ein Inkrement erbringen kann, sind aufgrund dieser Unterschiedlichkeit kaum möglich. Betrachtet man die Ergebnisse aus den Regressionsanalysen und deren Kreuzvalidierungen insgesamt, so scheint eine inkrementelle Validität mit den hier verwendeten Verfahren noch am ehesten über verbale Intelligenz möglich zu sein, zumindest dann, wenn man die Ergebnisse der Regressionsanalysen bezüglich der Klausur TestVL und deren stabile Kreuzvalidierungsanalyse heranzieht.

Die Vorhersagebeiträge aller übrigen Tests (mit Ausnahme der Kreativitätstsskala) fielen in der Vorhersage von Klausurergebnissen insignifikant aus. In allen Analysen zu prädiktiven Validitäten muss man allerdings angesichts der durchweg breiten Konfidenzintervalle für alle Parameterschätzungen den großen Stichprobenfehler bei der Interpretation hinzuziehen und die gezogenen Schlüsse als vorläufig betrachten.

Einstweilen kann die Aussage dieser Analysen in Übereinstimmung mit bisheriger Forschung somit nur lauten, dass die Abiturdurchschnittsnote auch in einer nach ihr hoch selegierten Stichprobe den validesten und stabilsten Einzelprädiktor darstellt. Die in dieser Untersuchung ermittelte hohe prädiktive Kraft der Abiturdurchschnittsnote für die Vordiplomnote von $r = .50$ unterstützt diese Interpretation.
Inkrementelle Validitäten können darüber hinaus mit Testverfahren erreicht werden; die Verallgemeinerbarkeit solcher Ergebnisse ist jedoch wegen ihrer Abhängigkeit von den herangezogenen Kriterien und wegen der Stichprobenabhängigkeit jeweils zu überprüfen.

Eine verlässlichere Aussage bezüglich der inkrementellen Validität der Testverfahren gerade auch in Bezug auf mittel- und längerfristige Studienerfolgskriterien lässt sich allerdings erst

dann geben, wenn die Studierenden des ersten Semesters aus der Kohorte des Wintersemesters 2004/05 ihr Vordiplom und Diplom abgelegt haben werden.

18. Analysen der Persönlichkeitsskalen unter einer Normal- und Faking-Good-Instruktion

Die folgenden Analysen beziehen sich auf die Untersuchung der in Kapitel 8.1.5.7 erläuterten Fragestellung, wie verfälschbar Persönlichkeitsfragebögen im Selektionskontext sind und welche Auswirkungen sich hieraus insbesondere auf ihre Kriteriumsvalidität ergeben. Zunächst werden hierzu die *Analysen der generellen Verfälschbarkeit* hinsichtlich Mittelwerts- und Varianzunterschiede von Fragebogendaten alle jener Studienteilnehmer dargestellt, welche die Fragebögen komplett ausgefüllt hatten (n = 427). Hieraus resultierten Stichprobengrößen von n = 209 (Normalinstruktion) bzw. n = 218 (Faking-good-Instruktion). Darauf folgend werden die *Analysen zu den Kriteriumsvaliditäten* im Studienfach Psychologie berichtet.

18.1 Analyse von Mittelwerts- und Varianzunterschieden zwischen Normal- und Faking-good-Instruktion

Einen ersten Eindruck der Unterschiede hinsichtlich Mittelwerts- und Varianzunterschieden und interner Konsistenzen nach Cronbachs Alpha unter beiden Instruktionsversionen gibt Tabelle 111.

Tabelle 111: Deskriptive Statistiken und interne Konsistenzen der z-standardisierten Persönlichkeitsskalenwerten unter Normal- (n = 209) und Faking-good-Instruktion (n = 218) in der Gesamtstichprobe (Fortsetzung der Tabelle auf folgender Seite)

Skala	**Testheft-Version**	***M***	***SD***	***d***	***α***
LMI	Normalinstruktion	-0.32	0.91	0.67	.88
	Faking-good-Instruktion	0.31	0.97		.89
FZA	Normalinstruktion	-0.19	0.95	0.39	.79
	Faking-good-Instruktion	0.19	1.00		.81
HZV	Normalinstruktion	-0.23	0.94	0.48	.79
	Faking-good-Instruktion	0.23	0.99		.79
OFF	Normalinstruktion	-0.07	0.94	0.17	.64
	Faking-good-Instruktion	0.09	0.99		.64

Skala	Testheft-Version	*M*	*SD*	*d*	*α*
GEW	Normalinstruktion	-0.28	0.94	0.58	.81
	Faking-good-Instruktion	0.27	0.97		.84
NEURO	Normalinstruktion	0.24	0.99	0.46	.85
	Faking-good-Instruktion	-0.21	0.95		.84
LÜG	Normalinstruktion	-0.20	0.97	0.40	.84
	Faking-good-Instruktion	0.19	0.99		.85

Anmerkung: LMI: Leistungsmotivationsinventar; FZA: Skala Flexible Zielanpassung; HZV: Skala Hartnäckige Zielverfolgung; OFF: Skala Offenheit; GEW: Skala Gewissenhaftigkeit; NEURO: Skala Neurotizismus; LÜG: Lügenskala

Auf Analyseebene der Mittelwerte sind rein deskriptiv betrachtet in allen Skalen Differenzen zwischen den Gruppen zu beobachten. Die Größe der Mittelwertsdifferenz unter der Faking--good-Instruktion beträgt für die meisten Skalen ca. ein Drittel bis ein Viertel der Standardabweichung der Normalinstruktion, was gut mit bisherigen Ergebnissen zum Ausmaß der Verfälschbarkeit auf Mittelwertsebene korrespondiert (Viswesvaran & Ones, 1999). Die Effektstärken streuen vom niedrigen bis in den hohen Bereich. Am deutlichsten fällt der Effekt für das LMI aus, am geringsten für die Offenheitsskala. Die Standardabweichungen hingegen zeigen in keiner Skala auffällige Abweichungen. Auch die internen Konsistenzen der Skalen bleiben unter der Faking-good-Bedingung erhalten.

Zur inferenzstatistischen Analyse der Mittelwertsdifferenzen wurde eine einfaktorielle multivariate Varianzanalyse mit den z-standardisierten Skalen als abhängige Variablen und den Instruktionsversionen als Faktor nach dem Allgemeinen Linearen Modell berechnet. Das Ergebnis der multivariaten Teststatistik zeigt Tabelle 112:

Tabelle 112: Ergebnisse der multivariaten Teststatistik zu Mittelwertsdifferenzen unter der Normal- (n = 209) und Faking-good-Instruktion (n = 218) anhand der Gesamtstichprobe

Teststatistik	Wert	*F*	Hypothese *df*	Fehler *df*	*p*
Wilks-Lambda	.86	9.16	7.00	419.00	.00

Es ergibt sich ein signifikanter Einfluss des Faktors Instruktion, durch den 14% der Varianz aufgeklärt werden können. Auf Ebene der einzelnen Skalen ergab die Varianzanalyse außer für die Skala „Offenheit für Erfahrungen“ (p > .05) signifikante Mittelwertsdifferenzen in der

erwarteten Richtung (s. Anhang S). Somit zeigen sich die Skalen mit Ausnahme der Offenheitsskala erwartungsgemäß verfälschbar, wobei Unterschiede im Ausmaß der Verfälschbarkeit beobachtbar sind. Skalen, die leistungsrelevante Aspekte betonen (LMI und Gewissenhaftigkeit), sind hiervon besonders stark betroffen.

Die Analyse auf Skalenvarianzunterschiede zwischen den Instruktionsversionen zeigt Tabelle 113 (hierbei wurde das Alpha-Niveau zur Verringerung des unbekannten Beta-Fehlers auf 20% festgelegt).

Tabelle 113: Levene-Test zur Prüfung der Varianzhomogenität zwischen den z-standardisierten Persönlichkeitsskalenwerten unter einer Normal- (n = 209) und einer Faking-good-Instruktion (n = 218) anhand der Gesamtstichprobe

Skala	**Levene-Statistik**	***df*1**	***df*2**	***p***
LMI	0.00	1	425	.98
FZA	1.28	1	425	.25
OFF	0.18	1	425	.67
GEW	0.28	1	425	.59
NEURO	0.00	1	425	.99
LÜG	0.18	1	425	.67

Anmerkung: LMI: Leistungsmotivationsinventar; FZA: Skala Flexible Zielanpassung; HZV: Skala Hartnäckige Zielverfolgung; OFF: Offenheit; GEW: Skala Gewissenhaftigkeit; NEURO: Skala Neurotizismus; LÜG: Lügenskala

Es ergeben sich keine signifikanten Abweichungen von der Annahme homogener Varianzen. Dieses Ergebnis scheint auf den ersten Blick gegenintuitiv, da hypothetisiert werden könnte, dass es unter der Faking-good-Instruktion zu einer Homogenisierung des Antwortverhaltens kommt und demnach die Testwertvarianzen niedriger ausfallen müssten. Allerdings ist eine Varianz*homogenität* keine notwendige Voraussetzung dafür, dass *keine* Antwortverfälschung stattgefunden hat. Zum einen ist nämlich davon auszugehen, dass die Personen auch unter der Faking-good-Bedingung unterschiedlich stark verfälschend antworteten, wie dies auch für die Normalinstruktions-Bedingung nicht ausgeschlossen werden kann und es dadurch zu einer Angleichung der Testwertvarianzen zwischen den Instruktionsversionen kam. Zum anderen setzt sich die Varianz der Testwerte in der Faking-good-Bedingung im Wesentlichen sowohl aus konstrukti*rrelevanter* Varianz zulasten der Faking-Dimension zusammen als auch aus konstrukt*relevanter* der jeweils eigentlich zu messenden Persönlichkeitsdimension. In der

Folge kann Varianzhomogenität vorliegen, obgleich die Varianz*quellen* jeweils verschieden sind. Die Analyse der Kriteriumsvalidität im folgenden Kapitel gibt hierzu weiteren Aufschluss.

18.2 Analyse der Kriteriumsvalidität unter Normal- und Faking-good- Instruktion in der Hauptstudiumsstichprobe

Für die Analyse der Auswirkungen auf die Validität der Fragebögen standen als Kriteriumswerte zum einen die bereits zurückliegenden Vordiplomnoten der Hauptstudiums-Stichprobe in Psychologie als retrospektive Validitäten zur Verfügung. Zum anderen dienten zur Bestimmung der prädiktiven Validität die Leistungen der Psychologie-Studierenden mit Studienbeginn im Wintersemester 2004/05 in der Orientierungsprüfungsklausur. Die folgenden Darstellungen geben generell in einem ersten Schritt die Analyseergebnisse hinsichtlich Mittelwerts- und Varianzunterschiede der Skalen zwischen den Instruktionsversionen wieder, um Hinweise auf das Ausmaß der Antwortverfälschung zu erhalten. Daran schließen sich die eigentlichen Validitätsanalysen an.

Im Folgenden werden zunächst die Ergebnisse für die Hauptstudiums-Stichprobe berichtet. Zur Bestimmung des Ausmaßes der Antwortverfälschung durch die Faking-good-Instruktion in dieser Stichprobe gibt die folgende Tabelle 114 eine vergleichenden Überblick über deskriptive Kennwerte der Skalen.

Tabelle 114: Deskriptivstatistiken der z-standardisierten Persönlichkeitsskalen unter der Normal- (n = 37) gegenüber der Faking-good-Instruktion (n = 38) in der Hauptstudiumsstichprobe (Fortsetzung der Tabelle auf folgender Seite)

Skala	**Testheft-Version**	***M***	***SD***	***d***
LMI	Normalinstruktion	-.67	0.98	
	Faking-good-Instruktion	.88	0.74	1.81
FZA	Normalinstruktion	-.49	0.88	
	Faking-good-Instruktion	.76	0.89	1.40
HZV	Normalinstruktion	-.61	1.10	
	Faking-good-Instruktion	.53	0.82	1.30

Skala	Testheft-Version	*M*	*SD*	*d*
OFF	Normalinstruktion	.00	1.03	
	Faking-good-Instruktion	.46	0.92	0.47
GEW	Normalinstruktion	-.38	1.02	
	Faking-good-Instruktion	.65	0.75	1.20
NEURO	Normalinstruktion	.63	1.12	
	Faking-good-Instruktion	-.80	0.80	1.57
LÜG	Normalinstruktion	-1.10	0.87	
	Faking-good-Instruktion	-0.29	1.09	0.85

Anmerkung: LMI: Leistungsmotivationsinventar; FZA: Skala Flexible Zielanpassung; HZV: Skala Hartnäckige Zielverfolgung; OFF: Skala Offenheit; GEW: Skala Gewissenhaftigkeit; NEURO: Skala Neurotizismus; LÜG: Lügenskala

Es ergeben sich rein deskriptiv betrachtet auffallende Mittelwerts- und in dieser Stichprobe nunmehr auch Varianzunterschiede (vgl. hierzu auch Tabelle 111). Die Effektstärkemaße der Mittelwertsdifferenzen fallen am deutlichsten für das LMI und die Neurotizismusskala aus, liegen mit Ausnahme der Offenheits-Skala generell hoch bis sehr hoch. Das LMI und die Neurotizismusskala scheinen somit besonders durchschaubar und somit stärker verfälschbar zu sein. Wie in den Analysen anhand der Gesamtstichprobe (s. Tabelle 112) ist die Skala „Offenheit für Erfahrungen“ am geringsten von Mittelwertsunterschieden betroffen. Hierbei ist es offenbar schwerer zu durchschauen, was die „Schlüsselrichtung“ sozial erwünschter Antworten ist. Vermutlich drückt sich hierin sogar eine generell größere idiosynkratische Auffassung dieses Konstruktes aus. Diese Interpretation kann in Analogie zu der auch in der Literatur zu findenden größeren Uneinigkeit über die Benennung dieses Faktors als „Culture“, „Intellect“, „Unabhängigkeit der Meinungsbildung“ oder eben „Offenheit für Erfahrungen“ gesehen werden (s. hierzu z. B. Amelang & Bartussek, 1997, S. 366).

Die Ergebnisse einer multivariaten Varianzanalyse nach dem Allgemeinen Linearen Modell zur Identifizierung signifikanter Mittelwertsdifferenzen zeigt Tabelle 115.

Tabelle 115: Ergebnisse der multivariaten Teststatistik zu Mittelwertsdifferenzen unter der Normal- (n = 37) und Faking-good-Instruktion (n = 38) anhand der Hauptstudiumsstichprobe

Teststatistik	**Wert**	***F***	**Hypothese *df***	**Fehler *df***	***p***
Pillai-Spur	.53	10.80	7.00	66.00	.00
Wilks-Lambda	.46	10.80	7.00	66.00	.00

Da der Stichprobenumfang jeweils gering ist, wurde zusätzlich die hierbei robustere Teststatistik Pillai-Spur berechnet (s. hierzu Bortz, 1999, S. 575). Beide Teststatistiken sind signifikant. Nach Wilks-Lambda können durch den Faktor Instruktionsversion 54% der Varianz aufgeklärt werden. In dieser Stichprobe fällt der Effekt also deutlich höher aus als in der Gesamtstichprobe mit 14% aufgeklärter Varianz, was sehr wahrscheinlich auf die größere Testvertrautheit dieser Stichprobe zurückzuführen ist. Die Einzelergebnisse der multivariaten Varianzanalyse zeigen außer für die Skala „Offenheit für Erfahrungen“ (p > .05) durchweg signifikante Mittelwertsdifferenzen im Sinne der Hypothese (s. Anhang T).
Da sich allerdings besonders Varianzunterschiede auf die Kriteriumsvaliditäten der Skalen auswirken können, wurde die Varianzhomogenität der Skalen zwischen den Instruktionsversionen mit dem Levene-Test überprüft. Das Alpha-Niveau wurde zur Verringerung des unbekannten Beta-Fehlers wieder auf 20% festgelegt. Die Ergebnisse zeigt Tabelle 116.

Tabelle 116: Levene-Test zur Überprüfung der Varianzhomogenitätsannahme der z-standardisierten Persönlichkeitsskalenwerten unter der Normal- (n = 209) und Faking-good-Instruktion (n = 218) in der Hauptstudiumsstichprobe

Skala	**Levene-Statistik**	***df*1**	***df*2**	***p***
LMI	1.11	1	72	.29
FZA	0.12	1	72	.72
HZV	1.55	1	72	.21
OFF	0.83	1	72	.36
GEW	6.83	1	72	.01
NEURO	4.09	1	72	.04
LÜG	1.35	1	72	.24

Anmerkung: LMI: Leistungsmotivationsinventar; FZA: Skala Flexible Zielanpassung; HZV: Skala Skala Hartnäckige Zielverfolgung; OFF: Skala Offenheit; GEW: Skala Gewissenhaftigkeit; NEURO: Skala Neurotizismus; LÜG: Lügenskala

Für die Gewissenhaftigkeits- und die Neurotizismusskala muss die Varianzhomogenitätsannahme verworfen werden. Nach Tabelle 114 fallen die Standardabweichungen unter der Faking-good-Instruktion deutlich geringer aus. In dieser mit Persönlichkeitsfragebögen „geübteren“ Stichprobe führt die Faking-good-Instruktion in diesen Skalenwerten also zu einer deutlichen Homogenisierung des Antwortverhaltens.

Wie bereits in Kap. 8.1.5.7 hingewiesen, müssen sich sozial erwünschte Antworten nicht notwendigerweise alleine in Mittelwerts- und/oder Varianzunterschieden zwischen den Instruktionstypen niederschlagen. Ein weiterer wichtiger Hinweis sind Korrelationsanalysen zwischen Skalen zur sozialen Erwünschtheit und den Persönlichkeitsskalen. Im vorliegenden Fall wurde daher in der Hauptstudiumsstichprobe der Gesamtwert der Lügen- und Leugnungsskalen von Ling (1967) mit den Rohwerten der übrigen Skalen korreliert. Hierbei sollten, um einen Effekt der Instruktion nachzuweisen, die Korrelationen unter der Faking--good-Bedingung höher mit den Skalen in der jwls. erwartbaren Richtung liegen als unter der Normalinstruktion. Die Ergebnisse dieser Analyse zeigt Tabelle 117.

Tabelle 117: Korrelationen der Lügenskalenwerte mit denen der Persönlichkeitsfragebögen aus der Hauptstudiumsstichprobe ($N_{\text{Normalinstruktion}} = 37$, $N_{\text{Faking-good.Instruktion}} = 38$)

Instruktions-Version	**LMI**	**FZA**	**HZV**	**OFF**	**GEW**	**NEUR**
Lügenskala Normalinstruktion	-.19	.00	-.19	-.14	.02	-.29*
Lügenskala Faking-good-Instruktion	.27*	.22	.25	.19	.37*	-.11

Anmerkung. Signifikanztestung ohne Korrektur für multiples Testen

*: $p < .05$

LMI: Leistungsmotivationsinventar; FZA: Skala Flexible Zielanpassung; HZV: Skala Hartnäckige Zielverfolgung; OFF: Skala Offenheit; GEW: Skala Gewissenhaftigkeit; NEURO: Skala Neurotizismus; LÜG: Lügenskala

Aufgrund der jeweils geringen Stichprobengrößen fallen die meisten Korrelationen insignifikant aus. Allerdings erkennt man unter der Faking-good-Bedingung die tendenziell höheren Korrelationen im Sinne der Erwartung: höhere Werte auf der Lügenskala gehen (außer für die Neurotizismusskala) tendenziell mit höheren Ausprägungen auf den Persönlichkeitsskalen einher. Interessanterweise verhält es sich bezüglich der Neurotizismusskala genau umgekehrt. Hier zeigt sich unter der Normalinstruktion ein deutlicherer Zusammenhang im Sinne der Erwartung. Eine mögliche Erklärung hierfür ist, dass die Lügenskala zu einem

gewissen Anteil auch echte Traitvarianz von Neurotizismus miterfasst bzw., wegen des negativen Vorzeichens, emotionaler Stabilität: Je emotional stabiler, desto tendenziell weniger „sozial erwünscht“ bzw. weniger sozial ängstlich wird geantwortet. Jedoch fällt dieser Zusammenhang nicht besonders hoch aus, sodass die Validität der Lügenskala als Instrument zur Erfassung sozial erwünschter Antworttendenz unterstützt wird.

Insgesamt ergeben sich somit in dieser Stichprobe zunächst rein deskriptiv betrachtet auf Ebene der Mittelwerte in allen Skalen deutliche Verschiebungen hin zu sozial erwünschten Antworten unter der Faking-good-Bedingung. Auf Ebene der Varianzen zeigen sich allerdings lediglich für die Skala Gewissenhaftigkeit und Neurotizismus statistisch bedeutsame Unterschiede. Diese Skalen scheinen für diese „fragebogengeübtere“ Stichprobe besonders durchschaubar bzw. verfälschbar zu sein, was zu einer deutlichen deutlichen Varianzminderung aufgrund homogenen Antwortverhaltens führt. Auch auf korrelativer Ebene ergibt sich, dass höhere Ausprägungen auf der Lügenskala unter der Faking-good-Bedingung mit sozial erwünschteren Antworten auf den Persönlichkeitsskalen einhergehen als unter der Normalinstruktion.

Um nun die eigentlich interessierenden Auswirkungen der Antwortverfälschung auf die Kriteriumsvaliditäten zu untersuchen, wurden die Skalenrohwerte jeweils mit der Vordiplomdurchschnittsnote der Probanden korreliert. Diese Ergebnisse zeigt Tabelle 118.

Tabelle 118: Korrelationen der z-standardisierten Persönlichkeitsskalenrohwerte mit der Vordiplomnote sowie Skalen-Standardabweichungen in der Hauptstudiumsstichprobe unter Normal- (n = 36) und Faking-good-Instruktion (n = 37)

Instruktions-Version	**LMI**	**FZA**	**HZV**	**OFF**	**GEW**	**NEURO**	**LÜG**
Normalinstruktion	-.39**	-.23	-.40**	-.47**	-.30*	.36*	.15
SD	0.98	0.88	1.10	1.03	1.02	1.12	0.87
Faking-good-Instruktion	.02	.26	.08	.14	.09	-.05	.24
SD	0.74	0.89	0.82	0.92	0.75	0.80	1.09

Anmerkung. Signifikanztestung ohne Korrektur für multiples Testen

*: $p < .05$

LMI: Leistungsmotivationsinventar; FZA: Skala Flexible Zielanpassung; HZV: Skala Hartnäckige Zielverfolgung; OFF: Skala Offenheit; GEW: Skala Gewissenhaftigkeit; NEURO: Skala Neurotizismus; LÜG: Lügenskala

Es zeigen sich nur unter der Normalinstruktion signifikante Kriteriumszusammenhänge im Sinne der Erwartung. Am höchsten fällt die Korrelation der Skala „Offenheit für neue Erfahrungen“, am geringsten (und insignifikant) diejenige der „flexiblen Zielanpassung“ aus. Die Lügenskala zeigt erwartungsgemäß in keiner Bedingung Zusammenhänge mit der Vordiplomnote. Wie bereits in Tabelle 116 ausgewiesen, bestanden nur für die Skalen Gewissenhaftigkeit und Neurotizismus signifikant unterschiedliche Varianzen. Nur für diese Skalen ist daher die Erklärung eines Validitätsabfalls aufgrund von Varianzminderung zutreffend. Dies kann letztlich nur bedeuten, dass durch die experimentelle Induktion von sozial erwünschtem Antwortverhalten die Eindimensionalität der Skalen zerstört wird. Der zusätzliche Einfluss des Fakings führt zu einer Anreicherung konstrukt- und kriteriums-irrelevanter Antwortvarianz, wodurch die Kriteriumsvalidität bis zur Insignifikanz abfällt.

18.3 Analyse der Kriteriumsvalidität unter Normal- und Faking-good- Instruktion in der Erstsemesterstichprobe

Die im Folgenden berichteten Ergebnisse beziehen sich im Gegensatz zum vorherigen Kapitel auf die Analyse die Auswirkungen der Instruktionsversionen auf die *prädiktive* Validität der Fragebogenwerte von Psychologiestudierenden mit Studienbeginn im Wintersemester 2004/05 in Bezug auf die Ergebnisse der OP-Klausur. Analog zum vorigen Kapitel werden vor den Kriteriumsvaliditätsanalysen zuerst die Ergebnisse zu Mittelwerts- und Varianzunterschieden berichtet.

Tabelle 119 gibt zunächst einen Überblick über die Deskriptivstatistiken.

Tabelle 119: Deskriptive Statistiken der z-standardisierten Persönlichkeitsskalenrohwerte unter Normal- (n = 40) und Faking-good-Instruktion (n = 41) in der Erstsemesterstichprobe (Fortsetzung der Tabelle auf folgender Seite)

Skala	**Testheft-Version**	***M***	***SD***	***d***
LMI	Normalinstruktion	-0.20	0.76	0.44
	Faking-good-Instruktion	0.14	0.79	
FZA	Normalinstruktion	-0.09	0.96	0.23
	Faking-good-Instruktion	-0.12	0.85	
HZV	Normalinstruktion	-0.10	0.77	0.34
	Faking-good-Instruktion	0.21	1.12	
OFF	Normalinstruktion	0.37	0.71	0.28
	Faking-good-Instruktion	0.15	0.90	

Skala	Testheft-Version	*M*	*SD*	*d*
GEW	Normalinstruktion	0.00	0.90	0.35
	Faking-good-Instruktion	0.30	0.79	
NEURO	Normalinstruktion	0.27	0.93	0.20
	Faking-good-Instruktion	0.08	0.93	
LÜG	Normalinstruktion	-0.07	0.70	0.33
	Faking-good-Instruktion	0.18	0.81	

LMI: Leistungsmotivationsinventar; FZA: Skala Flexible Zielanpassung; HZV: Skala Hartnäckige Zielverfolgung; OFF: Skala Offenheit; GEW: Skala Gewissenhaftigkeit; NEURO: Skala Neurotizismus; LÜG: Lügenskala

Die Mittelwertsdifferenzen fallen in dieser Stichprobe deskriptiv betrachtet geringer als diejenigen in der Hauptstudiums-Stichprobe aus (vgl. Tabelle 114) und liegen für alle Skalen im mittleren Effektstärke-Bereich. Ein Grund hierfür liegt wahrscheinlich in der geringeren Vertrautheit dieser Probandenstichprobe mit Fragebögen, wodurch es weniger durchschaubar war, welche Persönlichkeitsdimensionen die Fragebögen überhaupt erfassen und in welcher Richtung deshalb die Antworten überhaupt zu verfälschen waren.

Für die inferenzstatistische Überprüfung der Mittelwertsdifferenzen wurde eine multivariate Varianzanalyse nach dem Allgemeinen Linearen Modell berechnet. Die Ergebnisse der Globaltests zeigt Tabelle 120.

Tabelle 120: Ergebnisse der multivariaten Teststatistik zu Mittelwertsdifferenzen unter der Normal- (n = 40) und Faking-good-Instruktion (n = 41) anhand der Erstsemesterstichprobe

Teststatistik	Wert	*F*	Hypothese *df*	Fehler *df*	*p*
Pillai-Spur	.10	0.10	7.00	73.00	.29
Wilks-Lambda	.89	1.23	7.00	73.00	.29

Beide Teststatistiken fallen insignifikant aus. Da in den multivariaten Teststatistiken bzw. in Wilks Lambda die Determinante der Varianz-Kovarianzmatrix als *gemeinsames* Varianzmaß der Skalen (als die sog. generalisierte Varianz) einfließt, lässt sich das Ergebnis dahingehend interpretieren, dass sich die Instruktionsversionen im durch alle Skalen gebildeten *Gesamt*maß nicht unterscheiden. Gleichwohl ergibt eine Varianzanalyse (s. Anhang U) einen signifikanten ($p < .05$) *Einzel*effekt in hypothetisierter Richtung für das LMI; alle übrigen Mittelwerts-

differenzen fallen insignifikant aus ($p > .05$). Wie bereits oben ausgeführt, ist für die weitgehende Insignifikanz der Mittelwertsdifferenzen die geringere Testvertrautheit dieser Stichprobe im Vergleich zur Hauptstudiums-Stichprobe plausibel, wodurch es insgesamt zu geringeren Mittelwertsdifferenzen kommt. Darüber hinaus liegt die post-hoc berechnete Teststärke wegen des geringen Stichprobenumfangs für die meisten Einzeleffekte deutlich unter 80% (s. Anhang U). So lassen sich eindeutige inferenzstatistische Aussagen über die beobachteten Mittelwertsdifferenzen anhand dieser Stichprobe nicht mit der nötigen statistischen Sicherheit treffen.

Die Ergebnisse der besonders für Kriteriumsvaliditäten kritische Überprüfung der Varianzhomogenität zeigt Tabelle 121; hierzu wurde wieder das Alpha-Niveau auf 20% gesetzt, um den unbekannten Beta-Fehler zu reduzieren.

Tabelle 121: Levene-Test zur Überprüfung der Varianzhomogenitätsannahme der z-standardisierten Persönlichkeitsskalenrohwerte anhand der Erstsemesterstichprobe zwischen einer Normal- ($n = 40$) und einer Faking-good-Instruktion ($n = 41$)

Skala	**Levene-Statistik**	***df*1**	***df*2**	***p***
LMI	0.00	1	79	.95
FZA	0.41	1	79	.51
HZV	0.89	1	79	.34
OFF	1.90	1	79	.17
GEW	0.93	1	79	.33
NEURO	0.00	1	79	.98
LÜG	0.32	1	79	.56

Anmerkung: LMI: Leistungsmotivationsinventar; FZA: Skala Flexible Zielanpassung; HZV: Skala Hartnäckige Zielverfolgung; OFF: Skala Offenheit; GEW: Skala Gewissenhaftigkeit; NEURO: Skala Neurotizismus; LÜG: Lügenskala

Im Unterschied zur Hauptstudiumsstichprobe resultiert auf dieser Analyse-Ebene lediglich für die Skala „Offenheit für Erfahrungen“ ein signifikanter Effekt. Im Vergleich der Standardabweichungen aus Tabelle 119 resultiert allerdings unter der Faking-good-Instruktion eine *größere* Varianz als unter der Normalinstruktion. Man könnte, wie bereits im Kapitel 18.2 ausgeführt, vermuten, dass idiosynkratische Auffassungen der Iteminhalte dieser auf Konstruktebene heterogenen Skala zur Varianzerhöhung gerade in der Faking-good-Gruppe führten. Besonders in einer „fragebogen*un*geübten“ Stichprobe dürften die Meinungen unter-

schiedlich ausfallen, in welcher Richtung bspw. ein Item wie „Mich begeistern die Motive, die ich in der Kunst und in der Natur finde“ zu verfälschen ist. Überspitzt formuliert könnte man diese Aussage als zu wenig ziel- und erfolgsorientiert interpretieren oder aber als Indikator eines weit gefassten Interessenbereiches.

Allerdings sind, wie bereits in Kapitel 18.2 dargelegt, nicht identifizierbare Mittelwertsdifferenzen und Varianzinhomogenitäten noch kein hinreichender Beleg dafür, dass die Probanden *nicht* sozial erwünscht geantwortet hatten. Korrelative Analysen mit sozialen Erwünschtheitsskalen können weitere Hinweise liefern. Daher wurden wie in der Analyse der Hauptstudiums-Stichprobe die Rohwerte der Lügenskala mit den Persönlichkeitsskalen in den beiden Instruktionsgruppen korreliert. Die Ergebnisse hierzu zeigt Tabelle 122.

Tabelle 122: Korrelationen der Lügenskalenrohwerte mit denen der Persönlichkeitsskalen in der Erstsemesterstichprobe unter Normal- (n = 40) und Faking-good-Instruktion (n = 41)

Instruktions-Version	**LMI**	**FZA**	**HZV**	**OFF**	**GEW**	**NEURO**
Lügenskala Normalinstruktion	.28*	.19	.10	-.08	.29*	-.41**
Lügenskala Faking-good-Instruktion	.40**	.31*	.38**	.09	.43**	-.47**

Anmerkung. Signifikanztestung ohne Korrektur für multiples Testen

*: $p < .05$

**: $p < .01$

LMI: Leistungsmotivationsinventar; FZA: Skala Flexible Zielanpassung; HZV: Skala Hartnäckige Zielverfolgung; OFF: Skala Offenheit; GEW: Skala Gewissenhaftigkeit; NEURO: Skala Neurotizismus; LÜG: Lügenskala

Hypothesenkonform zeigen sich unter der Faking-good-Bedingung deutlich höhere Zusammenhänge in erwarteter Richtung. Numerisch am höchsten fallen die Beziehungen für die Skalen „Neurotizismus“ und „Gewissenhaftigkeit“ aus. Wie in der Haupstudiums--Stichprobe (vgl. Tabelle 117) ist die negative Korrelation der Lügenskala mit der Neurotizismusskala auch unter der Normalinstruktion auffällig, was abermals darauf hindeutet, dass die Lügenskala Trait-Varianzanteile mit Neurotizismus bzw. mit emotionaler Stabilität teilt, welche sich bei Beantwortung der Lügenskala in sozial ängstlicherem Antwortverhalten niederschlagen. Im Vergleich fällt diese Korrelation numerisch sogar noch deutlich höher als diejenige in der Hauptstudiumsstichprobe mit r = -.29 aus. Überhaupt resultieren hier engere Zusammenhänge auch mit den Skalen Gewissenhaftigkeit und Leistungsmotivation. Dies ließe

sich wiederum damit erklären, dass diese Stichprobe das Konzept einer Lügenskala als Erfassungsinstrument sozialer Erwünschtheit noch nicht hinreichend durchschaute und sich die Skala daher als valider als in der Hauptstudiumsstichprobe erweist.
Somit ergeben sich für diese Stichprobe besonders durch korrelative Analysen deutliche Instruktionseffekte, welche sich durch die Vergleiche von Mittelwerten und Varianzen nicht oder lediglich schwach identifizieren ließen. Verglichen mit der testerfahreneren Haupt-studiumsstichprobe bedeutet dies, dass Testvertrautheitseffekte eine weitere wichtige Quelle der Antwortverfälschung darstellen.

Für die abschließende Analyse der Auswirkungen sozialer Erwünschtheit auf die prädiktive Validität wurden die Skalenrohwerte unter beiden Bedingungen mit den Punktwerten der OP-Klausur korreliert. Das Ergebnis gibt Tabelle 123 wieder.

*Tabelle 123***:** Kriteriumskorrelationen und Standardabweichungen der z-standardisierten Persönlichkeitsskalenrohwerte in der Erstsemesterstichprobe unter Normal- (n = 40) und Faking-good-Instruktion (*n* = 41)

Instruktions-Version	**LMI**	**FZA**	**HZV**	**OFF**	**GEW**	**NEURO**	**LÜG**
Normalinstruktion	.25	-.43*	-.14	-.17	-.35*	.26	.02
SD	0.76	0.96	0.77	0.71	0.90	0.93	0.70
Faking-good-Instruktion	.18	.02	-.29	.05	.09	-.06	.03
SD	0.79	0.85	1.12	0.90	0.79	0.93	0.81

Anmerkung. Signifikanztestung ohne Korrektur für multiples Testen

*: $p < .05$

LMI: Leistungsmotivationsinventar; FZA: Skala Flexible Zielanpassung; HZV: Skala Hartnäckige Zielverfolgung; OFF: Skala Offenheit; GEW: Skala Gewissenhaftigkeit; NEURO: Skala Neurotizismus; LÜG: Lügenskala

In Übereinstimmung der Analysen zu retrospektiven Validitäten ergeben sich, außer für die Skala „Gewissenhaftigkeit“, nur unter der Normalinstruktion signifikante moderate Zusammenhänge in der erwarteten Richtung. Keinerlei substanzielle Korrelationen zeigen sich für die Lügenskala. Die moderat negative Korrelation der Gewissenhaftigkeits-Skala mit dem Gesamtrohwert der Orientierungsprüfung fällt hier deutlich aus dem Rahmen der Erwartung. Man muss dieses Ergebnis vor dem Hintergrund der geringen Stichprobengröße bzw. des großen Stichprobenfehlers als wenig verlässlich ansehen, zumal dies deutlich entgegen

bisheriger Befunde steht, wonach Gewissenhaftigkeit *positive* Korrelationen zu Leistungskriterien aufwies (s. u.a. Barrick & Mount, 1991; Judge, Higgins, Thoresen & Barrick, 1999; Mount, Barrick & Strauss, 1994). Statistisch betrachtet lässt hier sich das Vorliegen einer Mischgruppenverteilung unter der Normalinstruktion vermuten, wobei in der einen Gruppe ein Nullzusammenhang bestehen müsste, in der anderen hingegen ein negativer, sodass der Gesamtzusammenhang negativ ausfällt. Die negative Korrelation in der einen Gruppe spräche sodann auch für eine unter der Normalinstruktionsbedingung vorhandenen Substichprobe mit sozial erwünschter Antworttendenz. Zur Identifizierung einer Mischverteilung wurde deshalb eine hierarchische Cluster-Analyse über die Personen der Normalinstruktion nach dem Ward-Verfahren berechnet (Distanzmaß: quadrierte euklidische Distanz). Es resultierten zwei Personencluster mit $n_1 = 13$ und $n_2 = 19$. Die Korrelationen innerhalb dieser Cluster von Gewissenhaftigkeits-Rohwerte mit dem Punktwert der OP-Klausur ergaben für Personencluster 1 $r = -.19$, für Cluster 2 $r = -.02$. Die Nullkorrelation in Cluster 2 war auf eine deutliche Varianzeinschränkung aufgrund eines Deckeneffektes rückführbar ($M = 31.20$, $SD = 2.27$). Insgesamt wird damit die post-hoc aufgestellte Hypothese einer Subgruppe mit sozial erwünschter Antworttendenz auch unter der Normalinstruktion (Cluster 1) in dieser Stichprobe unterstützt.

Insgesamt betrachtet muss man, wie bereits zuvor bei den retrospektiven Validitätsuntersuchungen, auch für diese Analysen feststellen, dass die unter der Normalinstruktion vorhandene Validität der Verfahren durch sozial erwünschte Antworten unterminiert wird. Dies ist umso interessanter, da sich ein Einfluss von sozialer Erwünschtheit über Mittelwertsdifferenzen und der besonders wichtigen Frage nach Varianzunterschieden in den Skalenwerten zwischen den beider Instruktionsversionen *nicht* eindeutig nachweisen ließ. Lediglich auf korrelativer Ebene ließ sich durch soziale Erwünschtheitsskalen ein Effekt der Antwortverfälschungstendenz nachweisen.
Das Ergebnis kann gerade vor diesem Hintergrund nur dahingehend interpretiert werden, dass die Konstruktvalidität bzw. Eindimensionalität der Skalen durch eine Faking-good-Instruktion soweit verändert wurde, dass ein Faking-Faktor das eigentlich zu messende Persönlichkeitsmerkmal durch konstruktirrelevante Varianz überdeckte und es somit zu einer Kriteriumsinvalidierung der Skalen kam.

18.4 Zusammenfassung der Analysen von Persönlichkeitsfragebögen unter Normal- und Faking-good-Bedingungen

Die hier durchgeführten Analysen zeichnen insgesamt ein pessimistisches Bild der Brauchbarkeit von Fragebogendaten im Selektionskontext. In einer Analyse über die Gesamtstichprobe konnten außer für die Skala „Offenheit für Erfahrungen“ bedeutsame Mittelwertsdifferenzen im Sinne der Hypothese festgestellt werden, sodass die Annahme der Verfälschbarkeit von Fragebogenantworten einmal mehr unterstützt wurde. Hinsichtlich der Varianzunterschiede zeichnen die Analysen ein uneinheitliches Bild. In der Gesamtstichprobenanalyse konnten keine Varianzunterschiede der Skalenrohwerte zwischen den Instruktionsbedingungen nachgewiesen werden. Sie traten allerdings bei einer Substichprobe von Psychologie-Hauptstudiums-Studierenden in den Skalen „Neurotizismus“ und „Gewissenhaftigkeit“ deutlich hervor, wie auch die Mittelwertsdifferenzen in dieser Stichprobe insgesamt deutlich ausgeprägter waren. Die Korrelationen der Lügenskala mit den Persönlichkeitsfragebögen lagen hier lediglich tendenziell in der erwartbaren Richtung; wahrscheinlich, weil diese Stichprobe auch das Konzept der Lügenskala durchschaute. Da es sich bei dieser Stichprobe um eine mit der größten Fragebogenkenntnis handelt, kann man in einer Extrapolation der Ergebnisse von starker Antwortverfälschbarkeit durch Testvorbereitung ausgehen.

Hinsichtlich der Kriteriumsvalidität kam es in der Faking-good-Version zur vollständigen Invalidierung der unter der Normalinstruktion vorhandenen Validität mit der retrospektiv erhobenen Vordiplomnote. Besonders in Anbetracht von weitgehend erhaltener Rohwertestreuung in den meisten Skalen auch unter der Faking-good-Bedingung kann dieses Ergebnis nur über eine Mehrdimensionalität der Skalen durch einen Faking-Faktor erklärt werden, welcher sowohl die Konstrukt- als auch die Kriteriumsvalidität zerstört. In einer mit Testverfahren vergleichsweise weniger erfahrenen Stichprobe von Psychologie-Studierenden aus dem ersten Semester waren deutlich schwächer ausgeprägt Verfälschbarkeitseffekte durch die Instruktionsmanipulation zu beobachten. Lediglich im LMI kam es zu signifikanten Mittelwertsdifferenzen.
Signifikante Varianzunterschiede der Skalenwerte zwischen den Versionen waren in dieser Stichprobe nur für die Skala „Offenheit für Erfahrungen“ feststellbar und dort in hypothesenentgegengesetzter Richtung: Die Faking-good-Instruktion führte nicht zu einer Homogenisierung des Antwortverhaltens und vergleichsweise geringerer Varianz, sondern zu einer

größeren. Dies lässt vermuten, dass die Faking-good-Instruktion idiosynkratische Interpretationen dieser an sich bereits heterogenen Skala noch beförderte.
Überdies ließen sich auf Ebene von Korrelationen der Lügenskalen mit den Persönlichkeitsskalen Effekte zwischen den Instruktionsversionen im Sinne der Hypothese feststellen. Unter der Faking-good-Instruktion lagen die erwarteten Korrelationen mit den Lügenskalen deutlich höher als unter der Normalinstruktion. Auch waren hierbei die Zusammenhänge enger als in der Hauptstudiumsstichprobe, sodass die Validität von Skalen zur Erfassung sozialer Erwünschtheit bei testunerfahrenen Probanden größer ausfällt als bei testerfahrenen. Anders ausgedrückt verlieren allerdings auch diese Skalen durch Testerfahrung oder Testschulung ihre Validität.

In der Analyse der prädiktiven Validität zeigten sich in analoger Weise zu den Ergebnissen der retrospektiven Validität unter der Faking-good-Bedingung keine Zusammenhänge der Persönlichkeitsskalen mit den Ergebnissen der OP-Klausur. Eine negative Korrelation von Gewissenhaftigkeit und Klausurpunkten unter der Normalinstruktionsbedingung war auf das Vorliegen einer Mischverteilung mit einer, durch Tendenz zu sozial erwünschten Antworten gekennzeichneten Subgruppe, rückführbar.

Die Ergebnisse legen insgesamt den Schluss nahe, dass die Konstrukt- und Kriteriumsvalidität der Fragebögen durch Antwortverfälschungen zerstört wird. Angesichts der vorliegenden Ergebnisse zur Verfälschbarkeit, und auch der lediglich niedrigen bis moderaten Zusammenhänge unter der Normalinstruktion, ist der Einsatz von Persönlichkeitsfragebögen im Selektionskontext mit seinen weitreichenden Konsequenzen für das Individuum daher nicht zu befürworten.

19. Diskussion

Das Ziel dieser Arbeit lag in der theoriegeleiteten Konstruktion sowie der nachfolgenden Evaluation eines fachspezifischen Studierfähigkeitstest für Psychologie insbesondere hinsichtlich eines *Validitätszuwachses*, der von einer Ergänzung der Abiturdurchschnittsnote als Prädiktor durch ihn zu erwarten wäre. Ferner interessierte die *Differenzierungsfähigkeit* des Verfahrens gegenüber Studienanfängern anderer Fächer, um hierüber ein empirisches Indiz für die „Psychologie-Spezifität" der Testdimensionen zu erhalten. Schließlich galt es, die Einsatzmöglichkeiten von *Persönlichkeitsfragebögen im Selektionskontext* zu überprüfen. Im Folgenden sollen die Ergebnisse zur Beantwortung dieser in Kapitel 6 ausgeführten Hypothesen und Fragstellungen diskutiert werden, um abschließend eine Gesamtbeurteilung der Ergebnisse und einen Ausblick zu geben.

19.1 Zur inkrementellen Validität der Testverfahren

Hypothese I und II zu angenommenen substanziellen Zusammenhängen von Schulnoten und Testergebnissen mit Studienleistungskriterien fanden Unterstützung anhand der vorliegenden Daten. Insbesondere die Abiturdurchschnittsnote zeigte in allen untersuchten Stichproben bedeutsame Zusammenhänge in der hypothetisierten Richtung mit den jeweiligen Studienleistungen, was im Einklang mit bestehenden Befunden zu Abiturleistungen und Studienerfolg steht (vgl. z. B. Hell et al., 2005 sowie Schuler et al., 1990). Aber auch die Testergebnisse korrelierten im Sinner der Erwartung mit den jeweiligen Kriteriumsausprägungen und im Falle der Intelligenztests in der in der Literatur berichteten Höhe (vgl. Amelang, 1975; Gasch, 1971; Trost & Bickel, 1979; Trost, 1975). Je nach Kriterium traten hierbei allerdings Unterschiede in den Zusammenhängen auf, d.h. nicht alle Tests korrelierten jeweils substanziell mit den verschiedenen Kriterienmaßen.

Hypothese III zur inkrementellen Validität der Testverfahren als Ergänzung zur Abiturdurchschnittsnote wurde anhand von zwei Stichproben untersucht. Zum einen an Probanden aus dem Hauptstudium in Psychologie, wodurch sich retrospektive Validitäten der Testverfahren zur Vordiplomnote ergaben, zum anderen von Psychologie-Studierenden mit Studienbeginn im Wintersemester 2004/05.

In der *Hauptstudiumsstichprobe* ergaben sich außer für den Subtest zur verbalen Intelligenz signifikante Zusammenhänge der Testverfahren mit der Vordiplomnote. Allerdings fiel die

Kriteriumskorrelation der Abiturdurchschnittsnote deutlich höher aus. In einer simultanen Analyse der Prädiktor- und Kriteriumsvariablen mit einer kanonischen Korrelation war die Abiturdurchschnittsnote am stärksten mit den Kriterien auf Variablenseite assoziiert. Ebenso zeigte sie in Regressionsanalysen zur Analyse der inkrementeller Validität von Testverfahren in Bezug auf die durchschnittliche Vordiplomnote den größten Vorhersagebeitrag. Ein Inkrement war in dieser Stichprobe über einen Test zur fluiden Intelligenz möglich, alle weiteren Testverfahren zeigten keine statistisch bedeutsamen prädiktiven Anteile.
Als problematisch für eine Generalisierung dieses Befundes erwiesen sich Kreuzvalidierungsanalysen der ermittelten Regressionsgleichungen. Sie verwiesen über die Instabilität der Schätzgüte der Regressionsgleichung in den Kreuzvalidierungsstichproben auf eine starke Stichprobenabhängigkeit der Ergebnisse.

Für die Interpretation dieses Befundes weiterhin kritisch erwiesen sich Hinweise auf das Vorliegen von Testwiseness-Effekten in der vorliegenden Stichprobe. Insbesondere die nahe an Null liegenden bzw. insignifikanten konvergenten Validitäten der Intelligenztestleistungen mit den schulischen Leistungen widersprachen denen in der Literatur durchgängig zu findenden mittleren Zusammenhängen (s. etwa Amelang & Bartussek, 1997, S. 246). In der Stichprobe aus dem ersten Semester waren höhere Zusammenhänge zu beobachten. Möglicherweise kam es aufgrund von Testwiseness insgesamt zu einer Verzerrung sowohl der Kriteriums- als auch Konstruktvalidität. Ist jedoch gerade letztere nicht mehr sichergestellt, messen also die Tests zusätzlich noch die Dimension der Testvorerfahrung, bleibt fraglich, wie viel konstruktrelevante oder konstruktirrelevante Varianzanteile in die Test-Kriteriumskovarianz eingeflossen sind. Auch vor diesem Hintergrund kann nicht mit der nötigen Sicherheit auf ein tatsächliches Inkrement durch fluide Intelligenz geschlossen werden. Zudem bleibt fraglich, wie repräsentativ die durch freiwillige Testteilnahme gewonnene Stichprobe aus dem Hauptstudium tatsächlich war.

Auf einer solideren Basis stehen hingegen die Ergebnisse der *Probanden aus dem ersten Semester* vom Wintersemester 2004/05 in Psychologie, zumal hier eine Vollerhebung möglich war. Hier standen zum einen die Ergebnisse der Orientierungsprüfungsklausur (OP-Klausur) für die Analyse prädiktiver Validität zur Verfügung, zum anderen zwei im ersten Semester geschriebene Klausuren aus dem Fach Methodenlehre.
Neben der Abiturnote zeigte nur der Test der verbalen Intelligenz und derjenige zur Kreativität statistisch signifikante Korrelationen mit dem OP-Klausurergebnis. In einer simultanen Analyse aller Prädiktoren und den Noten der OP-Klausur anhand einer kanonischen

Korrelation wiesen allerdings ausschließlich die Abiturnote und die (um die Kovarianzanteile der Abiturdurchschnittsnote bereinigten) Abitureinzelfachnoten in Deutsch und Englisch prädiktive Beiträge auf.
Regressionsanalysen der Prädiktoren mit dem Rohwert der OP-Klausur als abhängige Variable ergaben jedoch nur für die Abiturdurchschnittsnote einen signifikanten Vorhersagebeitrag. Dieser erwies sich auch in Kreuzvalidierungsanalysen der Regressionsergebnisse als stabil. Uneinheitlicher erwiesen sich die Ergebnisse bezüglich der beiden Methodenlehre-Klausuren. In beiden korrelierte die Abiturdurchschnittsnote signifikant mit den Kriterien. Für die erste Klausur (TestTutVL) ergab sich lediglich für den Kreativitätstest ein statistisch bedeutsamer Zusammenhang, der zudem in einer Regressionsanalyse ein Inkrement zur Abiturnote aufwies. In der zweiten Methodenklausur (TestVL) allerdings waren aus der Reihe der Testverfahren mehr signifikante Kriteriumskorrelationen zu beobachten. Hier zeigten sich Beziehungen zur verbalen Intelligenz, Kreativität und dem Test zum empiriebezogenen Denken. Jedoch ergab eine Regressionsanalyse lediglich für die Skala zur verbalen Intelligenz einen allerdings deutlichen zusätzlich aufgeklärten Varianzanteil gegenüber der Abiturdurchschnittsnote. Gleichwohl erwies sich die Stabilität dieser regressionsanalytischen Ergebnisse in Kreuz-validierungsanalysen als gering, insbesondere in Bezug auf das Inkrement durch den Kreativitätstest; hinsichtlich des Tests zur verbalen Intelligenz war die Stabilität befriedigender.

Gründe für die Uneinheitlichkeit der Befunde in Bezug auf das Fach Methodenlehre können zunächst in den unterschiedlichen Messbereichen beider Klausuren gefunden werden. Sie beinhalteten jeweils verschiedene Themen aus der Methodenlehre und waren daher auch lediglich moderat miteinander korreliert. Womöglich lag darüber hinaus die Reliabilität der zweiten Klausur insgesamt höher, sodass es hier zu mehr statistisch substanziellen Zusammenhängen mit den Testverfahren kam. Auffällig war in diesem Zusammenhang auch die höhere Korrelation der Abiturnote mit der zweiten gegenüber der ersten Klausur.

In Bezug auf die unter Kapitel 6 genannte *Hypothese III zur inkrementellen Validität* durch die Testverfahren sind die Ergebnisse somit insgesamt kritisch zu bewerten. Mögliche Gründe hierfür sollen im Folgenden diskutiert werden.
Zunächst sei darauf hingewiesen, dass die Ergebnisse in Anbetracht einer nach schulischen Leistungsmaßen hochselegierten und daher sehr leistungshomogenen Stichprobe zu relativieren sind. Die Homogenität der Stichproben zeigte sich sowohl in den Abiturleistungen als auch in den geringen Varianzen der Vordiplom- und Klausurleistungen. Die jeweiligen

Verteilungen zeigten eine ausgeprägte Schiefe (s. hierzu Tabelle 74, Tabelle 76, Tabelle 90 und Tabelle 92). Die an die Testverfahren gestellten Anforderungen, in einer nach überwiegend kognitiven Merkmalen selegierten Probandenstichprobe ein Inkrement zu erzielen, lagen demzufolge sehr hoch.
In diesem Zusammenhang sind auch Bedenken hinsichtlich der psychometrischen Güte der Testverfahren zu nennen. Da es sich in den meisten Fällen um Testkonstruktionen des Autors handelte, stellen sie reine Vorversionen dar. Insbesondere die hieraus resultierenden relativ geringen Reliabilitäten der Subtests mögen mit ein Grund gewesen sein, dass die Zuwächse an prognostischer Validität insgesamt betrachtet vergleichsweise gering und mitunter inkonsistent ausfielen.

Die Unreliabilität der Kriterien als weitere Ursache für die inkonsistenten Befunde zur inkrementellen Validität zu betrachten, kann für die Vordiplomnoten und die Orientierungsprüfungsklausur nicht geltend gemacht werden, da deren internen Konsistenzen jeweils als gut bezeichnet werden konnten (vgl. Kap. 17.2.1.2 und 17.2.4.2). In Bezug auf die Methodenlehre-Klausuren trifft dieser Einwand hingegen zu (vgl. Kap. 17.2.5.1.1). Für diese wäre in der Tat bei höherer Reliabilität mit stärkeren Prädiktor-Kriteriumskorrelationen zu rechnen. Wie sich dies jedoch in der Folge auf die Befunde zur inkrementellen Validität auswirken könnte, bleibt unbestimmt, da sodann ebenso mit einer (noch) höheren Prädiktionskraft der Abiturdurchschnittsnote zu rechnen wäre.

Darüber hinaus ist in Bezug auf die Testkonstruktion an sich kritisch anzumerken, dass die Kriteriumsvalidierungen an derselben Stichprobe vorgenommen wurden, die bereits für die Skalen- und Itemanalysen nach Rasch-Modellen diente bzw. welche die größte Substichprobe bildete. Die Generalisierbarkeit der Ergebnisse ist daher auch von dieser Seite her kritisierbar. Hierbei ist jedoch daran zu erinnern, dass die Kreuzvalidierungsanalysen der Itemselektion (s. Kap. 15) zu weitgehend übereinstimmenden Ergebnissen gelangten. Zwar ist hierbei Stelzl (2005) zuzustimmen, dass eine Kreuzvalidierung der am Gesamtdatensatz gewonnenen Hypothesen über die Itempassung anhand von Zufallsstichproben aus den Gesamtdaten keine Kreuzvalidierung im strengen Sinne darstellt, doch als erste „Approximation“ strenger Kreuzvalidierungsanalysen an echt unabhängigen Stichproben aus Neuerhebungen erschien dieses Vorgehen als unabdingbar, um überhaupt Hinweise über die Stabilität der Ergebnisse zu erhalten.
Dass allerdings jede Itemselektion, die nicht auf rein technische Itemkonstruktionsfehler rückführbar ist, ganz gleich, ob sie in einer vorläufigen oder strengen Kreuzvalidierung

repliziert wurde, im Rahmen von Rasch-Analysen kritisch zu sehen ist, wurde bereits in Kapitel 9.2 erörtert. Die Forderung nach Unabhängigkeit der Vergleiche zwischen Personen von der Itemauswahl ist sodann aus logischen Gründen anzweifelbar. Dementsprechend ist auch die hier für die Subtests nach Itemselektion und teilweise zusätzlicher Spezifikation eines Speed-Parameters angenommene Modellgeltung als vorläufig anzusehen.

Einschränkungen der Ergebnisse im Hinblick auf Studierfähigkeitstests ergeben sich jedoch besonders seitens der Inhalte der Testverfahren. Das wohl grundlegenste Problem stellt hierbei die – entgegen dem ursprünglichen Anspruch – große Unspezifität der Verfahren dar. Die über den Workshop zur Anforderungsanalyse abgeleiteten erfolgskritischen Personenmerkmale spiegeln zum überwiegenden Teil letztlich doch sehr *allgemeine* studienerfolgskritische Personenmerkmale wieder und sind daher als weitaus weniger *studienfachspezifisch* zu bezeichnen. Diese ließen sich rein rational betrachtet ohne weiteres auch als notwendige allgemeine Personeneigenschaften etwa für ein Studium der Humanmedizin oder der Rechtswissenschaften übertragen. Den aus dieser Anforderungsanalyse abgeleiteten Testverfahren mangelt es also erheblich an der geforderten Studienfachspezifität. Mit anderen Worten: Was als Idee eines studienfachspezifischen Auswahlverfahrens begann, endete als allgemeiner Studierfähigkeitstest. Allenfalls lässt sich als studienfachspezifische Komponente der Test zum empiriebezogenen Denken nennen. In fast schon tragikomischer Weise muss man sich aber daran erinnern, dass dieser nicht auf Basis der Ergebnisse der Anforderungsanalyse gewonnen wurde, sondern aus rationalen Überlegungen des Autors. Ein Grund für die mangelnde Fachspezifität der Verfahren liegt sicher in der raschen Art der Anforderungsanalyse über einen Workshop. Für eine genauere Erschließung tatsächlich fachrelevanter Personenmerkmale hätten mehr Zeit und mehr Erhebungsinstrumente verwendet werden müssen.
Daher bedürfen die abgeleiteten Aussagen über inkrementelle Validitäten der Einschränkung auf *allgemeine* studienerfolgskritische Personenmerkmale. Es ist nicht auszuschließen, dass sehr viel spezifischere Testverfahren ein Mehr an inkrementeller Validität erbracht hätten. Die hier teilweise identifizierbaren Inkremente bewegen sich daher auch im Wertebereich von denen in der Literatur berichteten zu allgemeinen Studierfähigkeitstest (vgl. hierzu Kap. 4.2.3.2). Gleichwohl zeigt auch die Literatur zu studienfachspezifischen Auswahlverfahren lediglich geringe inkrementelle Validitäten über die Schulabschlussnote hinaus (Bolwahnn, 2003; Hell et al. 2005; Trost et al. 1998). Da jedoch bislang sehr wenig Forschung

zur Konstruktion eines studienfachspezifischen Auswahlverfahrens für das Fach Psychologie vorliegt, obliegt die Beantwortung dieser Frage letztlich weiterer Forschung.

Im Zusammenhang mit der mangelnden Studienfachspezifität der Testverfahren sind auch die Befunde hinsichtlich *Hypothese IV* zur Diskriminationsfähigkeit zwischen Studienteilnehmern verschiedener Fächer zu sehen. Hierbei ließ sich kein typisches und abgrenzbares Leistungsprofil von Psychologiestudierenden finden. Lediglich in den Tests spezifischerer Leistungsbereiche von Facetten der Kreativität und des empiriebezogenen Denkens erwiesen sich die Psychologiestudierenden aus dem ersten Semester denjenigen anderer Fächer als überlegen. Hierbei war jedoch lediglich die Dimension der Kreativität direkt aus den anforderungsanalytischen Ergebnissen abgeleitet worden; der Test zum empiriebzogenen Denken entstand, wie bereits weiter oben ausgeführt, aus Überlegungen des Autors.
Die Testverfahren sind also zu allgemein gehalten, um gemeinhin über die getesteten Inhalte eine „fächerdiskriminante“ Validität aufzuweisen.

Bezüglich der Generalisierbarkeit der Ergebnisse sei einschränkend darauf hingewiesen, dass die hier zur Analyse herangezogenen Stichproben jeweils nicht repräsentativ für das Bundesgebiet waren, sondern zum überwiegenden Teil aus Baden-Württemberg stammten bzw. dort ihr Abitur abgeleistet hatten (Alexandra Hohneder, persönliche Mitteilung vom 14.8.2006). Da in Baden-Württemberg ein Zentralabitur durchgeführt wird und somit einheitliche Prüfungsanforderungen und Auswertungsmaßstäbe vorliegen, dürfte die Reliabilität der Abiturnoten hier höher liegen als in Bundesländern ohne Zentralabitur (vgl. hierzu auch Prenzel et al., 2005). Hinzu tritt das (neben Bayern) höhere Anspruchsniveau des Baden-Württembergischen Abiturs (Prenzel et al., 2005). Eine höhere Reliabilität bzw. geringere Fehlervarianz der Abiturleistungen sowie das höhere Anspruchsniveau der Abituraufgaben mindert konsequenterweise die Wahrscheinlichkeit, durch Auswahltests einen zusätzlichen Vorhersagebeitrag zu leisten. In diesem Sinne soll – wie bereits in Kapitel 17.2 vorgebracht – noch einmal betont werden, dass die hier berichteten Ergebnisse nur auf eine Population zu verallgemeinern sind, welche nach dem bisherigen Auswahlmodus an der Universität Heidelberg zum Psychologiestudium zugelassen wurde. Sie treffen demnach keine Aussagen über den zusätzlichen Nutzen von Testverfahren an psychologischen Fachbereichen anderer Universitäten, etwa an sogenannten „Massenuniversitäten“ mit deutlich heterogener Bewerberzusammensetzung als in Heidelberg (z. B. aufgrund von mehr Zulassungen über Wartesemestern, Bewerbern aus verschiedenen Bundesländern wie auch unterschiedlicher

Nationen etc.). Hier könnten Auswahltests womöglich eine höhere inkrementelle Validität aufweisen.

19.1.1 Fazit der Ergebnisse zur inkrementellen Validität

Betrachtet man die Analysen an der Stichprobe von Psychologie-Studierenden mit Studienbeginn im Wintersemester 2004/05 wegen ihrer größeren Repräsentativität und aufgrund prädiktiver Aussagen (gegenüber retrospektiven) als gewichtiger, so muss in Bezug auf die OP-Klausurergebnisse die Hypothese einer inkrementellen Validität durch die Testverfahren klar verworfen werden. Dies wiegt für die Testverfahren umso schwerer, als dass die OP--Klausur unter den Kriterien, die zur Berechnung prädiktiver Validitäten zur Verfügung standen, einen höheren Rang als die Methodenklausuren einnimmt, in welchen ein vergleichsweise stabiles Inkrement der verbalen Intelligenz zu identifizieren war. Ihr Bestehen entscheidet über eine Fortsetzung oder einen Abbruch des Studiums. Nimmt man diese Klausur also als vorläufige „Nagelprobe“ inkrementeller Validität, ist der Nutzen dieser Testverfahren neben der Abiturdurchschnittsnote nicht gegeben.

In diesem Zusammenhang ist ebenso festzustellen, dass die Abiturdurchschnittsnote *trotz* einer deutlichen Varianzeinschränkung durch die Numerus-clausus-Regelung den sowohl hinsichtlich der Prädiktionskraft als auch im Hinblick auf die Vorhersagestabilität besten Einzelprädiktor darstellte. Insbesondere die hohe Validität von $r = .50$ in Bezug auf die Vordiplomnote als auch ihre in einzelnen Grundstudiumsklausuren durchgängig signifikanten und stabilen Vorhersagebeiträge unterstützen diese Interpretation.

Die ersten Ergebnisse zu Analysen der inkrementellen Validität zusammenfassend betrachtet lässt sich feststellen, dass als wesentlichster Befund der Status der Abiturnote als vergleichsweise bester und stabilster Prädiktor für Studienerfolg repliziert wurde (vgl. u.a. Hell, Trapmann, Weigand, Hirn & Schuler, 2005; Rindermann & Oubaid, 1999; Schuler, Funke & Baron-Boldt, 1990; Trost, 1975, 2003; Trost & Bickel, 1979). Durch die Hinzunahme von Testergebnissen in multiplen Regressionsanalysen konnte in Klausuren aus dem Grundstudium im Fach Methodenlehre eine Verbesserung der Vorhersage um 7% bzw. 14% zusätzlich aufgeklärter Varianz erzielt werden. Für eine Orientierungsprüfungsklausur ergab sich kein Inkrement über die Abiturnote. Ergebnisse aus Kreuzvalidierungen der

Regressionsschätzungen erwiesen sich, außer im Falle der Abiturdurchschnittsnote, als instabil.

Eine abschließende Beurteilung der Hypothese zur inkrementellen Validität der Testverfahren kann allerdings erst erfolgen, wenn die Probandenstichprobe von Psychologie-Studierenden mit Studienbeginn im Wintersemester 2004/05 ihr Vordiplom abgelegt hat, ohnehin am verlässlichsten erst, wenn die Hauptdiplomnote als Kriterium mit dem höchsten Rang vorliegt.

Die Ergebnisse der Validitätsanalysen dieser Arbeit auch im Lichte bisheriger Forschung betrachtend lässt sich feststellen, dass die Vorhersage von Studienerfolg weiterhin nicht in befriedigendem Ausmaß gelingt. Die Ergebnisse insbesondere aus der Verknüpfung von Prädiktoren über multiple Regressionsanalysen in der vorliegenden Arbeit stimmen in ihrer Gesamtheit gut mit den in der Literatur berichteten Befunden überein, wonach maximal ca. ein Drittel der Kriterienvarianz durch die Kombination von Prädiktoren aufgeklärt wird (vgl. hierzu Trost, 1975, S. 46 sowie Tabelle 5). Sie bestätigen damit ein weiteres Mal die Problematik einer unter statistischen wie auch unter Aspekten individueller Entscheidungen nicht befriedigenden Vorhersagegenauigkeit des Studienerfolgs (vgl. z. B. Trost, 1975, S. 46 sowie Trost, 2003, S. 37f.). In diesem Sinne ist Amelang (1978) zuzustimmen, wenn er in Bezug auf die Verbesserung der Vorhersagegenauigkeit bereits vor nunmehr beinahe 30 Jahren anmerkte: „In dieser Hinsicht sind über den Erscheinungsjahren der publizierten Untersuchungen keine substanziellen Fortschritte zu registrieren“. Eine Gesamtvarianzaufklärung von maximal einem Drittel scheint weiterhin die obere Grenze dessen zu sein, was mit bisherigen Auswahlverfahren innerhalb der facettenreichen Struktur des Studienerfolgs möglich ist. Sollte sich an diesem Ergebnis auch zukünftig nicht viel ändern, und nach Meinung des Autors besteht hierin vor dem Hintergrund der in Kapitel 4 berichteten Ergebnisse wenig Hoffnung, so wäre wohl insgesamt zu fordern, dass nicht nur auf einen einzigen Bedingungsaspekt von Ausbildungserfolg fokussiert wird, nämlich der Selektion von Bewerbern, sondern ebenso auf die Studienqualität und -bedingungen. Treffend bemerken Rindemann und Oubaid (1999) nämlich in diesem Zusammenhang:

> Schließlich kann es nicht Ziel der Universitäten sein, gute Studienanfänger nach ein paar Jahren nur als gealterte und – auch ohne die Universitäten – gereifte Personen in die Gesellschaft zu entlassen ... ebenso müssen Veränderungen in der universitären

> Ausbildung, der institutionellen Gewichtung der Hochschullehre und in den universitären Rahmenbedingungen eingeleitet werden. (S. 188)

Die Ergebnisse der vorliegenden Studie verweist somit ganz im Sinne von Rindermann und Oubaid (1999) sowie Amelang und Funke (2005) darauf, dass eine Auswahl alleine über die Kombination aus Abiturdurchschnittsnote und Testergebnissen nicht in dem Maße zur Verbesserung der Vorhersage beiträgt, wie dies eigtl. erwünscht und nötig wäre. Offenbar bedarf es hierzu in der Tat sequenzieller Auswahlstrategien, auch wenn diese von einem erheblichen Aufwand geprägt sind, doch durchaus auch schon mit positiven Erfahrungen z. B. an der medizinischen Fakultät der Ben-Gurion-Universität in Israel durchgeführt werden (Sacks, 2000, S. 301f.).

Im Weiteren weisen die Ergebnisse der vorliegenden Studie jedoch ebenso einmal mehr darauf hin, dass die Verengung der Diskussion um die Krisensymptome der deutschen Universitätslandschaft auf Eingangsmerkmale der Bewerber alleine nicht zu den erhofften Effekten in der Vorhersagegenauigkeit führen können und werden. Studierendenauswahl, ganz gleich durch welchen Modus durchgeführt, tragen nur soweit, wie Universitäten auch die nötigen Rahmenbedingungen zur Verbesserung der Ausbildung gewährt bekommen und gewillt sind, diese auch zur Verbesserung der Lehre zu nutzen. Potenziell geeignete Bewerber zu finden ist ein Aspekt der Thematik, das Potenzial dieser Bewerber auch angemessen zu fördern, um Studienerfolg mit zu ermöglichen, ein weiterer.

19.2 Zum Einsatz von Persönlichkeitsfragebogen im Selektionskontext

Die Fragestellung, ob die hypothetisch für den Erfolg im Psychologiestudium relevanten Persönlichkeitsmerkmale auch gegenüber experimentell induziertem sozial erwünschtem Antwortverhalten ihre Validität behalten, lässt sich anhand der vorliegenden Ergebnisse klar negativ beantworten. Zum einen zeigten sich nahezu alle Skalen als verfälschbar, wie Mittelwertsanalysen ergaben, wobei sich das Ausmaß der Verfälschbarkeit auch als eine Funktion der Testgeübtheit erwies. Der durch die Instruktion (Normalinstruktion vs. Faking--good-Instruktion) aufgeklärte Varianzanteil in den Fragebogen-Rohwerten fiel bei den Probanden aus dem Hauptstudium Psychologie deutlich höher aus als in der Kohorte aus dem ersten Semester in Psychologie und in der Gesamtstichprobe.

Zum anderen kam es auf Ebene der Kriteriumsvaliditäten unter der Faking-good-Bedingung in beiden Stichproben gegenüber der Normalinstruktion durch ein Absinken der Validitäten aller Skalen zu durchweg insignifikanten und nahe an null liegenden Werten. Dieser Effekt war für einige Skalen über eine Verminderung der Rohwertevarianz durch stereotyperes Antwortverhalten unter der Faking-good-Bedingung zu erklären, für andere Skalen hingegen musste man annehmen, dass es trotz erhaltener Varianz jeweils im Vergleich zur Normalinstruktion hierüber zu „Verwerfungen“ auf Ebene der Konstruktvalidität kam, was über ein Einfließen konstruktirrelevanter Faking-Varianz zu starken Verringerungen der Kriteriumsvaliditäten führte.

Die hier berichteten Ergebnisse stehen somit im Kontrast zu metaanalytisch berichteten Befunden, wonach die Kriteriumsvalidität von Persönlichkeitsfragebögen zwar sinkt, doch aber substanziell bleibt (Viswesvaran & Ones, 1999). Die Ergebnisse und Schlussfolgerungen der vorliegenden Studie hierzu bedürfen daher durchaus kritischer Anmerkungen. Diese betreffen zum einen die jeweils kleinen und irrepräsentativen Stichproben, aus welchen sich nur ungenaue Populationsschätzungen ergeben. Zum anderen stellt das hier gewählte experimentelle Between-subject design ein „worst-case-scenario“ der Verfälschbarkeit dar (s. hierzu z. B. Viswesvaran & Ones, 1999). In wie weit Personen in einer tatsächlichen (Studierenden-)Auswahlsituation verfälschen, kann daher nur grob abgeschätzt werden, zumal *intra*individuelle Antwortvarianz unter Normal- und Faking-good-Bedingungen mit einem Between-subject-design nicht kontrollierbar ist.
Zum anderen bedürfen allerdings auch metaanalytische Befunde einer kritischen Überprüfung. Auch weitgehend erhaltene Validitäten unter Faking-good-Bedingungen treffen keine Aussage darüber, wie Antwortverfälschungen in Richtung sozialer Erwünschtheit Einfluss darauf nehmen, wer zugelassen wird. Zickar et al. (1996, zit. nach Rosse et al., 1998, S. 637) konnten insbesondere zeigen, dass eine derartige Antwortverfälschung sehr wahrscheinlich die Rangordnung der Personenmesswerte besonders im oberen Bereich der Verteilung verändert und sich dies drastischer auswirkt, wenn die Selektionsrate geringer als 50% ist (Rosse et al., 1998, S. 641). Die Wahrscheinlichkeit der Zulassung eigentlich ungeeigneten Kandidaten erhöht sich demnach. Davon unberührt bestehen ohnehin berechtigte Zweifel an der Verlässlichkeit der Metaanalyse als Forschungsmethode an sich, besonders angesichts der File-drawer-Problematik, also der Tendenz, überwiegend signifikante Resultate zu publizieren (Johnson & Eagly, 2000; Vevea & Woods, 2005). Metaanalytische Befunde stellen sich nach neueren Untersuchungen als weitaus instabiler dar, als man gemeinhin annahm (Scargle, 2000; Schönemann & Scargle, 2006, zur Publikation eingereicht). Dass insignifikante Befunde

gerade in diesem Forschungszweig mit seinen im Durchschnitt lediglich niedrigen bis moderaten Kriteriumsvaliditäten alleine auf der Beurteilungsgrundlage der Häufigkeitsverteilung von Validitätskoeffizienten weitaus zahlreicher vorkommen und somit weniger wahrscheinlich publiziert werden, lässt den Einsatz von Persönlichkeitsfragebögen zu Selektionszwecken einmal mehr als fragwürdig erscheinen.

19.3 Ausblick: Inkrementelle Validität von Zulassungstests: Überhaupt das adäquateste Evaluationskriterium?

Der Betrachtung der inkrementellen Validität kommt in der Literatur ein vorrangiger Stellenwert in der Beurteilung des Nutzens von Testverfahren zu (vgl. hierzu z. B. die Literatur und deren Ergebnisse anhand Tabelle 5). Es lässt sich jedoch fragen, ob dieses Evaluationskriterium überhaupt das entscheidende für Zulassungstests darstellt. Nimmt man die Definition von Studierendenauswahlverfahren als Entscheidungsinstrumente über eine Zulassung oder Ablehnung von Bewerbern wörtlich, so scheint es für die Evaluation eines solchen Verfahrens vielmehr weitaus stringenter zu sein, danach zu fragen, wie ein Test die Auswahl an sich beeinflusst. Crouse und Trusheim (1988, S. 139) führen in Bezug auf den SAT dazu aus: „...the important issue for college admissions is not how much the SAT increases a multiple correlation coefficient, but how much the test affects selection“ (S. 139). Man muss also vielmehr danach fragen, mit welcher Wahrscheinlichkeit ein Test korrekte Auswahlentscheidungen trifft, also sowohl potenziell erfolgreiche als auch potenziell nicht erfolgreiche Bewerber identifizieren kann. Und in der Tat werden mitunter derartige Evaluationen eines nunmehr dichotomen Kriteriums (erfolgreich gegenüber nicht erfolgreich) bei festgelegtem Testtrennwert anhand der Taylor-Russel-Tafeln (Taylor & Russel, 1938, 1939) evaluiert (s. hierzu etwa Stemmler, 2005, sowie Hell et al., 2005). Hiebei wird unter Beachtung der Basisrate (Anteil der erfolgreichen Probanden zur Gesamtzahl aller Messwertträger), der Selektionsrate (Anteil der Ausgewählten Bewerber zur Gesamtzahl der Messwertträger) und der Testvalidität bestimmt, wie groß die prozentuale Verbesserung der Vorhersage von später auch erfolgreichen Bewerbern (True positives) durch die zusätzliche Anwendung eines Tests ausfällt. Nun stellt ein solches Vorgehen bereits gegenüber der alleinigen Betrachtung der inkrementellen Validität eine Verbesserung dar, da bei der Testevaluation ebenso die Parameter Basisrate, Selektionsrate und Testvalidität beachtet werden. Allerdings stellt sich die Frage, warum bei einem solchen Vorgehen lediglich die Verbesserung der Vorhersage von True positives beachtet wird und nicht zugleich auch diejenige von potenziell

nicht Erfolgreichen, den True negatives. Dies stellt insofern eine Eigentümlichkeit innerhalb der Psychologie dar, da z. B. jede psychologisch-klinische Studie darum bemüht ist, die Diagnosegüte daran zu messen, wie gut das Vorliegen eine Störung als auch ihr Nicht--Vorliegen erkannt wird, und wie dies ebenso eine der Interpretationsgrundlagen der allerorten durchgeführten Signifikanztestung darstellt. Es gehört jedoch zu den Besonderheiten der Psychodiagnostik, dass deren Verfahren nicht durchweg gemäß Cronbachs Forderung nach „Need for *critical* scrutiny of test validation“ (Cronbach, 1984; Hervorhebung vom Verfasser) untersucht werden.

Man mag nun gegen diese Methode einwenden, zu reinen Selektionszwecken einer Institution reiche die Identifizierung von potenziell Geeigneten aus. Dies mag für die Bewerberauswahl bspw. von Firmen, welche bestrebt sind, ihren eigenen Profit zu steigern auch durchaus berechtigt sein. Doch im Falle von Universitäten als Institutionen mit einem *Ausbildungs-auftrag* erscheint dieses Argument zweifelhaft, da sie einen gesamtgesellschaftlichen Auftrag *auch zu deren Nutzen* erfüllen müssen, und daher auch die Wahrscheinlichkeit von Entscheidungen von fälschlich als ungeeignet klassifizierten Personen möglichst niedrig zu halten ist. Wie Sternberg (1993) es im Zusammenhang mit der GRE ausdrückt:

> I’m concerned about the misses, not the false alarms. The false alarms, those with high GRE scores that can’t do anything else, merely waste some money and some faculty time .... the misses ... will never have the chance to make the contributions to society of which they might be capable. (S. 21)

Eine Verringerung sowohl der Fehlerraten von fälschlich als geeignet als auch von fälschlich als ungeeignet klassifizierten Personen lässt sich aber verständlicherweise nur über eine Betrachtung der Verbesserung *sämtlicher* korrekter Klassifikationen (True positives und True negatives) durch eine Testanwendung analysieren.

Aus diesem Grunde sind auch Kosten-Nutzen-Rechnungen zu Studierendenauswahlverfahren, welche lediglich auf die Einsparungen seitens der Universitäten über die Verbesserung der Vorhersage von später auch erfolgreichen Bewerbern berechnen, zu kurz gegriffen, da sie die geschätzten Folgekosten aus einer falsch negativen Klassifikation unbeachtet lassen (s. z. B. Schmidt-Atzert & Krumm, 2006).

Alles, was man für eine solche Testvalidierung tun müsste, wäre, den Test einer Zufalls-stichprobe aus dem Bewerberpool vorzugeben und alle Bewerber unabhängig vom Tester-

gebnis zuzulassen. Man würde somit die Basis für eine Testevaluation erhalten, die sowohl die Bestimmung der True positives als auch der True negatives ermöglichte und deren Kriteriumskorrelationen nicht mit fragwürdigen Selektionskorrekturen aufgewertet werden müssten (s. hierzu Kap. 4.2.3.4). Ein derartiges Stichprobendesign war bereits Grundlage für die vorbildhafte Entwicklung des Tests für medizinische Studiengänge (s. hierzu Deidesheimer Kreis, 1997) und wäre demnach prinzipiell auch für andere Studierfähigkeitstests zu fordern.

Methoden zur Berechnung einer Verbesserungsrate korrekter Klassifikationen bei verschiedenen Basis- und Selektionsraten und Testvaliditäten finden sich bei Schönemann (1997b) sowie Schönemann und Thompson (1996). Das Problem, das sich allerdings aus der Betrachtung korrekter Klassifikationen für die Testdiagnostik ergibt, ist, „...that, for the realistic validity range ($r \leq .5$), use of a test ensures an increase in correct decisions uniformly only for base rates near .5 .... For base rates outside the (.3, .7) range, no test near the modal validity .3 improves over random admissions, regardless which quota is used“ (Schönemann, 1997b, S. 184). Ruft man sich in diesem Zusammenhang noch einmal die fast durchweg in diesem Validitätsbereich liegenden Koeffizienten von Zulassungstests in Erinnerung (vgl. z. B. Tabelle 5, sowie Hell. et al., 2005) und bedenkt zudem, dass durch die gesetzlich vorgeschriebene Beachtung der Abiturdurchschnittsnote bei der Auswahl von Studierenden zwangsläufig ein Basisratenproblem entsteht, so scheint in der Tat die zusätzliche Verwendung von psychometrischen Tests bei weitem *nicht* ausreichend, um eine Verbesserung der Zulassungsentscheidungen im Sinne korrekter Klassifikationen zu gewährleisten. Dies verweist wiederum darauf, dass eine wirklich effiziente Studierendenauswahl ein Unterfangen ist, welches größerer Anstrengungen in Form sequenzieller Auswahlstrategien erfordert, soll die Gesamtvalidität des Verfahrens tatsächlich so hoch liegen, dass eine Verbesserung korrekter Entscheidungen möglich ist (zu Berechnungen von Mindestvaliditäten unter Beachtung verschiedenen Basisraten und Selektionsquoten s. Schönemann, 1997b). Konzeptionen solcher sequenzieller Auswahlstrategien mit internetbasierten Informationssystemen über Self-Assessments bis hin zur Kombination aus Abiturleistungen und Testergebnissen liegen bereits vor (s. insbesondere Hornke & Zimmerhofer, 2005, sowie Amelang & Funke, 2005). Dass diese aufwendiger und kostspieliger sind, als lediglich ein zusätzlicher Einsatz von Studierfähigkeitstests neben der Abiturdurchschnittsnote, ist unmittelbar einsichtig. Sicher mag die Durchführung sequenzieller Auswahlverfahren für manche Universitäten unrealistisch erscheinen. Alleine von Studierfähigkeitstests zu erwarten, dass sie die Probleme der Studierfähigkeitsdiagnostik lösen können, jedoch ebenso.

Gerade zum letztgenannten Punkt bedürfte es nach Meinung des Autors intensiverer kritischer fachinterner Diskussionen über den Nutzen von Studierfähigkeitstests unter weiterführenden Aspekten als lediglich demjenigen der inkrementellen Validität, damit sich die Testdiagnostik in Anlehnung an ein Zitat von Minton (1988, S. 74) später nicht vorwerfen lassen muss: While testing may not have made a significant contribution to the student admission process, the student admission process had made a significant contribution to testing.

20. Zusammenfassung

Die vorliegende Studie befasste sich mit der Konstruktion und Evaluation eines fachspezifischen Studierendenauswahlverfahrens für das Fach Psychologie an der Universität Heidelberg. Zunächst wurde ein Überblick über bisherige Befunde zu Studierendenauswahlverfahren im internationalen und nationalen Kontext gegeben.
Der empirische Teil der Arbeit basierte auf den Ergebnissen einer Anforderungsanalyse, in der Personenmerkmale abgeleitet wurden, welche für das erfolgreiche Absolvieren des Psychologiestudiums an der Universität Heidelberg hypothetisch unabdingbar sind. Hieraus resultierten Komponenten der kognitiven Leistungsfähigkeit zu verbaler, numerischer und fluider Intelligenz, zur Kreativität sowie zum erfahrungswissenschaftlichen, empiriebezogenen Denken. Als hypothetisch erfolgskritische Persönlichkeitsvariablen ergaben sich Leistungsmotivation, hartnäckige Zielverfolgung und flexible Zielanpassung, Gewissenhaftigkeit, emotionale Stabilität und Offenheit für Erfahrungen. Mit Ausnahme der Aufgaben zur Kreativität und der Persönlichkeitsvariablen wurden vom Autor Testverfahren konstruiert, im Übrigen Standardverfahren eingesetzt. Das Gesamtverfahren wurde anhand der Testergebnisse von $N = 434$ Testteilnehmern aus verschiedenen Studienfächern an der Universität Heidelberg und Substichproben aus dem Studienfach Psychologie konstruiert und Kriteriumsvalidierungen an Studierendenstichproben aus dem Fach Psychologie vorgenommen.

Die für diese Studie zentrale Hypothese war, dass durch die zusätzliche Beachtung der Testergebnisse neben der Abiturdurchschnittsnote eine substanzielle Verbesserung in der Vorhersagegenauigkeit von Studienleistungen in Form von Studiennoten zu erreichen ist (inkrementelle Validität durch Testverfahren).
Weiterhin wurde untersucht, in wie weit sich Kriteriumsvaliditäten von Daten aus Persönlichkeitsfragebögen, welche hypothetisch erfolgsrelevante Personenmerkmale erfassen, gegenüber einer im Selektionskontext sehr wahrscheinlichen Antwortverfälschung als robust erweisen. Hierzu war die Hälfte der Persönlichkeitsfragebögen mit einer Instruktion versehen, nach

welcher die Studienteilnehmer sich ihrer Meinung nach möglichst positiv in den Itemantworten darstellen sollten (Faking-good-Instruktion), die andere Hälfte der Persönlichkeitsfragebögen hingegen mit einer Normalinstruktion.
Auch wurde die studienfachspezifische Differenzierungsfähigkeit der Testbatterie in den Subtests zu kognitiven Leistungen über Mittelwertsvergleiche zwischen Studienteilnehmern aus dem ersten Semester verschiedener Fächer analysiert. Hierbei sollte sich nach der Hypothese ein für Psychologiestudierende typisches Leistungsprofil ergeben, das sich auch unabhängig vom Einfluss der Abiturdurchschnittsnote zeigen musste, wenn die Testbatterie fachspezifische Fähigkeiten zu erfassen vermag.

Hinsichtlich der fachspezifischen Differenzierungsfähigkeit konnte die Testbatterie lediglich im Subtest zu Kreativitätsfacetten und einem Test zur Erfassung empiriebezogenen Denkens im Sinne der Hypothese zwischen Teilnehmern aus den ersten Semestern verschiedener Fächer unterscheiden: Psychologiestudierende zeigten hierin im Durchschnitt bessere Leistungen gegenüber allen verglichenen Fächern. Allerdings waren nach statistischer Kontrolle der Abiturdurchschnittsnote nur noch die Hälfte der vorher identifizierbaren Mittelwertsdifferenzen vorhanden, sodass auch diese Variablen lediglich eine befriedigende fachspezifische Differenzierungsfähigkeit aufweisen.

Die Ergebnisse zur inkrementellen Validität der Testverfahren kognitiver Variablen zeigten zunächst, dass insbesondere die Abiturdurchschnittsnote einen vergleichsweise hohen und stabilen prädiktiven Vorhersagebeitrag bezüglich der durchschnittlichen Vordiplomnote als auch in Ergebnissen einer Orientierungsprüfungsklausur und zweier Klausuren im Fach Methodenlehre aus dem Grundstudium aufwies. Die Testergebnisse korrelierten ebenso in der erwarteten Richtung, wobei unterschiedliche Prädiktionsbeiträge der Subtest über die verschiedenen Kriterien zu beobachten waren. Gleichwohl erwiesen sich schulische Leistungsfähigkeitsmaße, allen voran die Abiturnote, in simultanen Analysen aller Prädiktoren über eine kanonische Korrelation als vergleichsweise beste Prädiktoren.

Die Hinzunahme der Testergebnisse zur Abiturdurchschnittsnote in multiplen Regressionen als zentrale Analyse der inkrementellen Validität erhöhte die prädiktive Güte bei drei von vier Kriterien (Vordiplomnote sowie Leistungen in den zwei Methodenlehre-Klausuren). Die Zugewinne an Prädiktionskraft lagen hierbei im Rahmen dessen, was aus internationalen Studien zu inkrementellen Validitäten bislang bekannt ist. Jedoch erbrachte jeweils nur ein einziger und jedes Mal ein anderer Subtest aus der Testbatterie eine Verbesserung der Vorher-

sage. Zudem erwiesen sich die Regressionsschätzungen in Kreuzvalidierungsanalysen als instabil, sodass generalisierte Aussagen zum Inkrement über die hier verwendeten Testverfahren nur eingeschränkt möglich sind.

Die Analysen zur Validität von Fragebögen unter experimentell induzierter Antwortverfälschung zeichneten ein pessimistisches Bild für deren Einsetzbarkeit im Selektionskontext. Allgemein ließen sich Effekte der Antwortverfälschung nachweisen, was sich insbesondere in Mittelwertsunterschieden zwischen den Instruktionsversionen auf den Persönlichkeitsskalen zeigte. Dieser Effekt war dabei umso stärker ausgeprägt, je erfahrener die jeweilige Stichprobe mit Persönlichkkeitsfragebögen war. Psychologiestudierende aus dem Hauptstudium waren im Vergleich zu Studierenden anderer Fächer am deutlichsten befähigt, die Antworten zu verfälschen.
Auf Ebene der Kriteriumsvalidität zeigte sich hauptbefundlich, dass unter der Normalinstruktion beobachtbare Kriteriumsvaliditäten unter der Faking-good-Instruktion bis zur Insignifikanz abfielen. Dieser Effekt war überwiegend nicht über korrelationsmindernde Varianzunterschiede zwischen den Skalenrohwerten beider Instruktionsversionen erklärbar, da sich diese in den wenigsten Fällen nachweisen ließ. Kriteriumsirrelevante Faktoren der sozialen Erwünschtheit scheinen hier die eigentlich zu messenden Dimensionen zu invalidieren. Das Ergebnis legt somit den Schluss nahe, Persönlichkeitsfragebögen nicht im Selektionskontext von Studierendenauswahlverfahren einzusetzen.

Insgesamt bestätigen die im Fokus dieser Arbeit stehenden Ergebnisse zur Validität von Abiturdurchschnittsnote und Testleistungen bisherige Befunde, welche die Abiturnote als besten Prädiktor für universitäre Leistungen ausweisen. Inkrementelle Validitäten sind durch Testverfahren zu erzielen, jedoch sollten diese, um einen praktisch bedeutsamen Vorhersagebeitrag im Sinne korrekter Klassifikationen (True Positives und True Negatives) zur Studienerfolgsprognose zu leisten, in einem sequenziellen Auswahlverfahren integriert werden, welches aus einem internetbasierten Selbstinformationssystem und darauf aufbauenden Selektionsschritten unter Beachtung der Abiturdurchschnittsnote und Testleistungen besteht. Entsprechende Konzeptualisierungen für die Universität Heidelberg liegen bereits vor (Amelang & Funke, 2005). Wegen der Aggregierung verschiedener Informationen über die Studieneignung eines Probanden verspricht ein solches Verfahren in seiner Gesamtheit noch die besten Aussichten auf eine allgemein notwendige Verbesserung in der Vorhersage der Studienleistung.

21. LITERATURVERZEICHNIS

ACT Incorporated. (2005). Description of the ACT: ACT Incorporated. http://www.actstudent.org/testprep/descriptions/

Adams, E. & Fagot, R. (1959). A model of riskless choice. *Behavioral Science, 4*, 1-10.

Amelang, M. (1974). Ausbildung und Prüfung im Fach Psychologie – Versuch einer Bewertung angestellt am Beispiel eines Institutes. *Psychologische Rundschau, 25*, 169-182.

Amelang, M. (1975). *Validierung von Anforderungsprofilen für das Studium der Medizin, Zahnmedizin, Pharmazie und Psychologie* (Vorläufiger Abschlussbericht eines Forschungsprojektes). Hamburg: Psychologisches Institut II der Universität Hamburg.

Amelang, M. (1978). Der Hochschulzugang. In T. Herrmann (Hrsg.), *Hochschulentwicklung. Aufgaben und Chancen* (S. 88-105). Heidelberg: Asanger.

Amelang, M. (1997). Differentielle Aspekte der Hochschulzulassung: Probleme, Befunde, Lösungen. In T. Herrmann (Hrsg.), *Hochschulentwicklung. Aufgaben und Chancen*. Heidelberg: Asanger.

Amelang, M. & Bartussek, D. (1970). Research on validity of a new lie scale. *Diagnostica, 16*(3), 103-122.

Amelang, M. & Bartussek, D. (1997). *Differentielle Psychologie und Persönlichkeitsforschung* (4. Auflage). Stuttgart: Kohlhammer.

Amelang, M. & Funke, J. (2005). Entwicklung und Implementierung eines kombinierten Beratungs- und Auswahlverfahrens für die wichtigsten Studiengänge an der Universität Heidelberg. *Psychologische Rundschau, 56*(2), 135-137.

Amelang, M. & Hoppensack, T. (1977). Persönlichkeitsstruktur und Hochschulbesuch. Vorhersage des Studienerfolges bei Studierenden verschiedener Fachrichtungen. *Psychologie in Erziehung und Unterricht, 24*, 193-204.

Amelang, M., Schäfer, A. & Yousfi, S. (2002). Comparing verbal and non-verbal personality scales: Investigating the reliability and validity, the influence of social desirability, and the effects of fake good instructions. *Psychologische Beitrage, 44*, 24-41.

Amelang, M. & Zielinski, W. (1997). *Psychologische Diagnostik und Intervention* (2. Auflage). Berlin: Springer.

Amthauer, R., Brocke, B., Liepmann, D. & Beauducel, A. (2001). I-S-T 2000 R Intelligenz-Struktur-Test 2000 R Intelligence Structure Test 2000 (revised)/zpid Synonym(e): IST 2000 R.

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38*(1), 123-140.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573.

Andrich, D. (2004). Controversy and the Rasch Model: A Characteristic of Incompatible Paradigms? *Medical Care, 42*(1; Supplement:I-7).

Baird, L. L. (1985). Do grades and tests predict adult accomplishment? *Research in Higher Education, 23*(1), 3-85.

Bak, P. M. & Brandtstädter, J. (1998). Flexible Zielanpassung und hartnäckige Zielverfolgung als Bewältigungsressourcen: Hinweise auf ein Regulationsdilemma

Flexible goal adjustment and tenacious goal pursuit as coping resources: Hints to a regulatory dilemma. *Zeitschrift für Psychologie, 206*(3), 235-249.

Baron-Boldt, J. (1989). *Die Validität von Schulabschlussnoten für die Prognose von Ausbildungs- und Studienerfolg. Eine Metaanalyse nach dem Prinzip der Validitätsgenerierung* (Bd. 280). Frankfurt am Main: Lang.

Baron-Boldt, J., Schuler, H. & Funke, U. (1988). Prädiktive Validität von Schulabschlußnoten: Eine Metaanalyse. *Zeitschrift für Pädagogische Psychologie, 2*(2), 79-90.

Baron, J. & Norman, M. F. (1992). SATs, achievement tests, and high-school class rank as predictors of college performance. *Educational & Psychological Measurement, 52*(4), 1047-1055.

Barrick, M. R. & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*(1), 1-26.

Barrick, M. R., Mount, M. K. & Strauss, J. P. (1993). Conscientiousness and performance of sales representatives: Test of the mediating effects of goal setting. *Journal of Applied Psychology, 78*(5), 715-722.

Baumert, J. & Watermann, R. (2000). Institutionelle und regionale Variabilität und die Sicherung gemeinsamer Standards in der gymnasialen Oberstufe. In J. Baumert, W. Bos & R. Lehmann (Hrsg.), *TIMSS/III. Band 2: Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe* (Bd. 2, S. 317-371). Opladen: Leske & Budrich.

Bell, P., Staines, P. & Mitchell, J. (2001). *Evaluating, doing and writing research in psychology: A step-by-step guide for students*. Thousand Oaks, CA, US: Sage Publications, Inc.

Beller, M. (1993). Zulassungsverfahren an israelischen Universitäten: Psychometrische und soziale Betrachtungen. In G. Trost, K. Ingenkamp & R. S. Jäger (Hrsg.), *Tests und Trends. 10. Jahrbuch der pädagogischen Diagnostik* (S. 115-142). Weinheim: Beltz.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Hrsg.), *Statistical Theories of Mental Test Scores.* (S. 397–545). Reading, Massachusets: Addison-Wesley.

Bock, R. D. & Jones, L. V. (1968). *The Measurement and Prediction of Judgment and Choice.* San Francisco: Holden Day.

Bolt, D. M., Cohen, A. S. & Wollack, J. A. (2002). Item parameter estimation under conditions of test speediness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement, 39*(4), 331-348.

Bolwahnn, B. L. (2003). Development and validation of a critical incident based rating instrument for psychology graduate student selection. *Dissertation Abstracts International: Section B: The Sciences & Engineering, 63*(7), 3502.

Bond, T. G. & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah: Erlbaum.

Borkenau, P. & Ostendorf, F. (1993). *NEO-FFI NEO-Fünf-Faktoren Inventar nach Costa und MCCrae - Deutsche Fassung*. Neo Five Factor Inventory (Costa, P.T. & McCrae, R.R., 1985).

Borsboom, D., Mellenbergh, G. J. & van Heerden, J. (2004). The Concept of Validity. *Psychological Review, 111*(4), 1061-1071.

Bortz, J. (1999). *Statistik für Sozialwissenschaftler* (5. Auflage). Berlin: Springer.

Bortz, J. & Döring, N. (1995). *Forschungsmethoden und Evaluation* (2., vollständig überarbeitete und aktualisierte Auflage). Berlin: Springer.

Bortz, J. & Lienert, G. A. (2003). *Kurzgefaßte Statistik für die klinische Forschung*. Berlin: Springer.

Bowen, C.-C., Martin, B. A. & Hunt, S. T. (2002). A comparison of ipsative and normative approaches for ability to control faking in personality questionnaires. *International Journal of Organizational Analysis, 10*(3), 240-259.

Bozdogan, H. (1987). Model selection for Akaike's information criterion (AIC). *Psychometrika, 53*(3), 345-370.

Brandstädter, J. (1984). Personal and social control over development: Some implications of an action perspective in life-span developmental psychology. In W. Strobe & M. Hewstone (Hrsg.), *European review of social psychology* (Bd. 6, S. 33-68). Chichester: Wiley.

Brandstädter, J. (Hrsg.). (2002). *Protective processes in later life: Maintaining and revising personal goals*. Taylor & Frances/Routledge.

Brandstädter, J. & Rothermund, K. (2002). The life-course dynamics of goal pursuit and goal adjustment: a two-process framework. *Developmental Review, 22*, 117-150.

Brandtstädter, J. & Renner, G. (1988). TEN/FLEX Fragebogen zur Erfassung von Flexibilitaet der Zielanpassung und Tenazitaet der Zielverfolgung (PSYNDEX Tests-Kurznachweis) Tenacious Goal Pursuit (TGP) and Flexible Goal Adjustment (FGA)/zpid Synonym(e): Fragebogen zum Zielverfolgungsverhalten; Fragen zum Umgang mit Problemen; HZ/FZ; HZV/FZA; TenFlex.

Bridgeman, B., Burton, N. & Cline, F. (2001). *Substituting SAT II: Subject tests for SAT I: Reasoning test: Impact on admitted class composition and quality* (Research Report No. 2001-3). New York: The College Board. http://www.collegeboard.com/research/pdf/rdreport200_3920.pdf

Bridgeman, B., Burton, N. & Cline, F. (2004). Replacing reasoning tests with achievement tests in university admissions: Does it make a difference? In R. Zwick (Hrsg.), *Rethinking the SAT: The future of standardized admission testing* (S. 277-288). New York: Routledge Farmer.

Bridgeman, B., McCamley-Jenkins, L. & Ervin, N. (2000). *Prediction of freshman grade point average from the revised and the recentered SAT I: Reasoning Test* (Research Report No. 2000-1): Educational Testing Service. http://www.collegeboard.com/research/pdf/rr0001_3917.pdf

Bultmann, T. (2001). Vom öffentlichen Bildungsauftrag zur privaten Dienstleistung. Hochschulpolitische Wende in Deutschland. In B. Hoff & P. Sitte (Hrsg.), *Politikwechsel der Wissenschaftspolitik.* Berlin: Dietz.

Burton, N. W. & Ramist, L. (2001). *Predicting success in college: SAT studies of classes graduating since 1980* (Research Report No. 2001-2): Educational Testing Service. http://www.collegeboard.com/research/pdf/rdreport200_3919.pdf

Burton, N. W. & Turner, N. (1983). *Effectiveness of Graduate Record Examinations for predicting first-year grades: 1981-82 summary report of the Graduate Record Examinations validity Study Service*. Princeton. NJ: Eduactional Testing Service.

Burton, N. W. & Wang, M.-m. (2005). *Predicting Long-Term Success in Graduate School: A Collaborative Validity Study* (No. RR-05-03). Princeton N.J.: Educational Testing Service. http://www.ets.org/Media/Research/pdf/RR-05-03.pdf

Camara, W. J. (2005). Broadening Criteria of College Success and the Impact of Cognitive Predictors. In W. J. Camara & E. W. Kimmel (Hrsg.), *Choosing students: Higher education admissions tools for the 21st century* (S. 53-79). Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Campbell, J. P. (1976). Psychometric Theory. In M. Dunnette (Hrsg.), *Handbook of Industrial and Organizational Psychology*. Chicago: Rand McNally.

Campbell, N. R. (1920). *Physics: The Elements*. Cambridge: University Press.

Campbell, N. R. (1921/1952). *What is Science?* New York: Dover.

Campbell, N. R. (1940). Physics and psychology. *British Association for the Advancement of Science, 2*, 347-348.

Chamorro-Premuzic, T. & Furnham, A. (2003a). Personality predicts academic performance: Evidence from two longitudinal university samples. *Journal of Research in Personality, 37*(4), 319-338.

Chamorro-Premuzic, T. & Furnham, A. (2003b). Personality traits and academic examination performance. *European Journal of Personality, 17*(3), 237-250.

Cliff, N. (1983). Some Cautions concerning the Application of causal Modeling Methods. *Mulivariate Behavioral Research, 18*, 115-126.

Cohen, J., Cohen, P., West, S. G. & Aiken, L. S. (2002). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (3. Auflage): Lawrence Erlbaum.

CollegeBoard. (2005). SAT Reasoning test: The College Board.

Cook, T. D. & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton Mifflin.

Costa, P. T. & McCrae, R. R. (1992). Normal personality assessment in clinical practice: The NEO Personality Inventory. *Psychological Assessment, 4*(1), 5-13.

Cowles, M., Darling, M. & Skanes, A. (1992). Some characteristics of the simulated self. *Personality and Individual Differences, 13*(5), 501-510.

Cox, D. R. & Hinckley, D. V. (1974). *Theoretical Statistics*. New York: Chapman & Hall.

Cronbach, L. J. (1982). Prudent aspiration for social inquiry. In W. H. Kruskal (Hrsg.), *The social sciences: Their nature and uses*. Chicago: University of Chicago Press.

Cronbach, L. J. (1984). *Essentials of psychological testing* (4. Auflage). New York: Harper & Row.

Crouse, J. & Trusheim, D. (1988). *The Case against the SAT*. Chicago: University of Chicago Press.

Daniel, H.-D. (1996). Korrelate der Fachstudiendauer von Betriebswirten: Ergebnisse einer Absolventenbefragung an der Universität Mannheim. *Zeitschrift für Betriebswirtschaft, 1*(E), 95-115.

Davier, M. v. & Rost, J. (1995). Polytomous mixed Rasch models. In G. H. Fischer &. I. W. Molenaar (Hrsg.), *Rasch model. Foundations, recent developments, and applications.* (S. 371-379). New York: Springer.

Deidesheimer Kreis. (1997). *Hochschulzulassung und Studieneingangstests: Studienfeldbezogene Verfahren zu Feststellung der Eignung für Numerus-clausus- und andere Studiengänge*. Göttingen, Zürich: Vandenhoeck & Ruprecht.

Dellaert, F. (2002). *The expectation maximization algorithm* (Technical Report No. GIT-GVU-02-20). Georgia: Institute of Technology. http://www-static.cc.gatech.edu/~dellaert/em-paper.pdf

Derous, E. & Born, M. P. (2005). Impact of face validity and information about the assessment process on test motivation and performance/Impact de la relation entre la validite apparente et l'information relative au processus de recrutement sur le test de motivation et la performance. *Travail Humain, 68*(4), 317-336.

Deutsches Institut für Normung. (2002). *DIN 33430: Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen*. Göttingen: Vandenhoeck & Ruprecht.

Dierkes, M. & Merkens, H. (2002). *Zur Wettbewerbsfähigkeit des Hochschulsystems in Deutschland*. Berlin: Wissenschaftszentrum Berlin für Sozialforschung.

Douglas, A. R. (2005). An investigation about whether restrictive time limits violate the fundamental assumptions of item response models. *Dissertation Abstracts International Section A: Humanities and Social Sciences, 65*(9), 3275.

Düchting, W., Ulmer, W. & Ginsburg, T. (1996). Cancer: a challenge for control theory and computer modelling. *European Journal of Cancer, 32A*, 1283-1292.

Eckes, T. (2005a). Evaluation von Beurteilungen: Psychometrische Qualitätssicherung mit dem Multifacetten-Rasch-Modell. Evaluation of ratings: Psychometric quality assurance via many-facet Rasch measurement. *Zeitschrift für Psychologie, 213*(2), 77-96.

Eckes, T. (2005b). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet rasch analysis. *Language Assessment Quarterly, 2*(3), 197-221.

Educational Testing Service. (2005). GRE General Test: Educational Testing Service.

Efron, B. & Tibshirani, R. (1985). The bootstrap method for assessing statistical accuracy. *Behaviormetrika, 17*, 1-35.

Egeln, J. & Heine, C. (2005). Die Ausbildungsleistungen der Hochschulen. Eine international vergleichende Analyse im Rahmen des Berichtssystems zur Technologischen Leistungsfähigkeit Deutschlands: HIS.

Eid, M. & Rauber, M. (2000). Detecting measurement invariance in organizationale surveys. *European Journal of Psychological Assessment, 16*(1), 20-30.

Ellingson, J. E., Smith, D. B. & Sackett, P. R. (2001). Investigating the influence of social desirability on personality factor structure. *Journal of Applied Psychology, 86*(1), 122-133.

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*(2), 93-112.

Engelhard, G., Jr. (1992). The measurement of writing ability with a Many-Faceted Rasch model. *Applied Measurement in Education, 5*(3), 171-191.

Engelhard, G., Jr. (1996a). Clarification to 'Examining rater errors in the assessment of written composition with a many-faceted Rasch model'. *Journal of Educational Measurement, 33*(1), 115-116.

Engelhard, G., Jr. (1996b). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement, 33*(1), 56-70.

Engelhard, G., Jr. (Hrsg.). (2002). *Monitoring raters in performance assessments*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational & Psychological Measurement, 58*(3), 357-381.

Farsides, T. & Woodfield, R. (2003). Individual differences and undergraduate academic success: The roles of personality, intelligence, and application. *Personality and Individual Differences, 34*(7), 1225-1243.

Fechner, G. (1987). Outline of a new principle of mathematical psychology (1851). *Psychologische Forschung, 49*(4), 203-207.

Fedrowitz, J. & Zempel, F. (1996). Bestandsaufnahme: Studierfähigkeit, Hochschulzugang und rechtliche Handlungsspielräume. In H. J. Meyer & D. Müller-Böling (Hrsg.), *Hochschulzugang in Deutschland. Status quo und Perspektiven* (S. 85-100). Gütersloh: Bertelsmann.

Ferguson, A., Meyers, R. J., Bartlett, H., Banister, F. C., Bartlet, W., Brown, N. R., et al. (1940). Quantitative estimates of sensory events: final report of the committee appointed to consider and report upon the possibility of quantitative estimates of sensory events. *Advancement of Science, 1*, 334-349.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, Bd. 37*(6), 359-374.

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests*. Bern Stuttgart Wien: Hans Huber.

Fisher, W. (2004). Meaning and method in the social sciences. *Human Studies, 30*, 1-26.

Fitzpatrick, A. R., Ercikan, K., Yen, W. M. & Ferrara, S. (1998). The consistency between raters scoring in different test years. *Applied Measurement in Education, 11*, 195-208.

Flanagan, J. (1954). The critical incident technique. *Psychological Bulletin, 51*, 327-358.

Flender, J., Christmann, U. & Groeben, N. (1999). Entwicklung und erste Validierung einer Skala zur Erfassung der passiven argumentativ-rhetorischen Kompetenz. *Zeitschrift für Differentielle und Diagnostische Psychologie, 20*(4), 309-325.

Flender, J., Christmann, U., Groeben, N. & Mlynski, G. (1996). *Argumentatiosnintegrität (XVII): Entwicklung und erste Validierung einer Skala zur Erfassung der passiven argumentativ-rhetorischen Kompetenz (SPARK)*. Heidelberg/Mannheim: Psychologisches Institut der Universität Heidelberg.

Forman, A. K. (1973). *Die Konstruktion eines neuen Matrizentests und die Untersuchung des Lösungsverhaltens mit Hilfe eines Linearen Logistischen Testmodells.*, Universität Wien.

Frey, M. C. & Detterman, D. K. (2004a). Scholastic assessment or g? The relationship between the scholastic assessment test and general cognitive ability. *Psychological Science, 15*(6), 373-378.

Frey, M. C. & Detterman, D. K. (2004b). Scholastic Assessment or g? The Relationship Between the Scholastic Assessment Test and General Cognitive Ability: Erratum. *Psychological Science, 15*(9), 641.

Frey, M. C. & Detterman, D. K. (2005). Regression Basics: Rejoinder to Bridgeman. *Psychological Science, 16*(9), 747.

Friday, A. S. (2005). Criterion-related validity of Big Five adolescent personality traits. *Dissertation Abstracts International: Section B: The Sciences and Engineering, 65*(9), 4885.

Furnham, A. (1986). Response bias, social desirability and dissimulation. *Personality and Individual Differences, 7*(3), 385-400.

Gallini, J. K. (1983). A Rasch analysis of Raven item data. *Journal of Experimental Education, 52*(1), 27-32.

Garson, D. (2005). Canonical Correlation. In *Statnotes: Topics in Multivariate Analysis.* http://www2.chass.ncsu.edu/garson/pa765/canonic.htm.

Gasch, B. (1971). *Erfolg im Psychologiestudium* (Bd. 21). Meisenheim am Glan.

Geiser, S. & Studley, R. (2002). UC and the SAT: Predictive Validity and Differential Impact of the SAT I and SAT II at the University of California. *Educational Assessment, 8*(1), 1-26.

Giesen, H. & Gold, A. (1996). Individuelle Determinanten der Studiendauer. Ergebnisse einer Längsschnittuntersuchung. In J. Lompscher & H. Mandl (Hrsg.), *Lehr- und Lernprobleme im Studium* (S. 86-99). Bern: Huber.

Giesen, H., Gold, A., Hummer, A. & Jansen, R. (1986). *Prognose des Studienerfolgs. Ergebnisse aus Längsschnittuntersuchungen. Schlussbericht zu dem mit Mitteln des Bundesministeriums für Bildung und Wissenschaft geförderten Forschungsprojekt „Längsschnittuntersuchungen zur Beobachtung und Analyse von Bildungslebensläufen“* (Schlussbericht). Frankfurt am Main: Universität Frankfurt.

Glas, C. A. W. & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Hrsg.), *Rasch models - Foundations, recent developments and applications* (S. 69-95). New York: Springer.

Gold, A. & Souvignier, E. (2005). Prognose der Studierfähigkeit. Ergebnisse aus Längschnittanalysen. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 37*(4), 214-222.

Goldberg, E. L. & Alliger, G. M. (1992). Assessing the validity of the GRE for students in psychology: A validity generalization approach. *Educational & Psychological Measurement, 52*(4), 1019-1027.

Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models. *British Journal of Mathematical and Statistical Psychology, 33*(2), 234-246.

Gordon, M. E. & Gross, R. H. (1978). A Critique of Methods for Operationalizing the Concept of Fakeability. *Educational & Psychological Measurement, 38*, 811-818.

Griffith, R. (1998). Faking of noncognitive selection devices: Red herring is hard to swallow. *Dissertation Abstracts International: Section B: The Sciences & Engineering, 58*(10), 5681.

Guilford, J. P. (1954). Psychometric methods (2. Auflage). New York: McGraw-Hill.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer (Hrsg.), *The American soldier. Studies in social psychology in World War II*. Princeton: Princeton University Press.

Guttman, L. (1977). What is not what in statistics. *The Statistican, 26*(2), 81-107.

Häcker, H. & Stapf, K. H. (1998). *Dorsch. Psychologisches Wörterbuch* (Bd. 13). Bern: Huber.

Hair, E. C. & Graziano, W. G. (2003). Self-esteem, personality and achievement in high school: A prospective longitudinal study in Texas. *Journal of Personality, 71*(6), 971-994.

Hänsgen, K.-D. (2006). *Eignungstest für das Medizinstudium (EMS)*. Vortrag an der Medizinischen Universität Wien, 4.5.2006. http://www.unifr.ch/ztd/ems/emswieninfo.pdf

Heckman, J. (1976). The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Sample Estimator of such Models. *Annals of Economics and Social Measurement, 5*.

Heckman, J. (1979). Sample Selection Bias as a Specification Error. *Econometrica, 47*.

Heene, M. (2003). Der Fragebogen zur Kompetenz- und Kontrollüberzeugung aus der Sicht der probabilistischen Testtheorie. Unveröffentlichte Diplomarbeit: Universität Trier.

Heene, M. & Funke, J. (2006). Why person homogeneity must be tested: A reanalysis of the TAI. *Submitted for publication*.

Heldmann, W. (1984). *Studierfähigkeit. Ergebnisse einer Umfrage*. Göttingen: Schwartz.

Hell, B., Trapmann, S., Weigand, S., Hirn, J. O. & Schuler, H. (2005). *Die Validität von Prädiktoren des Studienerfolgs – eine Metaanalyse*. Vortrag auf der 4. Tagung der Fachgruppe Arbeits- und Organisationspsychologie der Deutschen Gesellschaft für Psychologie in Bonn, 19.09.-21.09.2005. http://www.uni-hohenheim.de/studieneignung/publikationen/metaanalyse_verfahren_ao_2005.pdf

Hitpass, J. (1975). *Tests im "Besonderen Auswahlverfahren" für die Hochschulzulassung*. Opladen: Westdeutscher Verlag.

Hitpass, J., Ohlsson, R. & Thomas, E. (1984). *Studien- und Berufserfolg von Hochschulabsolventen mit unterschiedlichen Studieneingangsvoraussetzungen*. Opladen: Westdeutscher Verlag.

Hochschulrektorenkonferenz. (2004). Im Brennpunkt: Das Hochschul-Zulassungsrecht. http://www.hrk.de/de/brennpunkte/1946.php

Holland, P. W. & Wainer, H. (1993). *Differential Item Functioning*. Hilsdale, New Jersey: Erlbaum.

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics, 6*, 65-70.

Höppel, D. & Moser, K. (1993). Die Prognostizierbarkeit von Studiennoten und Studiendauer durch Schulabschlussnoten. *Zeitschrift für Pädagogische Psychologie, 7*(1), 25-32.

Hörner, W. (1999). Studienerfolgs- und Studienabbruchquoten im internationalen Vergleich. In M. Schröder-Gronostay & H.-D. Daniel (Hrsg.), *Studienerfolg und Studienabbruch. Beiträge aus Forschung und Praxis*. Neuwied: Luchterhand.

Hornke, L. F. & Habon, M. W. (1986). Rule-based item bank construction and evaluation within the linear logistic framework. *Applied Psychological Measurement, 10*(4), 369-380.

Hornke, L. F. & Rettig, K. (1988). Regelgeleitete Itemkonstruktion unter Zuhilfenahme kognitionspsychologischer Überlegungen. In K. D. Kubinger (Hrsg.), *Moderne Testtheorie* (S. 140-162). Weinheim: Beltz.

Hornke, L. F. & Zimmerhofer, A. (2005). Berufseignungsdiagnostische Entscheidungen. Zur Bewährung eignungsdiagnostischer Ansätze. *Psychologische Rundschau, 56*(2), 146-148.

Hosenfeld, I., Strauss, B. & Köller, O. (1997). Geschlechtsdifferenzen bei Raumvorstellungsaufgaben--eine Frage der Strategie? *Zeitschrift für Pädagogische Psychologie, 11*(2), 85-94.

House, J. D. (1998). Age differences in prediction of student achievement for Graduate Record Examination scores. *Journal of Genetic Psychology, 159*(3), 379-382.

House, J. D. & Johnson, J. J. (1998). Predictive validity of the graduate record examination for grade performance in graduate psychology courses. *Psychological Reports, 82*(3), 1235-1238.

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods, 5*(1), 64-86.

Hoyt, W. T. & Kerns, M.-D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods, 4*, 403-424.

Humphreys, L. G. (1968). The Fleeting Nature of the Prediction of College Academic Success. *Journal of Educational Psychology, 59*(5), 375-380.

Ingenkamp, K. (1971). *Die Fragwürdigkeit der Zensurengebung*. Weinheim: Beltz.

Jäger, A. O., Süss, H. M. & Beauducel, A. (1997). BIS-TEST Berliner Intelligenzstruktur-Test - Form 4 Berlin Intelligence Structure Test - Version 4/zpid Synonym(e): BIS-4.

Jensen, A. R. (1969). How much can we boost IQ and scholastic achievement? *Harvard Educationale Review, 39*, 1-123.

Jensen, A. R. (1983). The definition of intelligence and factor-score indeterminacy. *Behavioral and Brain Sciences, 6*(2), 313-315.

Jensen, A. R. (1985). The nature of the Black-White difference on various psychometric tests: Spearman's hypothesis. *Behavioral and Brain Sciences, 8*(2), 193-263.

Jensen, A. R. (1998). *The g Factor. The Science of Mental Ability*. Westport, CT: Praeger.

Johnson, B. T. & Eagly, A. H. (2000). Quantitative synthesis of social psychological research. In Reis, Harry T. & Judd, Charles M. (Hrsg). *Handbook of research methods in social and personality psychology.* (S. 496-528). New York: Cambridge University Press.

Judge, T. A., Higgins, C. A., Thoresen, C. J. & Barrick, M. R. (1999). The big five personaltiy traits, general mental ability, and career success across the life span. *Personnel Psychology, 52*(3), 621-652.

Karabatsos, G. (2001). The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. *Journal of Applied Measurement, 2*(4), 389-423.

Kerlinger, F. N. (1979). *Behavioral Research: A Conceptual Approach*. New York: Holt, Rinehart and Winston.

Klieme, E. & Nauels, H.-U. (1996). Wie hat sich der TMS bewährt? Korrelationsanalysen und Strukturgleichungsmodelle zur Vorhersage des Erfolgs bei den Vorprüfungen im Studiengang Zahnmedizin. In G. Trost (Hrsg.), *Test für medizinische Studiengänge (TMS): Studien zur Evaluation. 20. Arbeitsbericht* (S. 192-213). Bonn: Institut für Test- und Begabungsforschung.

Kobrin, J. L. & Michel, R. (2006). *The SAT as a Predictor of Different Levels of College Performance*. New York: College Board.

Köller, O. (1994). Psychometrische und psychologische Betrachtungen des Rateverhaltens in Schulleistungstests. *Empirische Pädagogik, 8*(1), 59-84.

Köller, O. & Baumert, J. (2002). Das Abitur – immer noch ein gültiger Indikator für die Studierfähigkeit? *Das Parlament, 26/02*.

Köller, O., Baumert, J. & Kai, U. S. (1999). Wege zur Hochschulreife. Offenheit des Systems und Sicherung vergleichbarer Stnadards. *Zeitschrift für Erziehungswissenschaft, 2*, 385-422.

Köller, O., Baumert, J. & Rost, J. (1998). Zielorientierungen: Ihr typologischer Charakter und ihre Entwicklung im frühen Jugendalter. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 30*(3), 128-138.

Köller, O., Rost, J. & Köller, M. (1994). Individuelle Unterschiede beim Lösen von Raumvorstellungsaufgaben aus dem IST-70 bzw. IST-70 Untertest "Würfelaufgaben". *Zeitschrift für Psychologie, 202*, 65-85.

Konegen-Grenier, C. (2001). *Studierfähigkeit und Hochschulzugang* (Bd. 61). Köln: Deutscher Instituts-Verlag.

Konradt, H.-J. (1997). Aufgaben der Hochschulreform und die aktuelle Diskussion. In T. Herrmann (Hrsg.), *Hochschulentwicklung - Aufgaben und Chancen* (S. 147-175). Heidelberg: Asanger.

Krampen, G. (2000). *Handlungstheoretische Persönlichkeitspsychologie*. Göttingen: Hogrefe.

Kuhn, T. S. (1970). *The Structure of Scientific Revolutions.* (2. Auflage). Chicago: The University of Chicago Press.

Kuhn, T. S. (1977). The function of measurement in modern physical science. In *The essential tension: Selected studies in scientific tradition and change* (S. 178-224). Chicago, IL: University of Chicago Press.

Kultusministerkonferenz (2004). *Netzwerk der Exzellenz*. http://www.kmk.org/aktuell/Exzellenznetzwerk.pdf

Kuncel, N. R., Hezlett, S. A. & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the graduate record examinations: Implications for graduate student selection and performance. *Psychological Bulletin, 127*(1), 162-181.

Lajoie, S. P. & Shore, B. M. (1986). Intelligence: The speed and accuracy tradeoff in high aptitude individuals. *Journal for the Education of the Gifted, 9*(2), 85-104.

Lemann, N. (1999). *The Big Test. The secret history of the American meritocracy*. New York: Farrar, Strauss and Giroux.

Levine, M. S. (1977). *Canonical analysis and factor comparison* (Bd. 6). Thousand Oaks, CA: Sage Publications.

Lewin, K. (1999). Studienabbruch in Deutschland. In M. Schröder-Gronostay & H.-D. Daniel (Hrsg.), *Studienerfolg und Studienabbruch. Beiträge aus Forschung und Praxis* (S. 17-49). Neuwied: Luchterhand.

Lienert, G. A. (1978). *Verteilungsfreie Methoden in der Biostatistik* (2. Auflage, Bd. 2). Meisenheim am Glan: Verlag Anton Hain.

Lin, T. H. & Dayton, C. M. (1997). Model selection information criteria for non-nested latent class models. *Journal of Educational and Behavioral Statistics, 22*(3), 249-264.

Linacre, J. M. (1997). KR-20 or Rasch Reliability: Which Tells the "Truth"?, *Rasch Measurement Transactions*. http://www.rasch.org/rmt/rmt113l.htm

Linacre, J. M. (2000). Redundant items, overfit and measure bias, *Rasch Measurement Transactions.* http://www.rasch.org/rmt/rmt143a.htm

Linacre, J. M. (2005a). Facets (Version 3.59). Chicago: Winsteps.com.

Linacre, J. M. (2005b). Facets Rasch measurement computer program (Version 3.59.0). Chicago: Winsteps.com.

Linacre, J. M. (2005c). Size vs. Significance: Infit and Outfit Mean-Square and Standardized Chi-Square Fit Statistic. *Rasch Measurement Transactions.* http://www.rasch.org/rmt/rmt171n.htm.

Linacre, J. M. (2005d). Winsteps (Version 3.57). Chicago: Winsteps.com.

Linacre, J. M. & Wright, B. D. (2002). Construction of measures from many-facet data. *Journal of Applied Measurement, 3*(4), 486-512.

Ling, M. (1967). LS Luegenskala Lying Scale/zpid Synonym(e): Luegen- und Leugnenskala.

Linn, R. L. (1968). Range Restriction Problems in the Use of Self-Selected Groups for Test Validation. *Psychological Bulletin, 69*(1), 69-73.

Linn, R. L. (2005). Evaluating College Applicants: Some alternatives. In W. J. Camara & E. W. Kimmel (Hrsg.), *Choosing Stundents. Higher Education Admission Tools for the 21st Century* (S. 141-156). Mahwah: Lawrence Erlbaum.

Lissmann, U. (1977). *Gewichtung von Abiturnoten und Studienerfolg*. Weinheim: Beltz.

Lord, F. M. & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores.* Reading, Massachusets: Addison-Wesley.

Luce, R. D. & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology, 1*(1), 1-27.

Ludlow, L. H. & Haley, S. M. (1992). Polytomous Rasch models for behavioral assessment: The Tufts assessment of motor performance. In M. Wilson (Hrsg.), *Objective Measurement: Theory into practice*. Norwood, New Jersey: Ablex.

MacCallum, R. C., Wegener, D. T., Uchino, B. N. & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin, 114*(1), 185-199.

MacDonald, P. & Paunonen, S. V. (2002). A monte marlo comparsison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement, 62*(6), 921-943.

Mackintosh, N. J. & Bennett, E. S. (2005). What do Raven's Matrices measure? An analysis in terms of sex differences. *Intelligence, 33*, 663–674.

Maraun, M. D. (1998). The nexus misconceived: Wittgenstein made silly. *Theory & Psychology, 8*(4), 489-501.

Marcus, B. (2003). Persönlichkeitstests in der Personalauswahl: Sind „sozial erwünschte" Antworten wirklich nicht wünschenswert? *Zeitschrift für Psychologie, 211*(3), 138-148.

Marcus, B. & Schütz, A. (2005). Who Are the People Reluctant to Participate in Research? Personality Correlates of Four Different Types of Nonresponse as Inferred from Self- and Observer Ratings. *Journal of Personality, 73*(4), 960-984.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174.

Masters, G. N. (1988). Item discrimination: When more is worse. *Journal of Educational Measurement, 25*(1), 15-29.

McCrae, R. R. & Costa, P. T. (1991). The NEO Personality Inventory: Using the Five-Factor Model in counseling. *Journal of Counseling & Development, 69*(4), 367-372.

McFarland, L. A. & Ryan, A. M. (2000). Variance in faking across noncognitive measures. *Journal of Applied Psychology, 85*(5), 812-821.

Meehl, P. E. & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52*, 194-216.

Michell, J. (1990). *An introduction to the logic of psychological measurement*. Hillsdale, New Jersey,England: Lawrence Erlbaum Associates, Inc.

Michell, J. (1997). Quantitative Science and the Definition of Measurement in Psychology. *British Journal of Psychology, 88*, 355-383.

Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept*. New York, NY: Cambridge University Press.

Michell, J. (2001). Teaching and misteaching measurement in psychology. *Australian Psychologist, 36*(3), 211-217.

Minton, H. L. (1988). *Lewis M. Terman. Pioneer in Psychological Testing*. New York: New York University Press.

Möller, J. & Köller, O. (1997). Kontexteffekte in Berichtszeugnissen. *Psychologie in Erziehung und Unterricht, 44*, 187-196.

Moosbrugger, H. & Hartig, J. (2002). Factor analysis in personality research: Some artefacts and their consequences for psychological assessment. *Psychologische Beiträge, 44*(1), 136-158.

Morgan, R. (1990). Analyses of the predictive validity of the SAT and high school grades from 1976 to 1985. In W. W. Willingham, C. Lewis, R. Morgan & L. Ramist (Hrsg.), *Predicting college grades: An analysis of institutional trends over two decades* (S. 103-115). Princeton, New Jersey: Educational Testing Service.

Morrison, T. & Morrison, M. (1995). A meta-analytic assessment of the predictive validity of the quantitative and verbal components of the Graduate Record Examination with graduate grade point average representing the criterion of graduate success. *Educational and Psychological Measurement, 55*(2), 309-316.

Moser, K., Schwörer, F., Eisele, D. & Haefele, G. (1998). Persönlichkeitsmerkmale und kontraproduktives Verhalten in Organisationen Ergebnisse einer Pilotstudie Personality and counterproductive behavior in organizations: Results of a pilot study. *Zeitschrift für Arbeits- und Organisationspsychologie, 42*(2), 89-94.

Moulton, M. H. (2004). Weighting and Calibration. Merging Rasch Reading and Math Subscale Measures into a Composite Measure. http://www.eddata.com/resources/publications/EDS_Rasch_Moulton_AERA_2004.pdf

Mount, M. K., Barrick, M. R. & Strauss, J. P. (1994). Validity of observer ratings of the big five personality factors. *Journal of Applied Psychology, 79*(2), 272-280.

Myford, C. M. (2002). Investigating design features of descriptive graphic rating scales. *Applied Measurement in Education, 15*(2), 187-215.

Myford, C. M. & Wolfe, E. W. (2002). When raters disagree, then what: Examining a third-rating discrepancy resolution procedure and its utility for identifying unusual patterns of ratings. *Journal of Applied Measurement, 3*(3), 300-324.

Myford, C. M. & Wolfe, E. W. (2003). Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part I. *Journal of Applied Measurement, 4*(4), 386-422.

Myford, C. M. & Wolfe, E. W. (2004). Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part II. *Journal of Applied Measurement, 5*(2), 189-227.

Nauels, H.-U. & Meyer, H. J. (1997). Untersuchungen zur Vorhersagekraft des TMS: differentielle Aspekte der Studienerfolgsprognose und Testfairneß. In G. Trost (Hrsg.), *Test für medizinische Studiengänge (TMS): Studien zur Evaluation. 21. Arbeitsbericht* (S. 76-141). Bonn: Institut für Test- und Begabungsforschung.

Nicholson, R. A. & Hogan, R. (1990). The construct validity of social desirability. *American Psychologist, 45*(2), 290-292.

Nunnaly, J. C. (1978). *Psychometric Theory* (2. Auflage). New York: McGraw-Hill.

OECD (2003). *Bildung auf einen Blick. OECD-Indikatoren 2003*. Bielefeld: Bertelsmann.

Oldfield, R. C. (1939). Some factors in the genesis of interest in psychology. *The British Journal of Psychology, 30*(109-123).

Ones, D. S. & Viswesvaran, C. (1998). The effects of social desirability and faking on personality and integrity assessment for personnel selection. *Human Performance, 11*(2), 245-269.

Orlik, P. (1961). Ein Beitrag zu den Problemen der Metrik und der diagnostischen Valenz schulischer Leistungsbeurteilungen. *Zeitschrift für Experimentelle und Angewandte Psychologie, 8*, 400-408.

Ostini, R. & Nering, M. L. (2005). *Polytomous item response models* (Bd. 144). London: Sage.

Paolillo, J. G. P. (1982). The predictive validity of selected admission variables relative to grade point average earned in a masters of business administration program. *Educational & Psychological Measurement, 42*, 1163-1167.

Paulhus, D. L. (Hrsg.) (1991). *Measurement and control of response bias*. San Diego, CA, US: Academic Press, Inc.

Peeters, H. & Lievens, F. (2005). Situational Judgment Tests and Their Predictiveness of College Students' Success: The Influence of Faking. *Educational & Psychological Measurement, 65*(1), 70-89.

Perline, R., Wright, B. D. & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement, 3*(2), 237-255.

Perron, O. (1907). Grundlagen für eine Theorie des Jacobischen Kettenbruchalgorithmus. *Mathematische Annalen, 64*, 11-76.

Petrill, S. A. & Deater-Deckard, K. (2004). The heritability of general cognitive ability: A within-family adoption design. *Intelligence, 32*(4), 403-409.

Pixner, J., Zapf, S. & Schüpbach, H. (2005). *Die Anforderungsanalyse im Entwicklungsprozess eines mehrstufigen e-Assessments*. auf der DPPD-Arbeitstagung 2005 in Marburg.

Plomin, R. & Spinath, F. M. (2004). Intelligence: Genetics, Genes, and Genomics. *Journal of Personality and Social Psychology, 86*(1), 112-129.

Preiser, G. (1975). Zur „Objektivität" mündlicher Prüfungen im Fach Psychologie. *Psychologische Rundschau, 26*, 256-281.

Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M., et al. (2005). *PISA 2003: Ergebnisse des zweiten Ländervergleichs Zusammenfassung*. Kiel: Leibniz-Institut für die Pädagogik der Naturwissenschaften.

Rahn, H., Trost, G., Bickel, H., Blum, F., Christian, H. & Raiser, H. (1976). *Modellversuch „Studienfeldbezogene Tests"*. Bonn - Bad Godesberg: Institut für Test- und Begabungsforschung.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.*: Copenhagen: Danish Intitute for Educational Research.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, 4*, 321-323.

Rasch, G. (1977). On specific objectivity: An attempt at formalising the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy, 14*, 58-94.

Raven, J. C., Raven, J. & Court, J. H. (1998). APM Advanced Progressive Matrices Advanced Progressive Matrices/author Synonym(e): Progressive-Matrices-Test (PMT); Raven's Progressive Matrices (RPM).

Read, T. R. C. & Cressie, N. A. C. (1988). *Goodness-of fit statistics for discrete multivariate data.* New York: Springer.

Reich, A. (2002). *Hochschulrahmengesetz. Kommentar* (Bd. 8). Bad Honnef: Bock.

Rindermann, H. (1996). *Untersuchungen zur Brauchbarkeit studentischer Lehrevaluationen*. Landau: Empirische Pädagogik.

Rindermann, H. & Oubaid, V. (1999). Auswahl von Studienanfängern durch Universitäten--Kriterien, Verfahren und Prognostizierbarkeit des Studienerfolgs *Zeitschrift für Differentielle und Diagnostische Psychologie, 20*(3), 172-191.

Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R. & Carlstrom, A. (2004). Do Psychosocial and Study Skill Factors Predict College Outcomes? A Meta-Analysis. *Psychological Bulletin, 130*(2), 261-288.

Robie, C., Born, M. P. & Schmit, M. J. (2001). Personal and situational determinants of personality responses: A partial reanalysis and reinterpretation of the Schmit et al (1995) data. *Journal of Business and Psychology, 16*(1), 101-117.

Rogers, H. J. & Hattie, J. A. (1987). A Monte Carlo investigation of several person and item fit statistics for item response models. *Applied Psychological Measurement, 11*(1), 47-57.

Roskam, E. E. (1989). Operationalization, a superfluous concept. *Quality and Quantity, 23*, 237-275.

Rosse, J. G., Stecher, M. D., Miller, J. L. & Levin, R. A. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. *Journal of Applied Psychology, 83*(4), 634-644.

Rost, J. (1988). *Quantitative und qualitative probabilistische Testtheorie*. Bern: Hans Huber.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches. *Applied Psychological Measurement, 14*, 75-92.

Rost, J. (1991). A logistic mixture distribution model for polychotomous item response. *The British Journal of Mathematical and Statistical Psychology, 44*(75-92).

Rost, J. (1996). Logistic mixture models. In W. v. d. L. R. Hambleton (Hrsg.), *Handbook of modern item response theory.* (S. 449-463). Berlin: Springer.

Rost, J. (2002). When personality questionnaires fail to be unidimensional. *Psychologische Beitrage, 44*(1), 108-125.

Rost, J. (2004). *Lehrbuch Testtheorie Testkonstruktion*. Bern: Verlag Hans Huber.

Rost, J., Carstensen, C. & Davier, M. v. (1997). Applying the mixed Rasch model to personality questionnaires. In J. R. R. Langeheine (Hrsg.), *Applications of latent trait and latent class models in the social sciences* (S. 324-332). New York: Waxman.

Rost, J. & Carstensen, C. H. (2002). Multidimensional Rasch measurement via item component models and faceted designs. *Applied Psychological Measurement, 26*(1), 42-56.

Rost, J., Carstensen, C. H. & von Davier, M. (1999). Sind die Big Five Rasch-skalierbar? Eine Reanalyse der NEO-FFI-Normierungsdaten. Are the Big Five Rasch scaleable? A reanalysis of the NEO-FFI norm data. *Diagnostica, 45*(3), 119-127.

Rost, J. & von Davier, M. (1994). A conditional item fit index for Rasch models. *Applied Psychological Measurement, 18*, 171-182.

Rothermund, K. & Brandstädter, J. (2003). Coping With Deficits and Losses in Later Life: From Compensatory Action to Accommodation. *Psychology and Aging, 18*(4), 896-905.

Rothstein, J. M. (2002). Admissions Bias: A New Approach to Validity Estimation in Selected Samples. *Center for Studies in Higher Education: Research & Occasional Paper Series, 3*(2), 1-16.

Russell, D. W. (2002). In search of underlying dimensions:The use (and abuse) of factor analysis in Personality and Social Psychology Bulletin. *Personality and Social Psychology Bulletin, 28*(12), 1629-1646.

Sacks, P. (2000). *Standardized Minds*. New York: Da Capo Press.

Scargle, J. D. (2000). Publication Bias: The "File-Drawer" Problem in Scientific Inference. *Journal of Scientific Exploration, 14*(1), 91-106. http://www.scientificexploration.org/jse/articles/pdf/14.1_scargle.pdf

Scheiblechner, H. (1999). Additive conjoint isotonic probabilistic models (ADISOP). *Psychometrika, 64*(3), 295-316.

Scheiblechner, H. (2002). *Testtheorie*. Unveröffentlichtes Vorlesungsskript. Universität Marburg.

Schmidt-Atzert, L. (2005). *Prädiktion von Studienerfolg bei Psychologiestudenten.* Vortrag auf dem 44. Kongress der DGPs in Göttingen, 29.9.2004. http://www.dgps.de/fachgruppen/diff_psy/PodiumsBeitraegeGoettingen.pdf

Schmidt-Atzert, L. & Krumm, S. (2006). Professionelle Studierendenauswahl durch die Hochschulen - Wege und Irrwege. *Report Psychologie*(6/7 2006), 297-308.

Schmidt, F. L. & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*(2), 262-274.

Schmidt, F. L. & Hunter, J. E. (1999). Theory testing and measurement error. *Intelligence, 27*, 183-198.

Schmidt, F. L. & Hunter, J. E. (2004). General Mental Ability in the World of Work: Occupational Attainment and Job Performance. *Journal of Personality & Social Psychology, 86*(1), 162-173.

Schmidt, F. L., Hunter, J. E. & Outerbridge, A. N. (1986). Impact of job experience and ability on job knowledge, work sample performance, and supervisory ratings of job performance. *Journal of Applied Psychology, 71*(3), 432-439.

Schmit, M. J. & Ryan, A. M. (1993). The Big Five in personnel selection: Factor structure in applicant and nonapplicant populations. *Journal of Applied Psychology, 78*(6), 966-974.

Schneider, A. M. & Briel, J. B. (1990). *Validity of the GRE: 1988-89 Summary Report* (No. GREB-90-01VSS). Princeton, N.J.: Educational Testing Service. http://www.ets.org/Media/Research/pdf/GREB-90-01VSS.pdf

Schönemann, P. H. (1981). Factorial definitions of intelligence: Dubious legacy of dogma in data analysis. In I. Borg (Hrsg.), *Multidimensional Data Representations: When & Why*. Ann Arbor: Mathesis Press. http://www2.psych.purdue.edu/~phs/pdf/33.pdf

Schönemann, P. H. (1994). Measurement: The reasonable ineffectiveness of mathematics in the social sciences. In I. Borg & P. Mohler (Hrsg.), *Trends and Perspectives in Empirical Social Research*. New York: Walter de Gruyter. http://www2.psych.purdue.edu/~phs/pdf/73.pdf

Schönemann, P. H. (1997a). Famous artefacts: Spearman's hypothesis. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition, 16*(6), 665-694. http://www2.psych.purdue.edu/~phs/pdf/83.pdf

Schönemann, P. H. (1997b). Some new results on hit rates and base rates in mental testing. *Chinese Journal of Psychology, 39*(2), 173-192. http://www2.psych.purdue.edu/~phs/pdf/82.pdf

Schönemann, P. H. (2005). Psychometrics of Intelligence. In *Enyclopedia of Social Measurement* (Bd. 3, S. 193-201). New York, NY, US: Elsevier. http://www2.psych.purdue.edu/~phs/pdf/89.pdf

Schönemann, P. H. (2006). Bias In Multiple Correlation Estimates as a Function of Data Pooling, with Applications to the GRE. Paper submitted for publication.

Schönemann, P. H. & Scargle, J. D. (2006). A Generalized Model for Publication Bias and the File-drawer Problem. Paper submitted for publication.

Schönemann, P. H. & Thompson, W. W. (1996). Hit-rate bias in mental testing. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition, 15*(1), 3-28. http://www2.psych.purdue.edu/~phs/pdf/78.pdf

Schrader, F. W. & Helmke, A. (1990). Lassen sich Lehrer bei der Leistungsbeurteilung von sachfremden Gesichtspunkten leiten? – Eine Untersuchung zu Determinanten diagnostischer Lehrerurteile. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 22*(4), 312-324.

Schuler, H. (2001). Arbeits- und Anforderungsanalyse. In H. Schuler (Hrsg.), *Lehrbuch der Personalpsychologie*. Göttingen: Hogrefe.

Schuler, H., Funke, U. & Baron-Boldt, J. (1990). Predictive validity of school grades: A meta-analysis. *Applied Psychology: An International Review, 39*(1), 89-103.

Schuler, H. & Prochaska, M. (2001). LMI Leistungsmotivationsinventar Achievment Motivation Inventory (AMI)/author Synonym(e): HLMT - Hohenheimer Leistungsmotivationstest.

Schumacker, R. E. (2003). *Reliability in Rasch Measurement: Avoiding the Rubber Ruler*. Paper presented at the annual meeting of the American Educational Research Association April 25, 2003 Chicago, Illinois.

Smith, D. B. & Ellingson, J. E. (2002). Substance versus style: A new look at social desirability in motivating contexts. *Journal of Applied Psychology, 87*(2), 211-219.

Smith, E. V., Jr. (2001). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. *Journal of Applied Measurement, 2*(3), 281-311.

Smith, E. V., Jr. (2005). Effect of Item Redundancy on Rasch Item and Person Estimates. *Journal of Applied Measurement, 6*(2), 147-163.

Smith, R. M. (1994). Detecting item bias in the Rasch rating scale mode. *Educational and Psychological Measurement, 54*, 886-896.

Smith, R. M. (1996). Acomparison of the Rasch separate calibration and between-fit methods of detecting item bias. *Educational and Psychological Measurement, 56*, 403-418.

Smith, R. M., Schumacker, R. E. & Bush, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement, 2*, 66-78.

Spearman, C. (1904). 'General intelligence', objectively determined and measured. *American Journal of Psychology, 15*(2), 201-293.

Spearman, C. (1927). *The abilities of man*. London: Mac Millan.

SPSS. (2004). SPSS (Version 12.0): SPSS Incorporation.

Statistisches Bundesamt. (2003). Aktuelle Ergebnisse aus der Studentenstatistik für das Wintersemester 2003/2004. Ergänzende Unterlagen zur Pressekonferenz „Hochschulstandort Deutschland 2003". Wiesbaden: Statistisches Bundesamt. http://www.destatis.de/presse/deutsch/pk/2003/hochschulstandort_2003i.pdf

Statistisches Bundesamt. (2005). Hochschulen: Statistisches Bundesamt.

Stelzl, I. (1979). Is the Rasch model suitable for testing homogeneity? A report on simulation studies with nonhomogeneous data. *Zeitschrift für Experimentelle und Angewandte Psychologie, 26*(4), 652-672.

Stelzl, I. (2005). *Fehler und Fallen der Statistik. Für Psychologen, Pädagogen und Sozialwissenschaftler*. Münster: Waxmann.

Stemmler, G. (2005). Studierendenauswahl durch Hochschulen: Ungewisser Nutzen. *Psychologische Rundschau, 56*(2), 125-127.

Sternberg, R. J. (1993). The concept of "giftedness": a pentagonal implicit theory. In C. Foundation (Hrsg.), *The Origins and Development of High Ability*. West Sussex: John Wiley & Sons.

Sternberg, R. J. & Expertise, Yale University New Haven C. T. U. S. (2004). Theory-based University admissions testing for a New Millennium. *Educational Psychologist, 39*(3), 185-198.

Sternberg, R. J. & Williams, W. M. (1997). Does the Graduate Record Examination predict meaningful success in the graduate training of psychology? A case study. *American Psychologist, 52*(6), 630-641.

Stevens, S. S. (1951). Mathematics, measurement and psychophysics. In S. S. Stevens (Hrsg.), *Handbook of Experimental Psychology* (S. 1-49). New York: Wiley.

Stevens, S. S. (1959). Measurement, psychophysics and utility. In C. W. Churchman & P. Ratoosh (Hrsg.), *Measurement:Definitions and Theories* (S. 18-63). New York: Wiley.

Steyer, R., Yousfi, S. & Würfel, K. (2005). Der Zusammenhang zwischen Schul- und Studiennoten im Diplomstudiengang Psychologie. *Psychologische Rundschau, 56*(2), 123-154.

Stumpf, H. & Nauels, H.-U. (1988). Untersuchung zur prognostischen Validität des „Tests für medizinische Studiengänge" (TMS) in Bezug auf den Abschluss des Studiums der Humanmedizin. In G. Trost (Hrsg.), *Test für medizinische Studiengänge (TMS): Studien zur Evaluation. 12. Arbeitsbericht*. Bonn: Institut für Test- und Begabungsforschung.

Tabachnick, B. G. & Fidell, L. S. (2001). *Using multivariate statistics* (4. Auflage): Boston, MA, Pearson A & B.

Taylor, H. C. & Russell, J. T. (1938). The relationship of validity coefficients to the practical effectiveness of tests in the employment situation. *Psychological Bulletin, 35*, 652.

Taylor, H. C. & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: discussion and tables. *Journal of Applied Psychology, 23*, 565-578.

Tent, L., Fingerhut, W. & Langfeld, H. P. (1976). *Quellen des Lehrerurteils*. Weinheim: Beltz.

Tett, R. P., Jackson, D. N. & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology, 44*(4), 703-742.

Thompson, B. (1995). Canonical correlation analysis. In L. Grimm & P. Yarnold (Hrsg.), *Reading and Understanding Multivariate Statistics* (Bd. 2). Washington, DC: American Psychological Association.

Thorndike, E. L. (1918). The nature, purposes, and general methods of measurement of educational products. In G. M. Whipple (Hrsg.), *Seventeenth Yearbook of the National Society for the Study of Education* (Bd. 2, S. 16-24). Bloomington, Illinois: Public School Publishing.

Thurstone, L. L. (1926). The scoring of individual performance. *Journal of Educational Psychology, 17*, 446-457.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33*, 529-554.

Thurstone, L. L. (1952). L.L. Thurstone. In L. Gardner (Hrsg.), *A History of Psychology in Autobiography Yates* (Bd. 6, S. 294-321). Englewood Cliffs, New Jersey: Prentice Hall.

Trapmann, S., Hell, B. & Schuler, H. (2005a). *Mehrdimensionale Studienerfolgsprognose: Die Bedeutung kognitiver, temperamentsbedingter und motivationaler Prädiktoren für die wichtigsten Erfolgskriterien*. Vortrag auf der 4. Tagung der Fachgruppe Arbeits- und Organisationspsychologie der Deutschen Gesellschaft für Psychologie in Bonn, 19.09.-21.09.2005. http://www.uni-hohenheim.de/studieneignung/publikationen/multi_studienerfolgsprognose_ddp_2005.pdf

Trapmann, S., Hell, B. & Schuler, H. (2005b). *Psychologische Konstrukte als Prädiktoren des Studienerfolgs – eine Metaanalyse*. Vortrag auf der 8. Arbeitstagung der Fachgruppe Differentielle Psychologie, Persönlichkeitspsychologie und Psychologische Diagnostik der Deutschen Gesellschaft für Psychologie in Marburg, 26.09.-27.09.2005. http://www.uni-hohenheim.de/studieneignung/publikationen/metaanalyse_konstrukte_ddp_2005.pdf

Trost, G. (1975). *Vorhersage des Studienerfolgs*. Braunschweig: Georg Westermann Verlag.

Trost, G. (2003). *Deutsche und internationale Studierfähigkeitstests. Arten, Brauchbarkeit, Handhabung*. Bonn: DAAD.

Trost, G. (2005). Studierendenauswahl durch die Hochschulen: Welche Verfahren kommen prinzipiell in Betracht, welche nicht? *Psychologische Rundschau, 56*(2), 138-140.

Trost, G. & Bickel, H. (1979). *Studierfähigkeit und Studienerfolg*. München: Minerva.

Trost, G., Blum, F., Fay, E., Klieme, E., Maichle, U., Meyer, M., et al. (1998). *Evaluation des Tests für medizinische Studiengänge (TMS): Synopse der Ergebnisse*. Bonn: Institut für Test- und Begabungsforschung.

Trost, G. & Freitag, G. (1991). Prognostische Validität studienfeldbezogener Tests zur Beratung von Studierwilligen. In H. Schuler & U. Funke (Hrsg.), *Eignungsdiagnostik in Forschung und Praxis*. Stuttgart: Verlag für angewandte Psychologie.

Trost, G. & Piel, H. P. (1991). *Evaluation des Auswahlverfahrens zur Zulassung an der Wissenschaftlichen Hochschule für Unternehmensführung Koblenz*. Bonn: Institut für Test- und Begabungsforschung.

Uekawa, K. (2006). What is an outfit statistic? http://www.estat.us/sas/Misfit.doc.

van den Berg, P. T. & Feij, J. A. (2003). Complex Relationships Among Personality Traits, Job Characteristics, and Work Behaviors. *International Journal of Selection and Assessment, 11*(4), 326-339.

Van den Wollenberg, A. L. (1979). *The Rasch model and time limited tests - an application and some theoretical contributions.* University of Nijmegen, Nijmegen.

Van den Wollenberg, A. L. (1985). Speed and precision in intelligence tests: Facts or artefacts? *Tijdschrift voor Onderwijsresearch, 10*(2), 69-81.

van der Linden, W. J. (1994). Fundamental Measurement and the Fundamentals of Rasch Measurement. In M. Wilson (Hrsg.), *Objective Measurement: Theory into Practice* (Bd. 2, S. 3-24). Norwood: Ablex Publishing Corporation.

van der Ven, A. H. G. S. & Ellis, J. L. (2000). A Rasch analysis of Raven's standard progressive matrices. *Personality and Individual Differences, 29*(1), 45-64.

Vevea, J. L. & Woods, C. M. (2005). Publication Bias in Research Synthesis: Sensitivity Analysis Using A Priori Weight Functions. *Psychological Methods, 10*(4), 428-443.

Vigneau, F. & Bors, D. A. (2005). Items in Context: Assessing the Dimensionality of Raven's Advanced Progressive Matrices. *Educational and Psychological Measurement, 65*(1), 109-123.

Viswesvaran, C. & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational & Psychological Measurement, 59*(2), 197-210.

von Davier, M. (2001). WINMIRA 2001 (Version 1.37). Kiel: IPN - Institute for Science Education.

von Davier, M. & Molenaar, I. W. (2003). A person-fit index for polytomous Rasch models, latent class models, and their mixture generalizations. *Psychometrika, 68*(2), 213-228.

von Davier, M. & Strauss, B. (2003). New Developments in Testing Probabilistic Models. *International Journal of Sport and Exercise Psychology, 1*(1), 61-81.

Wang, W.-C. & Chen, C.-T. (2005). Item Parameter Recovery, Standard Error Estimates, and Fit Statistics of the WINSTEPS Program for the Family of Rasch Models. *Educational & Psychological Measurement, 65*(3), 376-404.

Wang, W.-C., Cheng, Y.-Y. & Wilson, M. (2005). Local Item Dependence for Items across Tests connected by common stimuli. *Educational and Psychological Measurement,, 65*(1), 5-27.

Weber, H. & Schmidt-Atzert, L. (2005). Stellungnahme der Deutschen Gesellschaft für Psychologie e. V. (DGPs) zur Auswahl von Studierenden durch die Hochschulen *Psychologische Rundschau, 56*(2), 153-154.

Weiner, S. & Wolf, I. (2003). *How to prepare for the GRE test* (15. Auflage). New York: Barron's.

Weingardt, E. (1972). Untersuchungen über Korrelationen zwischen Reifeprüfungsnoten und Erfolg auf der Universität. In K. Ingenkamp (Hrsg.), *Die Fragwürdigkeit der Zensurengebung* (S. 252-255). Weinheim: Beltz.

Westermann, R., Heise, E., Spies, K. & Trautwein, U. (1996). Identifikation und Erfassung von Komponenten der Studienzufriedenheit. *Psychologie in Erziehung und Unterricht, 43*, 1-22.

Westmeyer, H. (2005). Einige Grundsätze zum Vorgehen bei der Auswahl von Studierenden. *Psychologische Rundschau, 56*(2), 142-144.

Wetzenstein, E., Becker, F., Brand, A., Feige, T., Fertl, J. & Geiger, C. (2004). *Anforderungsanalyse zum Psychologiestudium*. Berlin: Humboldt Universität.

Willingham, W. W., Lewis, C., Morgan, R. & Ramist, L. (1990). *Predicting college grades: An analysis of institutional trends over two decades*. Princeton, New Jersey: Educational Testing Service.

Wilson, M. (2003). On Choosing a Model for Measuring. *Methods of Psychological Research, 8*(3), 1-22.

Wirtz, M. & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität: Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen*. Göttingen: Hogrefe.

Wissenschaftsrat. (2004). *Empfehlungen zur Reform des Hochschulzugangs*. http://www.wissenschaftsrat.de/texte/5920-04.pdf

Wolfe, E. W., Moulder, B. C. & Myford, C. M. (2001). Detecting differential rater functioning over time (DRIFT) using a Rasch Multi-faceted Rating scale model. *Journal of Applied Measurement, 2*(3), 256-280.

Wollenberg, v. d. A. L. (1988). Testing a latent trait model. In R. Langeheine & J. Rost (Hrsg.), *Latent trait and latent class models*. New York: Plenum.

Wright, B. D. (1977). Misunderstanding the Rasch model. *Journal for the Educational Measurement, 14*(2), 97-116.

Wright, B. D. (1985). Additivity in Psychological Measurement. In E. E. Roskam (Hrsg.), *Measurement and Personality Assessment* (S. 101-111): Elsevier.

Wright, B. D. (1994). Reasonable mean-square fit values: *Rasch Measurement Transactions*. http://www.rasch.org/rmt/rmt83b.htm

Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues & Practice, 16*(4), 33-45,52.

Wright, B. D. & Masters, G. N. (1982). *Rating Scale Analysis*. Chicago: MESA Press.

Wright, B. D. & Stone, M. H. (1979). *Best Test Design*. Chicago: MESA Press.

Wu, M., Adams, R. & Haldane, S. (2005). ConQuest (Version 3.1). Berkeley: University of California.

Zickar, M. J. (1997). Identifying untraited individuals using model-based measurement. *Dissertation Abstracts International: Section B: The Sciences & Engineering, 58*(6), 3352.

Zimmerhofer, A. (2003). *Neukonstruktion und erste Erprobung eines webbasierten Self-Assessments zur Feststellung der Studieneignung für die Fächer Elektrotechnik, technische Informatik sowie Informatik an der RWTH Aachen.* Unveröffentlichte Diplomarbeit. RWTH Aachen.

22. ANHANG

Anhang A

Aus Anforderungsanalyse-Workshop abgeleitete kritische und typische Anforderungssituationen des Psychologiestudiums

Selbständiger Umgang mit wissenschaftlichen Fragestellungen	**Literaturarbeit**	**Methodenkennt-nisse erwerben und anwende**	**Umgang mit englischer Literatur**	**Transfer und Integration von Informationen**	**Prüfungsmanagement**	**Studienorganisation**
Eigene Fragestellungen entwickeln; Versuchspläne entwickeln; Diplomarbeitsthema finden;	Literaturrecherche, Literaturbearbeitung und -zusammenfassung	Statistik verstehen und anwenden; Umsetzung Fragestellungen in prüfbare Hypothesen	Engl. Texte lesen und zsm.-fassen; Diskussion nach Vortrag in Englisch	Übertragen von prinzipien; Fülle von Infrom. Verarbeiten und filtern; Anwendungen für Theorien finden	Auf unerwartete Fragen reagieren können; Prüfungsstoff strukturieren	Studium und Semester planen; Anforderungen kennen; Prüfungen planen

Präsentation	**Organisation von Ressourcen**	**Teamarbeit**
Referat halten; Studienergebnisse zsm.-fassen und präsentieren	Förderungsquellen auftun; Praktikumsplätze finden;	Gruppendiskussion; Gruppenarbeit (Experimente zusammen planen und durchführen)

Anmerkung:
Fettgedruckte Überschriften stellen die im Plenum gefundenen Oberbegriffe der darunter stehenden (durch Kartenabfragen gewonnenen) Begriffe dar.

Anhang B

Ergebnisse der Kartenabfrage zu Anforderungsmerkmalen des Studienerfolgs im Fach Psychologie

Intelligenzfaktoren	**Instrumentelle Intelligenz**	**Argumentationskompetenz**	**Problemsensitivität**	**Leistungsmotivation**	**Divergentes Denken**
Primärfaktoren (Induktion, verbal Ability; perceptual Speed, numeric Ability); logisches Denken; Konzentrationsfähigkeit; geistige Ausdauer; Sprachgefühl; Allgemeinbildung; Sprachkompetenz (Englisch, Deutsch); methodisches Denken; kognitive Flexibilität; schnell in Dinge einarbeiten können; gutes Gedächtnis; abstraktes Denken; Generalisieren können; formales Denken	Didaktische Kompetenz; EDV-Kenntnisse	Kritik äußern; Diskussionsbereitschaft; Kommunikationskompetenz; Argumentationsfähigkeit; Kritik annehmen; Selbstvertrauen.	Problembewusstsein; Probleme erkennen und kommunizieren, Hilfe einfordern; Problempersistenz; kritisches Denken; Reflexionsfähigkeit; Offenheit für Erfahrungen	Lernbereitschaft; Leistungsbereitschaft; Lernfähigkeit; Motivation; Gewissenhaftigkeit; Handlungsorientierung; Hartnäckigkeit.	Kreativität; Neugier; flexible Zielanpassung; Begeisterungsfähigkeit; Flexibilität; Offenheit; soziale Verantwortung; Rollenübernahme.

Anhang B (Fortsetzung)

Persistenz	**Stabile Persönlichkeit**	**Soziale Kompetenz**	**Selbständigkeit und Kooperation**
Durchhaltefähigkeit; Ausdauer; Geduld; Zielorientierung	Psychische Gesundheit; Realitätsorientierung; Frustrationstoleranz; Belastbarkeit	Kooperationsfähigkeit; soziale Verträglichkeit; Empathie; Hilfsbereitschaft; „Anstandsregeln“; Einfühlungsvermögen; Interesse an Menschen; soziales Engagement; Toleranz; soziale Sensibilität; soziale Kompetenz/Intelligenz; Kommunikationsfähigkeit (in Seminaren; mit Dozenten, in Gruppen; in Prüfungen; bei Präsentationen; Konfliktfähigkeit.	Organisationstalent; „Zeitmanagementfähigkeit“.

Anmerkung: Fettgedruckte Überschriften stellen die im Plenum gefundenen Oberbegriffe der darunter stehenden (durch Kartenabfragen gewonnenen) Begriffe dar.

Anhang C

Fragebogen zur Erfassung eingeschätzter Bedeutsamkeit verschiedener Eignungsmerkmale bezüglich Studienerfolg

Eignungsmerkmal	**Diagnostische Verfahren oder Aufgabentypen zum Erfassen des Eignungsmerkmals**	**Eingeschätzte Wichtigkeit des Eignungsmerkmals bezüglich einzelner Merkmale des Studienerfolgs**	**unwichtig**		**wenig wichtig**	**wichtig**	**sehr wichtig**
Intelligenzfaktoren		Studienabschlussnote	☐	☐	☐	☐	
		weniger Studienabbrüche	☐	☐	☐	☐	
		kürzere St.-Dauer	☐	☐	☐	☐	
		St.-Zufriedenheit	☐	☐	☐	☐	
		Allgemeine Wichtigkeit für das Psychologiestudium	☐	☐	☐	☐	
Instrumentelle Intelligenz		Studienabschlussnote	☐	☐	☐	☐	
		weniger Studienabbrüche	☐	☐	☐	☐	
		kürzere St.-Dauer	☐	☐	☐	☐	
		St.-Zufriedenheit	☐	☐	☐	☐	
		Allgemeine Wichtigkeit für das Psychologiestudium	☐	☐	☐	☐	
Argumentationskompetenz		Studienabschlussnote	☐	☐	☐	☐	
		weniger Studienabbrüche	☐	☐	☐	☐	
		kürzere St.-Dauer	☐	☐	☐	☐	
		St.-Zufriedenheit	☐	☐	☐	☐	
		Allgemeine Wichtigkeit für das Psychologiestudium	☐	☐	☐	☐	

Anhang C (Fortsetzung)

			unwichtig	wenig wichtig	wichtig	sehr wichtig
Problemsensitivität		Studienabschlussnote	☐	☐	☐	☐
		weniger Studienabbrüche	☐	☐	☐	☐
		kürzere St.-Dauer	☐	☐	☐	☐
		St.-Zufriedenheit	☐	☐	☐	☐
		Allgemeine Wichtigkeit für das Psychologiestudium	☐	☐	☐	☐
Leistungsmotivation		Studienabschlussnote	☐	☐	☐	☐
		weniger Studienabbrüche	☐	☐	☐	☐
		kürzere St.-Dauer	☐	☐	☐	☐
		St.-Zufriedenheit	☐	☐	☐	☐
		Allgemeine Wichtigkeit für das Psychologiestudium	☐	☐	☐	☐
Selbständigkeit		Studienabschlussnote	☐	☐	☐	☐
		weniger Studienabbrüche	☐	☐	☐	☐
		kürzere St.-Dauer	☐	☐	☐	☐
		St.-Zufriedenheit	☐	☐	☐	☐
		Allgemeine Wichtigkeit für das Psychologiestudium	☐	☐	☐	☐

Anhang C (Fortsetzung)

			unwichtig		wenig wichtig	wichtig	sehr wichtig
Kooperationsfähigkeit		Studienabschlussnote	☐	☐	☐	☐	
		weniger Studienabbrüche	☐	☐	☐	☐	
		kürzere St.-Dauer	☐	☐	☐	☐	
		St.-Zufriedenheit	☐	☐	☐	☐	
		Allgemeine Wichtigkeit für das Psychologiestudium	☐	☐	☐	☐	
Persistenz		Studienabschlussnote	☐	☐	☐	☐	
		weniger Studienabbrüche	☐	☐	☐	☐	
		kürzere St.-Dauer	☐	☐	☐	☐	
		St.-Zufriedenheit	☐	☐	☐	☐	
		Allgemeine Wichtigkeit für das Psychologiestudium	☐	☐	☐	☐	
Stabile Persönlichkeit		Studienabschlussnote	☐	☐	☐	☐	
		weniger Studienabbrüche	☐	☐	☐	☐	
		kürzere St.-Dauer	☐	☐	☐	☐	
		St.-Zufriedenheit	☐	☐	☐	☐	
		Allgemeine Wichtigkeit für das Psychologiestudium	☐	☐	☐	☐	

Anhang C (Fortsetzung)

			unwichtig		wenig wichtig	wichtig	sehr wichtig
Soziale Kompetenz		Studienabschlussnote	☐	☐	☐	☐	
		weniger Studienabbrüche	☐	☐	☐	☐	
		kürzere St.-Dauer	☐	☐	☐	☐	
		St.-Zufriedenheit	☐	☐	☐	☐	
		Allgemeine Wichtigkeit für das Psychologiestudium	☐	☐	☐	☐	
Divergentes Denken		Studienabschlussnote	☐	☐	☐	☐	
		weniger Studienabbrüche	☐	☐	☐	☐	
		kürzere St.-Dauer	☐	☐	☐	☐	
		St.-Zufriedenheit	☐	☐	☐	☐	
		Allgemeine Wichtigkeit für das Psychologiestudium	☐	☐	☐	☐	

Nennen Sie mögliche <u>Kriterien</u> des Studienerfolgs

Anhang D

Beurteilungsschema der Aufgabe zum empiriebezogenen Denken

Aufgabenteile	**Dimensionen der Aufgabenbeurteilung**	**Beispiel zu nennender Punkte**	**Ratingkategorien**
<u>Teilaufgabe a)</u> **Wie würde man empirisch generell vorgehen?**	Prinzip des Gruppenvergleichs	1= kein Gruppenvgl. Genannt 2= 2 Gruppen 3= EG vgl. mit KG 4=3 Gruppen (neuer vs. alter Drink vs. Placebo) 5= ausführliche Beschreibung von drei Gruppen	1: ungenügend 2: teilweise befriedigend 3: befriedigend 4: gut 5: sehr gut
	Störvariablenkontrolle und Fehlervarianzminimierung	**Randomisierte Zuweisung zu EG und KG**, Kontrolle Faktor Geschmacksvalenz; **messwiederholte Messungen**, um längerfristige von kurzfristigen Effekten zu unterscheiden. **Doppelblindversuch.** Mehrfachmessung zur Erfassung des Wirkungs-verlaufs	1: ungenügend 2: teilweise befriedigend 3: befriedigend 4: gut 5: sehr gut

<u>Teilaufgabe b)</u> **Verallgemeinerbarkeit der Ergebnisse**	Verallgemeinerbarkeit auf verschiedene Personen und Variablen	**Ausreichend große und repräsentative** Stichprobe. **Wahl eines sinnvollen Maßes** für Konzentration	1: ungenügend 2: teilweise befriedigend 3: befriedigend 4: gut 5: sehr gut
	Absicherung gegenüber dem Zufall	1= kein Test genannt 2= Inferenzstatistische Absicherung allgemein genannt 3= t-Test 4= ANOVA 5= ANVOVA	1: ungenügend 2: teilweise befriedigend 3: befriedigend 4: gut 5: sehr gut
	Externe Validität	**standardisiertes Verfahren zur Konzentrationsfähigkeit**; Zusätzliche **Kovariaten kontrollieren** wie Tageszeit, subjektive Befindlichkeit, Geschlecht, Trinkgewohn-heiten oder andere Variablen Mehrfach-Messungen zur Feststellung zeitlicher Wirkungsverläufe (mindestens zwei Messzeit-punkte)	1: ungenügend 2: teilweise befriedigend 3: befriedigend 4: gut 5: sehr gut

Anhang E

„Musterlösung“ zur Beurteilung der Aufgabe zum empiriebezogenen Denken (lediglich als konkretisierendes Beispiel der Beurteilungsdimensionen gedacht):

ad Teilaufgabe a):

- 3 Gruppen: 2.Neuer Energy-Drink 2. Alter E.-Drink, 3. Kontrollgruppe mit Placebo
- Randomisierte Zuweisung der Vpn zu den Bedingungen
- - Störvariablen kontrollieren: Geschmacksvalenz (Fragebogen), Mehrfachmessung über Zeit für Wirkverläufe, Faktor Geschlecht mit erheben und möglichst gleichviel Frauen wie Männer
- Doppelblindversuch (weder Versuchsleiter noch Versuchsperson wissen, was verabreicht wird)
- aV: Standardisierten Konzentrationstest bearbeiten lassen zu verschiedenen Zeitpunkten

b) Verallgemeinerung der Ergebnisse:

- valides Maß der aV gewählt (s. oben)
- repräsentative Stichprobe: genügend groß und so, dass sie Gegebenheiten der Population widerspiegelt (Jede Person der Population muss gleiche Wahrscheinlichkeit haben für den Versuch gewählt worden zu sein).
- Absicherung gegenüber dem Zufall:

  ANKOVA: Faktor Drink: 3 Stufen (neuer vs. alter Drink vs. Placebo); Faktor Geschlecht (als Kontrollfaktor); Faktor Zeit (gleich oder mehr als zwei Messzeitpunkte)

  Kovariate (stetige Variable): Geschmacksvalenz, Intelligenz (wegen möglichen Einflusses auf Konzentrationsleistung)

Anhang F

Beschreibung und Beispiele für Aufgabengruppe 01

Es werden Ihnen zwei Wörter vorgegeben.
Zwischen dem ersten und dem zweiten Wort besteht eine bestimmte Beziehung.
Sie sollen aus den vorgegebenen Antworten 1) bis 4) dasjenige Begriffspaar finden, das in der selben Beziehung steht wie die ersten beiden Wörter. Es gibt immer nur **eine** richtige Lösung.

Beispiel 1:

Lastwagen : Autobahn = ?

1) Straße : Fußweg
2) Helm : Kopf
3) Fahrrad : Fahrradweg
4) Waldweg : Bäume

Lösung:

„Fahrrad : Fahrradweg“ ist richtig. Als Lösung wäre in diesem Fall „3“ in das Antwortkästchen direkt unter der Aufgabe einzutragen.

Beispiel 2:

laut : leise = ?

1) hart : fest
2) dunkel : hell
3) verlobt : verheiratet
4) müde : schläfrig

Lösung:

Da „laut“ das Gegenteil von „leise“ ist, muss „dunkel : hell“ die richtige Lösung sein. Also ist 2) die richtige Lösung, die in das Kästchen direkt unter der Aufgabe einzutragen wäre.

Anhang G

Beschreibung und Beispiele für Aufgabengruppe 02

Es werden Ihnen fünf Wörter vorgegeben. Eines dieser Wörter passt **nicht** zu den anderen vier. Sie sollen aus den vorgegebenen Antworten 1) bis 5) dasjenige Wort herausfinden, das **nicht** zu den anderen passt. Es gibt immer nur **eine** richtige Lösung.

Beispiel:

Welcher Begriff passt nicht zu den anderen?

1) Samt
2) Baumwolle
3) Seide
4) Papier
5) Leinen

Lösung:

Da „Papier“ kein Stoff ist, passt dieser Begriff nicht zu den anderen. Als Lösung wäre in diesem Fall „4“ in das Antwortkästchen direkt unter der Aufgabe einzutragen.

Anhang H

Beschreibung und Beispiele für Aufgabengruppe 03

Es werden Ihnen Zahlenreihen vorgegeben, die nach einer bestimmten Regel aufgebaut sind. Sie sollen in jeder Reihe die nächstfolgende Zahl finden. Die Aufgaben sind entsprechend den folgenden Beispielen zu lösen. Es gibt immer nur **eine** richtige Lösung.

Beispiel 1: **3 6 9 12 15 18 21 24 27 ?**

In dieser Zahlenreihe ist jede folgende Zahl um 3 größer als die vorhergehende.
Die Lösung dieser Aufgabe lautet 30

Beispiel 2: **7 5 9 7 11 9 13 11 15 ?**

In dieser Zahlenreihe werden abwechselnd 2 abgezogen und 4 dazugezählt.
Die Lösung dieser Aufgabe lautet 13

Tragen Sie die Lösung **als Zahl** in das dafür vorgesehene Antwortkästchen direkt neben der jeweiligen Aufgabe ein.

Anhang I

Beschreibung und Beispiele für Aufgabengruppe 04

Es werden Ihnen Zahlentabellen vorgegeben, die nach einer bestimmten Regel aufgebaut sind. Es sind immer die fehlenden Zahlen in der Tabelle nach der jeweiligen Regel zu ergänzen. Die Aufgaben sind entsprechend den folgenden Beispielen zu lösen. Es gibt immer nur **eine** richtige Lösung.

Beispiel 1:

2	**4**	**6**
5	**2**	**7**
3	**2**	**?**

In dieser Zahlentabelle ergeben sich die Zahlen in den Zellen der rechten Spalte jeweils durch Zusammenzählen der Zahlen in den entsprechenden Zeilen: 2+4=6; 5+2=7; 3+2=5
Die Lösung dieser Aufgabe lautet 5

Beispiel 2:

12	**?**	**17**
5	**2**	**3**
2	**3**	**5**

In dieser Zahlentabelle ergeben sich die Zahlen der oberen Zeile durch Malnehmen der in den Spalten stehenden Zahlen und dem Hinzufügen einer 2: (2x5)+2=12; (5x3)+2=17.
Die Lösung dieser Aufgabe lautet 8

Tragen Sie die Lösung **als Zahl** in das dafür vorgesehene Antwortkästchen direkt neben der jeweiligen Aufgabe ein.
Mögliche Rechenoperatoren können Zusammenzählen (+), Abziehen (-), Malnehmen (*), Teilen (:), Quadrieren (X^2, Quadratwurzelziehen ($\sqrt{X}$) und Kombinationen dieser Operatoren sein.

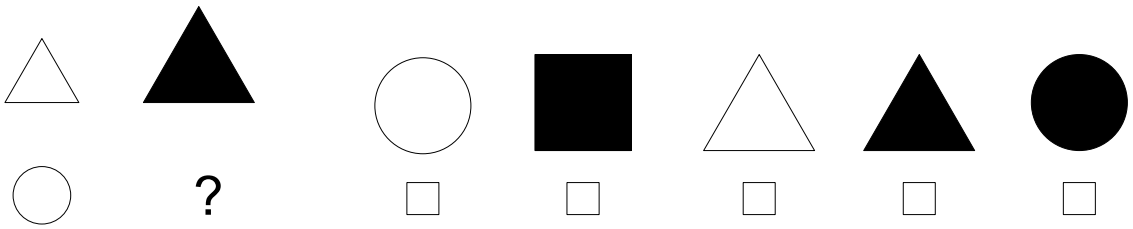
Anhang J

Beschreibung und Aufgabenbeispiele für Aufgabengruppe 03

Jede der folgenden Aufgaben zeigt Ihnen auf der linken Seite eine Reihe von Figuren, die einer bestimmten Regel entsprechend aufgebaut sind. Auf der rechten Seitenhälfte werden Ihnen fünf verschiedene Figuren zur Auswahl angeboten. Es gibt immer nur **eine** richtige Lösung.

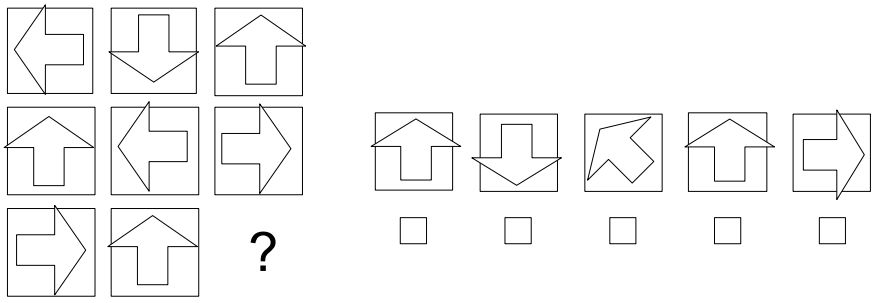
Sie sollen herausfinden, welche der fünf Auswahlfiguren anstatt des Fragezeichens eingesetzt werden muss.

Beispiel 1:



In der oberen Zeile dieses Beispiels verändert sich das kleine weiße Dreieck in ein großes schwarzes, d.h. der kleine weiße Kreis muss sich in einen großen schwarzen Kreis verändern. Damit ist die fünfte Figur der Auswahlfiguren die richtige. Diese Figur muss dann in dem darunter befindlichen Kästchen angekreuzt werden.

Beispiel 2:



In diesem Beispiel wird der obere Pfeil jeder Spalte jeweils um eine Vierteldrehung im Uhrzeigersinn gedreht. Daher muss der fehlende Pfeil in der rechten Spalte nach unten zeigen. Daher ist die zweite Auswahlfigur die richtige und wäre im Kästchen darunter anzukreuzen.
Es ist immer nur eine Lösung richtig!

Anhang K

Verteilung der Testteilnehmer auf Studienfächer

Studienfach	Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Anglistik	2	.5	.5	.5
Biologie	2	.5	.5	.9
Chemie	11	2.5	2.5	3.5
Erziehungswissensch	1	.2	.2	3.7
Ethnologie	14	3.2	3.2	6.9
Geographie	39	9.0	9.0	15.9
Germanistik	1	.2	.2	16.1
Geschichte	1	.2	.2	16.4
Japanologie	2	.5	.5	16.8
Jura	91	21.0	21.0	37.8
Kunstgeschichte	1	.2	.2	38.0
Mathematik-Dipl	1	.2	.2	38.2
Philosophie	1	.2	.2	38.5
Physik	2	.5	.5	38.9
Politik	13	3.0	3.0	41.9
Psychologie	183	42.2	42.2	84.1
Religion	2	.5	.5	84.6
Romanistik	1	.2	.2	84.8
Soziologie	24	5.5	5.5	90.3
Sport	1	.2	.2	90.6
Übersetzen	1	.2	.2	90.8
VWL	40	9.2	9.2	100.0
Gesamt	434	100.0	100.0	

Anhang L

Verteilung der Testteilnehmer auf die Anzahl eingeschriebener Semester je Diplom- oder Hauptfach zum Wintersemester 2004 (WS04).

Anzahl

	Semesteranzahl zum WS04														
Studienfach	1	2	3	4	5	6	7	8	9	10	11	12	13	15	Gesamt
Psychologie	81	0	27	1	24	1	15	1	11	5	7	6	2	2	183
Soziologie	21	1	2	0	0	0	0	0	0	0	0	0	0	0	24
Politik	2	1	9	1	0	0	0	0	0	0	0	0	0	0	13
Jura	78	12	0	0	0	0	0	0	1	0	0	0	0	0	91
Geographie	35	0	3	0	1	0	0	0	0	0	0	0	0	0	39
VWL	36	1	1	0	0	0	1	1	0	0	0	0	0	0	40
Chemie	10	0	0	0	0	0	0	0	1	0	0	0	0	0	11
Biologie	2	0	0	0	0	0	0	0	0	0	0	0	0	0	2
Ethnologie	9	0	3	0	1	0	0	1	0	0	0	0	0	0	14
Erziehungswissenschaft	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Germanistik	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
Geschichte	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Japanologie	1	0	0	0	0	0	0	1	0	0	0	0	0	0	2
Kunstgeschichte	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
Philosophie	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
Physik	0	0	0	0	2	0	0	0	0	0	0	0	0	0	2
Religion	0	1	1	0	0	0	0	0	0	0	0	0	0	0	2
Romanistik	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
Sport	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
Übersetzen	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
Gesamt	277	17	47	2	30	2	16	4	13	6	7	6	2	2	431

Anhang M:

Ergebnisse der multivariaten Varianzanalyse zum Vergleich von Mittelwerten der z-standardisierten Leistungstestskalen von Testteilnehmern aus dem 1.Semester der Studienfächer Psychologie, Jura, VWL, Geographie und Soziologie

Quelle	**Abhängige Variable**	**Quadratsumme vom Typ III**	***df***	**Mittel der Quadrate**	***F***	***p***	**Teststärke**
Korrigiertes Modell	z-verbale Intelligenz	14.86	4	3.71	3.44	.00	.85
	z-Numerische Intelligenz	6.72	4	1.68	1.68	.15	.51
	z-Matrizen	3.69	4	0.92	1.04	.38	.32
	z-empiriebezogenes Denken	14.75	4	3.68	4.08	.00	.91
	z-Kreativität	19.51	4	4.87	4.64	.00	.94
	z-SPARK	11.31	4	2.82	3.04	.01	.80
Intercept	z-verbale Intelligenz	4.16	1	4.16	3.85	.05	.49
	z-Numerische Intelligenz di	1.30	1	1.30	1.30	.25	.20
	z-Matrizen	0.08	1	0.08	0.09	.76	.06
	z-empiriebezogenes Denken	11.68	1	11.68	12.93	.00	.94
	z-Kreativität	9.38	1	9.38	8.94	.00	.84
	z-SPARK	4.17	1	4.17	4.49	.03	.56
Studienfach	z-verbale Intelligenz	14.86	4	3.71	3.44	.00	.85
	z-Numerische Intelligenz di	6.72	4	1.68	1.68	.15	.51
	z-Matrizen	3.69	4	0.92	1.04	.38	.32
	z-empiriebezogenes Denken	14.75	4	3.68	4.08	.00	.91
	z-Kreativität	19.51	4	4.87	4.64	.00	.94
	z-SPARK	11.3	4	2.82	3.04	.01	.80
Fehler	z-verbale Intelligenz	291.23	270	1.07			

Fortsetzung folgende Seite

	z-Numerische Intelligenz	270.25	270	1.00
	z-Matrizen	238.41	270	0.88
	z-empiriebezogenes Denken	243.87	270	0.90
	z-Kreativität	283.29	270	1.04
	z-SPARK	250.67	270	0.92
Gesamt	z-verbale Intelligenz	307.89	275	
	z-Numerische Intelligenz	277.04	275	
	z-Matrizen	242.11	275	
	z-empiriebezogenes Denken	266.35	275	
	z-Kreativität	306.47	275	
	z-SPARK	263.88	275	
Korrigierte Gesamtvariation	z-verbale Intelligenz	306.10	274	
	z-Numerische Intelligenz di	276.98	274	
	z-Matrizen	242.10	274	
	z-empiriebezogenes Denken	258.62	274	
	z-Kreativität	302.81	274	
	z-SPARK	261.98	274	

Anhang N: Korrelationen der Abiturnoten mit Testleistungen in der Gesamtstichprobe ($N = 434$) (für angegebene Abiturleistungen)

Abiturnoten	**Verbale Intelligenz**	**numerische Intelligenz**	**Matrizen**	**Intelligenzgesamtscore**	**SPARK**	**Kreativität**	**Empiriebezogenes Denken**
Abitur-durchschnitt	-.21**	-.18**	-.09*	-.27**	-.21**	-.14**	-.03
N	422	422	422	422	422	422	422
Mathematik	-.16**	-.23**	-.12**	-.26**	-.11*	-.003	-.03
N	389	389	389	389	389	389	389
Deutsch	-.11*	-.01	-.01	-.09*	-.09*	-.11*	-.01
N	379	379	379	379	379	379	379
Englisch	-.16**	-.04	-.02	-.12**	-.18	-.06	-.07
N	351	351	351	351	351	351	351

Anmerkung. Einseitige Signifikanztestung ohne Korrektur für multiples Testen
*: $p < .05$
**: $p < .01$

Anhang O: Korrelationen der Abiturnoten und Testleistungen mit Vordiplomnoten in Psychologie in der Hauptstudiumsstichprobe Psychologie (retrospektive Validität)

Vordiplomfach	**Abitur-durchschnitt**	**Mathematik**	**Deutsch**	**Englisch**	**verbale Intelligenz**	**numerische Intelligenz**	**Matrizen**	**Intelligenz-gesamtscore**	**SPARK**	**Kreativität**
Allgemeine Psychologie I	.33**	.13	.20	.09	-.22*	.01	-.18	-.17	-.36**	.07
N	55	46	45	55	55	55	55	55	55	55
Allgemeine Psychologie II	.34**	.06	.23	.32*	-.20	.00	-.32**	-.21	-.28*	.12
N	55	46	45	55	55	55	55	55	55	55
Biopsychologie	.21	-.13	.29*	.45**	-.21	.03	-.24*	-.14	-.34**	-.09
N	55	46	45	55	55	55	55	55	55	55
Differentielle Psychologie	.43**	.02	.51**	.21	-.03	-.04	-.38**	-.16	-.20	.13
N	55	46	45	55	55	55	55	55	55	55
Methodenlehre	.41**	.39**	.17	.19	-.17	-.17	-.27*	-.27*	-.14	-.09
N	55	46	45	55	55	55	55	55	55	55
Entwicklungs-psychologie	.40**	.21	.24	.22	-.12	-.33**	-.30*	-.35**	-.36**	.01
N	55	46	45	55	55	55	55	55	55	55
Sozialpsychologie	.31**	.13	.26*	.33	-.16	-.25*	-.40**	-.34**	-.10	-.09
N	55	46	45	55	55	55	55	55	55	55
Vordiplomnote	.54**	.24*	.39**	.39**	-.17	-.26*	-.40**	-.39**	-.31**	-.10
N	73	62	60	55	73	73	73	73	73	73

Anhang P: Levene-Test (auf dem Mittelwert basierend) zur Überprüfung der Varianzhomogenität der Testleistungen der Psychologie-Hauptstudiumsstichprobe mit denen des 1. Semesters in Psychologie (z-standardisierte Skalenwerte)

Skala	***F***	***df1***	***df2***	***p***
Verbale Intelligenz	0.76	1	153	.38
Numerische Intelligenz	0.00	1	153	.96
Matrizen	2.03	1	153	.15
SPARK	0.00	1	153	.99
Kreativität	0.03	1	153	.84

Anhang Q: Ergebnisse der multivariaten Varianzanalyse zum Vergleich von Mittelwerten der z-standardisierten Testleistungen von Testteilnehmern aus dem ersten Semester in Psychologie und denjenigen aus dem Hauptstudium

Quelle	**Abhängige Variable**	**Quadratsumme vom Typ III**	***df***	**Mittel der Quadrate**	***F***	***p***	**Teststärke**
Korrigiertes Modell	z-Verbale Intelligenz	3.12	1	3.12	3.91	.05	.50
	z-Numerische Intelligenz	2.52	1	2.52	2.50	.11	.34
	z-Matrizen	0.18	1	0.18	0.18	.67	.07
	z-SPARK	7.59	1	7.59	7.92	.00	.79
	z-Kreativität	0.08	1	0.08	0.12	.72	.06
Intercept	z-Verbale Intelligenz	9.95	1	9.95	12.46	.00	.93
	z-Numerische Intelligenz	0.51	1	0.51	0.51	.47	.11
	z-Matrizen	0.66	1	0.66	0.65	.41	.12
	z-SPARK	2.77	1	2.77	2.89	.09	.39
	z-Kreativität	10.58	1	10.58	15.34	.00	.97
Erstsem. vs. Hauptstudium	z-Verbale Intelligenz	3.12	1	3.12	3.91	.05	.50
	z-Numerische Intelligenz	2.52	1	2.52	2.50	.11	.34
	z-Matrizen	0.18	1	0.18	0.18	.67	.07
	z-SPARK	7.59	1	7.59	7.92	.00	.79
	z-Kreativität	0.08	1	0.08	0.12	.72	.06
Fehler	z-Verbale Intelligenz	122.24	153	0.79			
	z-Numerische Intelligenz	154.06	153	1.00			
	z-Matrizen	155.40	153	1.01			
	z-SPARK	146.66	153	0.95			
	z-Kreativität	105.55	153	0.69			

Fortsetzung folgende Seite

Gesamt	z-Verbale Intelligenz	134.84	155
	z-Numerische Intelligenz	157.00	155
	z-Matrizen	156.28	155
	z-SPARK	156.63	155
	z-Kreativität	116.16	155
Korrigierte Gesamtvariation	z-Verbale Intelligenz	125.36	154
	z-Numerische Intelligenz	156.58	154
	z-Matrizen	155.58	154
	z-SPARK	154.25	154
	z-Kreativität	105.64	154

Anhang R: Korrelationen der Abitur- und Testleistungen mit Klausurteil- und -gesamtleistungen der Orientierungsprüfung (Erstsemesterstichprobe Psychologie)

Abiturnoten und Testskalen	Note Wahrnehmung	Note Lernen	Note Gedächtnis	Note Denken	Note Emotion	Note Motivation	Rohwert Orientierungs-prüfung	Note Orientierungs-prüfung
Abiturdurchschnitt	.34**	.16	.37**	.29**	.14	.23*	-.32**	.31**
N	67	67	67	67	67	67	66	66
Mathematiknote	.13	-.02	.25*	.16	-.004	.07	-.16	.16
N	66	66	66	66	66	66	65	65
Deutschnote	.33**	.11	.43**	.41**	.40**	.41**	-.35**	.31**
N	65	65	65	65	65	65	64	64
Englischnote	.27*	.12	.46**	.42*	.46**	.35*	-.47	-.47
N	64	64	64	64	64	64	63	63
verbale Intelligenz	-.13	.00	-.28**	-.14	-.09	-.12	.29**	-.29**
N	67	67	67	67	67	67	66	66
Numerische Intelligenz	-.15	-.15	-.29**	-.14	.01	-.23*	.15	-.14
N	67	67	67	67	67	67	66	66
Matrizen	-.09	.11	-.22*	-.19	-.09	-.08	.08	-.12
N	67	67	67	67	67	67	66	66
SPARK	.07	.13	.04	.00	-.07	-.04	.01	-.02
N	67	67	67	67	67	67	66	66
Kreativität	-.19	-.09	-.22*	-.41**	-.32**	-.13	.23*	-.22*
N	67	67	67	67	67	67	66	66
Empiriebezogenes Denken	-.21*	-.00	-.10	-.27*	-.21*	-.03	.19	-.19
N	67	67	67	67	67	67	66	66

Anmerkung. Einseitige Signifikanztestung (ohne Korrektur für multiples Testen)
*: $p < .05$
*: $p < .01$

Anhang S: Ergebnisse der multivariaten Varianzanalyse zum Vergleich von Mittelwerten der z-standardisierten Persönlichkeitsfragebögen unter einer Normal- vs. Faking-good-Instruktion (Gesamtstichprobe)

Quelle	**Abhängige Variable**	**Quadratsumme vom Typ III**	***df***	**Mittel der Quadrate**	***F***	***p***	**Teststärke**
Korrigiertes Modell	z-LMI	43.66	1	43.66	48.70	.00	1.00
	z-FZA	16.06	1	16.06	16.69	.00	.98
	z-HZV	24.19	1	24.19	25.65	.00	.99
	z-OFF	3.05	1	3.05	3.23	.07	.43
	z-GEW	34.48	1	34.48	37.73	.00	1.00
	z-NEURO	22.37	1	22.37	23.74	.00	.99
	z-LUEG	16.49	1	16.49	17.12	.00	.98
Intercept	z-LMI	.00	1	0.00	0.00	.94	.05
	z-FZA	.00	1	0.00	0.00	.98	.05
	z-HZV	4.597E-05	1	4.597E-05	0.00	.99	.05
	z-OFF	.06	1	0.06	0.06	.79	.05
	z-GEW	.01	1	0.01	0.01	.91	.05
	z-NEURO	.05	1	0.05	0.06	.80	.05
	z-LUEG	.00	1	0.00	0.00	.93	.05
Nomal- vs. Faking-good-Instruktion	z-LMI	43.66	1	43.66	48.70	.00	1.00
	z-FZA	16.06	1	16.06	16.69	.00	.98
	z-HZV	24.19	1	24.19	25.65	.00	.99
	z-OFF	3.05	1	3.054	3.23	.07	.43
	z-GEW	34.48	1	34.48	37.73	.00	1.00
	z-NEURO	22.37	1	22.37	23.74	.00	.99
	z-LUEG	16.49	1	16.49	17.12	.00	.98
Fehler	z-LMI	381.01	425	0.89			
	z-FZA	408.85	425	0.96			

Fortsetzung folgende Seite

	z-HZV	400.76	425	0.94
	z-OFF	401.61	425	0.94
	z-GEW	388.44	425	0.91
	z-NEURO	400.46	425	0.94
	z-LUEG	409.50	425	0.96
Gesamt	z-LMI	424.68	427	
	z-FZA	424.91	427	
	z-HZV	424.97	427	
	z-OFF	404.75	427	
	z-GEW	422.93	427	
	z-NEURO	422.85	427	
	z-LUEG	426.00	427	

Anhang T: Ergebnisse der multivariaten Varianzanalyse zum Vergleich von Mittelwerten der z-standardisierten Persönlichkeitsfragebögen unter einer Normal- vs. Faking-good-Instruktion (Stichprobe Hauptstudium Psychologie)

Quelle	**Abhängige Variable**	**Quadratsumme vom Typ III**	**df**	**Mittel der Quadrate**	***F***	***p***	**Teststärke**
Korrigiertes Modell	z-LMI	45.51	1	45.518	58.794	.00	1
	z-FZA	30.11	1	30.116	37.440	.00	1
	z-HZV	24.33	1	24.338	25.494	.00	1
	z-OFF	3.46	1	3.466	3.614	.06	0.47
	z-GEW	19.63	1	19.637	24.056	.00	1
	z-NEURO	37.37	1	37.371	39.192	.00	1
	z-LUEG	12.20	1	12.201	12.449	.00	0.94
Intercept	z-LMI	.82	1	0.827	1.068	.30	0.17
	z-FZA	1.45	1	1.456	1.810	.18	0.26
	z-HZV	.104	1	0.104	.109	.74	0.06
	z-OFF	3.73	1	3.731	3.890	.05	0.49
	z-GEW	1.28	1	1.286	1.575	.21	0.24
	z-NEURO	.39	1	0.390	.409	.52	0.10
	z-LUEG	35.89	1	35.891	36.619	.00	1
Nomal- vs. Faking-good-Instruktion	z-LMI	45.51	1	45.518	58.794	.00	1
	z-FZA	30.11	1	30.116	37.440	.00	1
	z-HZV	24.33	1	24.338	25.494	.00	1
	z-OFF	3.46	1	3.466	3.614	.06	0.47
	z-GEW	19.63	1	19.637	24.056	.00	1
	z-NEURO	37.37	1	37.371	39.192	.00	1
	z-LUEG	12.20	1	12.201	12.449	.00	0.94
Fehler	z-LMI	55.74	72	0.774			
	z-FZA	57.91	72	0.804			
	z-HZV	68.73	72	0.955			
	z-OFF	69.04	72	0.959			
	z-GEW	58.77	72	0.816			
	z-NEURO	68.65	72	0.954			

Fortsetzung folgende Seite

	z-LUEG	70.57	72	0.980
Gesamt	z-LMI	102.07	74	
	z-FZA	89.48	74	
	z-HZV	93.17	74	
	z-OFF	76.24	74	
	z-GEW	79.69	74	
	z-NEURO	106.47	74	
	z-LUEG	118.66	74	
Korrigierte Gesamtvariation	z-LMI	101.26	73	
	z-FZA	88.03	73	
	z-HZV	93.07	73	
	z-OFF	72.51	73	
	z-GEW	78.41	73	
	z-NEURO	106.07	73	
	z-LUEG	82.77	73	

Anhang U: Ergebnisse der multivariaten Varianzanalyse zum Vergleich von Mittelwerten der z-standardisierten Persönlichkeitsfragebögen unter einer Normal- vs. Faking-good-Instruktion (Stichprobe Erstsemester zum Winstersemester 2004/05 in Psychologie)

Quelle	**Abhängige Variable**	**Quadratsumme vom Typ III**	**df**	**Mittel der Quadrate**	***F***	***p***	**Teststärke**
Korrigiertes Modell	z-LMI	2.57	1	2.57	4.23	.04	.52
	z-FZA	0.01	1	0.01	0.02	.88	.05
	z-HZV	2.09	1	2.09	2.22	.14	.31
	z-OFF	0.94	1	0.94	1.43	.23	.21
	z-GEW	1.80	1	1.80	2.48	.11	.34
	z-NEURO	0.71	1	0.71	0.81	.37	.14
	z-LUEG	1.27	1	1.27	2.19	.14	.31
Intercept	z-LMI	0.07	1	0.07	0.12	.72	.06
	z-FZA	1.04	1	1.04	1.25	.26	.19
	z-HZV	0.21	1	0.21	0.28	.63	.07
	z-OFF	5.57	1	5.57	8.41	.00	.81
	z-GEW	2.04	1	2.04	2.82	.09	.38
	z-NEURO	2.52	1	2.52	2.87	.09	.38
	z-LUEG	0.25	1	0.25	0.42	.51	.09
Nomal- vs.Faking-good-Instruktion	z-LMI	2.57	1	2.57	4.23	.04	.52
	z-FZA	0.01	1	0.01	0.02	.88	.05
	z-HZV	2.09	1	2.09	2.22	.14	.31
	z-OFF	0.94	1	0.94	1.43	.23	.21
	z-GEW	1.80	1	1.80	2.48	.11	.34
	z-NEURO	0.71	1	0.71	81	.37	.14
	z-LUEG	1.27	1	1.27	2.19	.14	.31
Fehler	z-LMI	48.07	79	0.60			
	z-FZA	65.82	79	0.83			
	z-HZV	74.20	79	0.93			
	z-OFF	52.36	79	0.66			
	z-GEW	57.20	79	0.72			
	z-NEURO	69.16	79	0.87			
	z-LUEG	46.13	79	0.58			
Gesamt	z-LMI	50.71	81				

Fortsetzung folgende Seite

	z-FZA	66.88	81
	z-HZV	76.53	81
	z-OFF	58.83	81
	z-GEW	61.09	81
	z-NEURO	72.36	81
	z-LUEG	47.67	81
Korrigierte Gesamtvariation	z-LMI	50.64	80
	z-FZA	65.84	80
	z-HZV	76.30	80
	z-OFF	53.31	80
	z-GEW	59.00	80
	z-NEURO	69.87	80
	z-LUEG	47.41	80