Was misst das strukturierte Einstellungsinterview?

Studien zur Konstruktvalidität des Multimodalen Interviews

Dissertation

zur

Erlangung des Doktorgrades

der Naturwissenschaften

(Dr. rer. nat.)

dem

Fachbereich Psychologie

der Philipps-Universität Marburg

vorgelegt von

Gerald Richter

aus Bietigheim

Marburg/Lahn 2003

Vom Fachbereich Psychologie
der Philipps-Universität Marburg als Dissertation angenommen.

Erstgutachter: Prof. Dr. Martin Kleinmann
Zweitgutachter: Prof. Dr. Ulrich Wagner

Tag der mündlichen Prüfung am 28.04.2003

Danksagung

Viele Menschen haben zum Gelingen der vorliegenden Arbeit beigetragen. Auch wenn ich hier nicht alle nennen kann, so gilt doch allen mein herzlicher Dank.

Meinem Chef Prof. Martin Kleinmann für die gute Zusammenarbeit und Betreuung der Arbeit. Ich hoffe, dass mein nächster Job mir ähnlich viel Spaß machen wird!

Meinem Kollegen im DFG-Projekt „Managementdiagnostik – Die gute Versuchsperson denkt“ Thomas Hartstein mit dem ich zusammen alle Höhen und Tiefen der Promotion durchlitten habe.

Meinen Kollegen Cornelius Koch und Klaus Lober für wertvolle Anregungen und gute Freundschaft. Viel Spaß und Erfolg in Zürich!

PD Dr. Thomas Staufenbiel der bereitwillig als kompetenter Ansprechpartner und Privatbibliothekar für die unterschiedlichsten Fragen zur Verfügung stand.

Herrn Prof. Ulrich Wagner für die Übernahme des Zweitgutachtens und wichtige Anmerkungen in der Schlussphase des Projektes.

Den Methodikern am Fachbereich Prof. Ingeborg Stelzl, Prof. Hans-Henning Schulze und Prof. Hartmann Scheiblechner für Geduld und gute Ratschläge.

Unseren Diplomanden Katja Nicht, Dorit Auge, Torsten Biemann, Peter Guzzardi, Sonia Canossa Garcia, Andrea Sturm, Daniela Bierwirth und Marten Voigt.

Den Hiwinen Diana Mucha, Latifa Baddour, Gesine Mührling und Yvonne Nestoriuc.

Den Herren Prof. Michael Frese (Gießen), Prof. Dieter Zapf (Frankfurt) und Prof. Walter Neubauer (Bonn) für Unterstützung bei der Datenerhebung.

Günther Kohlhaas von der ZAS in Marburg, der uns sowohl in Marburg als auch hessenweit wichtige Kontakte verschaffte.

Den Korrekturlesern Simone Volkmann, Alex Engelbach und Jörn Sparfeldt.

Allen Teilnehmern und Beobachtern.

Sowie Dr. Susanne Schilling, welche gleich zu Beginn dieses Projektes die richtigen Worte für mich fand: Nur eine fertige Dissertation ist eine gute Dissertation!

Bedanken möchte ich mich auch bei meinen Eltern und bei meiner Schwester, welche mich in den entscheidenden Situationen immer unterstützt haben sowie bei allen Freundinnen und Freunden, für ein Leben außerhalb von Konstrukt- und Kriteriumsvalidität.

Am meisten bedanke ich mich aber meinen Kindern, die mir immer wieder zeigen, was wirklich im Leben zählt.

Und natürlich bei meiner Frau, die ich über alles liebe und die für diese Arbeit bei ihren eigenen Zielen zurückstecken musste.

Marburg, im Januar 2003 Gerald Richter

Inhalt

1 Einleitung

Zusammenfassung. Ziel dieses einleitenden Kapitels ist es, einen Überblick über den Forschungsstand des Einstellungsinterviews zu geben und in ein bisher vernachlässigtes Thema – nämlich der Frage nach der Konstruktvalidität des strukturierten Interviews – einzuführen. Nach der Erläuterung grundlegender Begriffe und relevanter Methoden in diesem Forschungsbereich werden abschließend die drei Fragestellungen vorgestellt, die sich aus dem aktuellen Forschungsstand ergeben und die ihm Rahmen dieser Arbeit mit Hilfe von drei Studien (Kapitel 2 bis 4) bearbeitet werden. Sie lauten: 1. Wie steht es um die Konstruktvalidität von strukturierten Interviews? 2. Hat das Ausmaß, in dem Bewerber die Anforderungsdimensionen eines strukturierten Interviews erkennen, Einfluss auf die konvergente Validität und die gezeigte Leistung? 3. Hat die Bekanntgabe der Anforderungsdimensionen (Transparenz) Einfluss auf die Konstrukt- und Kriteriumsvalidität des strukturierten Interviews?

1.1 Das Einstellungsinterview

1.1.1 Akzeptanz und Verbreitung

Nach Analyse der Bewerbungsunterlagen ist das Einstellungsgespräch sowohl in Deutschland als auch international das am häufigsten verwendete Verfahren zur Auswahl von Bewerbern (vgl. Schuler, Frier & Kauffmann, 1993; Schulz, Schuler & Stehle, 1985). In einer Studie mit 959 Firmen in 20 Nationen kommen Ryan, McFarland, Baron und Page (1999) zu dem Ergebnis, dass, über alle Länder und Organisationen hinweg betrachtet, jeder akzeptierte Bewerber im Durchschnitt 2.78 Interviews durchläuft, bevor er eine endgültige Stellenzusage erhält.

Einstellungsinterviews sind aber nicht nur bei Entscheidern sehr beliebt, sondern werden auch von Bewerbern gut akzeptiert (Fruhner, Schuler, Funke & Moser, 1991; Schuler & Fruhner, 1993).

1.1.2 Kriteriumsvalidität und Moderatoren der Kriteriumsvalidität

Die große Akzeptanz und weite Verbreitung des Einstellungsinterviews erstaunt allerdings insofern, als dass die Kriteriumsvalidität – insbesondere die prognostische Validität – lange Zeit als sehr schlecht galt. In Sammelreferaten und Überblicksarbeiten (z.B. Reilly & Chao, 1982; Schmitt, 1976) wurde sie, bei großer Streuung, auf etwa r = .05 bis r = .25 geschätzt. In einer der ersten Metaanalysen kamen Hunter und Hunter (1984) auf eine maximale durchschnittliche Validität von r = .14. (Kriterium Vorgesetztenurteil). Einschränkend muss allerdings angefügt werden, dass in diese Berechnung lediglich zehn Korrelationskoeffizienten eingingen. Es folgten Jahre intensiver Forschung, wobei die Analyse der sozialen Prozesse während des Interviews und die Verbesserung der Kriteriumsvalidität im Zentrum des Bemühens standen (vgl. Eder, Kacmar & Ferris, 1989; Schuler, 2002).

Diese Bemühungen waren von Erfolg gekrönt. Seit der Studie von Hunter und Hunter (1984) sind eine ganze Reihe von Metaanalysen erschienen, welche die Reliabilität (Conway, Jako & Goodman, 1995) und Kriteriumsvalidität von Einstellungsinterviews für eine Vielzahl von Jobs und ethnischen Gruppen belegen (Huffcutt & Arthur, 1994; Schmidt & Hunter, 1998; Schmidt & Rader, 1999; Wiesner & Cronshaw, 1988; Wright, Lichtenfels & Pursell, 1989). Nach Schuler (1992) wurden dabei insbesondere Anzahl der Interviewer, Anforderungsbezug bzw. Arbeitsanalyse im Vorfeld und Standardisierung/Strukturierung des Interviews als Moderatoren der Kriteriumsvalidität diskutiert.

Für die zeitweise diskutierte Moderatorvariable Anzahl der Interviewer sind die Ergebnisse uneinheitlich. Cronshaw und Wiesner (1989) konnten in ihrer Metaanalyse zeigen, dass die mittlere Validität des unstrukturierten Einstellungsinterviews von r = .11 auf r = .21 steigt, wenn beim Gespräch nicht nur ein Interviewer, sondern zwei oder mehr Interviewer anwesend sind, die dann aufgrund einer Konsensentscheidung oder aufgrund eines statistischen Durchschnittswertes ein gemeinsames Urteil fällen. Bei strukturierten Interviews ergaben sich hingegen keine Unterschiede bzw. eher leicht niedrigere Validitäten für das Team- im Vergleich zum Einzelinterview (r = .33 vs. r = .35). Auch in der Metaanalyse von McDaniel, Whetzel, Schmidt und Maurer (1994) ergaben sich leichte Vorteile für Interviews, die durch einen einzelnen Interviewer durchgeführt wurden im Vergleich zu den Teaminterviews (mittlere Validität von r = .24 vs. r = .17). Dieses Ergebnis ist vor allem auf die strukturierten Teaminterviews zurückzuführen,

die mit r = .20 etwas schlechter abschnitten, als die strukturierten Einzelinterviews mit r = .25.

Als zweiter Moderator der Kriteriumsvalidität wurde die anforderungsbezogene Konstruktion des Interviews aufgrund einer Arbeitsanalyse diskutiert (vgl. Feild & Gatewood, 1989). McDaniel et al. (1986, zitiert nach Schuler, 1992) fanden in ihrer Metaanalyse mittlere Validitäten von r = .30 für anforderungsbezogen konstruierte Interviews und Werte von r = .21 für traditionelle (psychologische) Interviews ohne Arbeitsanalyse. Langdale und Weitz (1973) konnten in einem experimentellen Design zeigen, dass die Reliabilität des Interviews steigt, wenn der Interviewer im Vorfeld umfassende Informationen über die ausgeschriebene Stelle hat. Dies ist insofern interessant, als dass die Wurzel aus der Reliabilität eines Verfahrens eine Obergrenze seiner Validität darstellt. Weitere Studien, die einen positiven Zusammenhang zwischen Anforderungsbezug und Kriteriumsvalidität belegen, werden bei Feild und Gatewood (1989) diskutiert.

Als wichtigster Moderator der Kriteriumsvalidität hat sich die Standardisierung/Strukturierung des Interviews herausgestellt. So berichten beispielsweise Wiesner und Cronshaw (1988) in ihrer Metaanalyse mittlere Validitätskoeffizienten von r = .13 für unstrukturierte Interviews und von r = .40 für strukturierte Interviews. In einer Metaanalyse mit 114 Primärstudien unterscheiden Huffcutt und Arthur (1994) vier Ausprägungen von Strukturierungsgrad. Diese reichen von unstrukturiert (Level 1) bis hoch strukturiert (Level 4) und lassen sich auf die beiden Dimensionen Standardisierung der Interviewfragen und Standardisierung der Antwortevaluation zurückführen. Während in typischen Level 1 Interviews im Vorfeld keine expliziten Absprachen über die Art der Interviewfragen bzw. Bewertung der Antworten getroffen werden, sind die thematischen Inhalte bei Level 2 Interviews bereits vor Interviewbeginn geregelt. In Level 3 Interviews werden den Bewerbern wortwörtlich festgelegte Fragen gestellt, welche meist aus einem Fragen-Pool ausgewählt werden. In hochstrukturierten Level 4 Interviews ist darüber hinaus die Bewertung der Antworten genau festgelegt, so dass verschiedene Bewerber einer Stelle nach den selben Kriterien bewertet werden können. Je nach Ausmaß der Standardisierung berichten Huffcutt und Arthur (1994) mittlere Validitätskoeffizienten von r = .11 (Level 1), r = .20 (Level 2) und r = .34 (Level 3 bzw. Level 4). Auch in den Metaanalysen von McDaniel et al. (1994) sowie Schmidt und Hunter (1998) zeigen sich unstrukturierte Interviews den strukturierten unterlegen. In der ersten Studie

ergaben sich mittlere Validitätskoeffizienten von *r* = .18 vs. *r* = .25 (korrigiert: *r* = .33 vs. *r* = .44) und in der zweiten Studie mittlere korrigierte Validitätskoeffizienten von *r* = .38 vs. *r* = .51. Die relativ hohen mittleren Validitätskoeffizienten der unstrukturierten Interviews in der Studie von McDaniel et al. (1994) erklären die Autoren damit, dass ihre unstrukturierten Interviews im Vergleich zu den Interviews in früheren Metaanalysen relativ strukturiert gewesen seien.

Da strukturierte Interviews meist aufgrund einer Arbeitsanalyse wie z.B. der Critical Incident Technique (CIT) von Flanagan (1954) entwickelt werden (vgl. Campion, Pursell & Brown, 1988; Latham, 1989), ergibt sich eine Konfundierung zwischen den beiden letztgenannten Moderatorvariablen Anforderungsbezug und Standardisierungsgrad. Außerdem gehören Anforderungsbezug bzw. Arbeitsanalyse inzwischen zum theoretischen Standard bei der Entwicklung von Instrumenten zur Personalauswahl (Joiner, 2000), so dass dieser Moderator heute meist nicht mehr extra benannt wird.

1.2 Strukturierungsgrad und Beispiele strukturierter Interviews

1.2.1 Komponenten der Strukturierung

Entsprechend der oben diskutierten Ergebnisse bemühen sich in den letzten Jahren einige Forscher verstärkt darum, die verschiedenen Aspekte der Standardisierung/Strukturierung genauer zu differenzieren. So haben beispielsweise Campion, Palmer und Campion (1997) in ihrer Überblicksarbeit 15 verschiedene Komponenten benannt, die sich nachweislich positiv auf Reliabilität oder Validität auswirken und anhand derer sich der Strukturierungsgrad von Einstellungsinterviews unterscheiden lässt (vgl. Tabelle 1).

Tabelle 1: Elemente zur Strukturierung des Interviews (nach Campion, Palmer & Campion, 1997)

1.	Base questions on a job analysis
2.	Ask exact same questions of each candidate
3.	Limit prompting, follow-up questioning, and elaboration on questions
4.	Use better types of questions
5.	Use longer interview or larger number of questions
6.	Control ancillary information
7.	Do not allow questions from candidate until after the interview
8.	Rate each answer or use multiple scales
9.	Use detailed anchored rating scales
10.	Take detailed notes
11.	Use multiple interviewers
12.	Use same interviewer(s) across all candidates
13.	Do not discuss candidates or answers between interviews
14.	Provide extensive interviewing training
15.	Use statistical rather than clinical prediction

1.2.2 Das Biographische und das Situative Interview (BI und SI)

Unter den verschiedenen strukturierten Interviews haben sich in der wissenschaftlichen Literatur zwei in den 80iger Jahren entwickelte Verfahren als besonders erfolgreich herausgestellt. Es sind dies einerseits das Biographische Interview (BI; vgl. Janz, 1982, 1989)[1] sowie andererseits das Situative Interview (SI;

[1] Interviews mit vergangenheitsbezogenen bzw. biographischen Fragen wurden unter verschiedenen Namen veröffentlicht. Janz (1982) nennt sein Interview „patterned behavior description interview“ (PBDI oder BDI), während Pulakos und Schmitt (1995) ihre Fragen als “experience-based” bezeichnen und Motowidlo, Carter, Dunnette und Tippins (1992) für ihr Interview den Namen “structured behavior interview” (SBI) eingeführt haben. Allen diesen Verfahren gemeinsam ist, dass sie sich auf das tatsächliche Verhalten in vergangenen Situationen beziehen. Daher sollen im Folgenden die verschiedenen Namen als Synonyme aufgefasst werden, wie es auch Pulakos und Schmitt (1995) oder Huffcutt, Conway, Roth und Klehe (2002) vorschlagen. In Deutschland hat sich für vergangenheitsorientierte Interviews der Name Biographisches Interview (BI) durchgesetzt (vgl. Schuler, 1992), der auch im Folgenden verwendet werden soll.

vgl. Latham, 1989; Latham & Saari, 1984; Latham, Saari, Pursell & Campion, 1980; Weekley & Gier, 1987). Der Hauptunterschied zwischen beiden Interviews ist, dass biographische Fragen vergangenheitsbezogen sind und situative Fragen zukunftsbezogen.

Der zentrale Gedanke des BI ist, dass der beste Prädiktor für zukünftiges Verhalten, das in der Vergangenheit gezeigte Verhalten ist (vgl. Janz, 1982, 1989). Folglich werden die Bewerber danach befragt, wie sie sich in früheren arbeitsrelevanten Job- und Lebenssituationen verhalten haben. Die Fragen werden möglichst aufgrund einer Anforderungsanalyse, zumeist der CIT, ausgewählt und sollen für Berufserfolg relevantes und notwendiges Job-Wissen, Fertigkeiten und Fähigkeiten abfragen.

Situative Fragen basieren hingegen auf der Ziel-Setzungs-Theorie (Locke & Latham, 1990). Den Bewerbern werden daher hypothetische, jobbezogene Situationen dargestellt, in die sie sich hineinversetzen sollen. Anschließend werden die Bewerber gefragt, wie sie sich in solchen Situationen verhalten würden. Die Situationen wurden wiederum aufgrund einer Anforderungsanalyse (CIT) ausgewählt. Ein Vorteil der situativen Fragen ist, dass auch Quereinsteiger oder Jobneulinge, also Bewerber die keine direkten Erfahrungen mit den kritischen Situationen haben, eine Antwort geben können.

BI und SI gehören nach Huffcutt und Arthur (1994) zu den stark bis sehr stark strukturieren Verfahren und zeigen in aktuellen Metaanalysen hohe Kriteriumsvaliditäten (Huffcutt, Conway, Roth & Klehe, 2002; Latham & Sue Chan, 1999; Taylor & Small, 2000). So finden sich beispielsweise in der letztgenannten Studie von Huffcutt et al. (2002) mittlere Validitäten von .27 für das SI und von .31 für das BI. Innerhalb der strukturierten Interviews sind BI und SI die am häufigsten eingesetzten Verfahren. So wurden beispielsweise in der Metaanalyse von Huffcutt, Roth und McDaniel (1996) 22 hoch-strukturierte Interviews untersucht. 10 davon bestanden aus situativen und 7 aus biographischen Fragen.

Die starke Standardisierung und Strukturierung von SI und neueren Varianten des BI sind zwar zentrale Gemeinsamkeiten beider Verfahren und wesentliche Grundlagen für ihre gute Reliabilität und Validität, scheinen sich jedoch negativ auf die Akzeptanz durch die Bewerber und Anwender auszuwirken (z.B. Latham, 1989; Latham & Finnegan, 1993; van der Zee, Bakker & Bakker, 2002). So konnte Latham (1989) in einer Studie zeigen, dass Bewerber unstrukturierte Interviews bevorzugen, da sie diese als eher steuerbar erleben und mehr Möglichkeiten sehen, sich gut zu

verkaufen. Außerdem sahen sowohl Bewerber als auch Rechtsanwälte größere Chancen einer erfolgreichen Klage vor dem Arbeitsgericht, wenn die Ablehnung des Bewerbers aufgrund eines unstrukturierten und nicht aufgrund eines strukturierten Interviews geschehen ist. Hinzu kommt die massive Unterschätzung der Nützlichkeit strukturierter Interviews durch die verantwortlichen Entscheider in den Unternehmen, wie sie beispielsweise in der Studie von Terpstra und Rozell (1997) deutlich wird. Diese und ähnliche Ergebnisse führten zur Entwicklung eines weiteren Interviewverfahrens, dem sogenannten Multimodalen Interview (MMI; Schuler, 1989, 1992; Schuler & Funke, 1989).

1.2.3 Das Multimodale Interview (MMI)

Ziel des MMI ist die Verbesserung der prognostischen Validität bei gleichzeitiger Tauglichkeit zur Selbstselektion und unter Wahrung von Akzeptanz (sozialer Validität) und Praktikabilität (vgl. Schuler, 1992; Schuler & Moser, 1995). Um diese Ziele zu erreichen, werden im MMI bis zu acht verschiedene Komponenten eingesetzt, welche teilweise strukturiert und teilweise unstrukturiert sind. Die Komponenten im Einzelnen sind:

1. Gesprächsbeginn,
2. Selbstvorstellung des Bewerbers (Selbstpräsentation),
3. Berufsinteressen und Berufswahl,
4. Freier Gesprächsteil,
5. Biographiebezogene Fragen,
6. Realistische Tätigkeitsinformation,
7. Situative Fragen,
8. Fragen des Bewerbers und Gesprächsabschluss.

Eine Bewertung des Bewerbers, welche dann zur Prognose des späteren Berufserfolges genutzt werden kann, erfolgt in den standardisierten Komponenten Selbstvorstellung, Berufsinteressen, sowie mit Hilfe von biographischen und situativen Fragen. Dabei besitzt die Komponente Berufsinteressen den geringsten Vorhersagewert. Die Teile Gesprächsbeginn, freier Gesprächsteil, realistische Tätigkeitsinformation und Gesprächsabschluss dienen vor allem der Auflockerung des Interviews und tragen dadurch zu einer verbesserten Akzeptanz auf Seiten des

Bewerbers bei. Eine weitere Funktion der realistischen Tätigkeitsinformation ist die Unterstützung der Selbstselektion der Bewerber (vgl. Wanous, 1978). Zu einer Erhöhung der Akzeptanz tragen auch die standardisierten Komponenten Selbstvorstellung und Berufsinteressen bei, da diese Fragen umfassen, welche den typischen Erwartungen eines Bewerbers an ein Auswahlinterview entsprechen.
Die Kombination von SI und BI in Form des MMI hat seine Validität in einer Vielzahl von Studien unter Beweis gestellt. Zielgruppen waren u.a. Auszubildende im Bankenbereich (Schuler & Moser, 1995), Pächter von Brauereigaststätten (Schuler, 1999), potentielle Unternehmensgründer (Schuler & Rolfs, 2000), Ingenieure in Forschung und Entwicklung (Schuler, Funke, Moser, Donat & Barthelme, 1995) sowie mittlere und obere Führungskräfte in verschiedenen multinationalen Konzernen (Deller & Kendelbacher, 1998; Stahl, 1995). Inzwischen gibt es vermehrt auch englischsprachige Autoren, die eine Kombination von verschiedenen Fragetypen empfehlen (Salgado & Moscoso, 2002) bzw. einsetzen (Campion, Campion & Hudson, 1994; Conway & Peneno, 1999), ohne sich jedoch explizit auf das MMI zu beziehen.

1.3 Die Konstruktvalidität strukturierter Interviews

Wie oben dargestellt ist die Kriteriumsvalidität strukturierter Interviews gut belegt; es gibt jedoch erst wenige Studien zur Konstruktvalidität des strukturierten Interviews. Dies ist nach Huffcutt, Conway, Roth und Stone (2001) nicht überraschend, da Interviews job-spezifisch mit Hilfe von Arbeitsanalysen entwickelt werden. Daher unterscheiden sich Interviews in Art und Anzahl der erfassten Dimensionen bzw. Konstrukte. Andererseits ist der Mangel an Studien zur Konstruktvalidität sehr bedauerlich, da der Prozess der Konstruktvalidierung helfen könnte zu verstehen, warum und wann – d.h. unter welchen Bedingungen, wie z.B. Job oder Jobgruppe – Interviews als Instrument der Personalauswahl prognostisch valide sind (vgl. Cronshaw & Wiesner, 1989). Zur Veranschaulichung ein Beispiel. McDaniel et al. (1994) kommen zu dem Ergebnis, dass Interviews ein Konglomerat von kognitiven Fähigkeiten, manifester Motivation sowie sozialen und kommunikativen Fähigkeiten erfassen. Dieses Konstruktgemisch hat sich als dazu geeignet erwiesen Berufserfolg vorherzusagen. Unklar bleibt jedoch zu welchem Ausmaß die verschiedenen Faktoren erfasst werden. Es ist nun denkbar, dass die Vorhersage von Berufserfolg

bei einem Sachbearbeiter aufgrund der im Interview erfassten kognitiven Fähigkeiten gelingt. Dasselbe Interview könnte gleichzeitig in der Lage sein den Berufserfolg als Versicherungsvertreter zu prognostizieren, wobei nun jedoch die im Interview erfassten kommunikativen Fähigkeiten ausschlaggebend sein könnten. Wenn man nun die verschiedenen im Interview erfassten Konstrukte differenzieren könnte, dann sollte es möglich sein validere Instrumente zur Personalauswahl zu entwickeln. Fehlendes Wissen zur Konstruktvalidität eines Verfahrens birgt außerdem die Gefahr, das schon relativ kleine Veränderungen der Rahmenbedingungen massive Auswirkungen auf die prognostische Validität eines Verfahrens haben. Die Einführung eines kognitiven Fähigkeitstests zur Vorauswahl könnte im obigen Beispiel zu einem Absinken der prognostischen Validität des Interviews für die Gruppe der Sachbearbeiter führen, während die prognostische Validität für die Versicherungsvertreter nahezu unberührt bleiben könnte.

1.3.1 Externe und interne Konstruktvalidität

Theoretisch gibt es viele verschiedene Ansätze, die im Prozess der Konstruktvalidierung eines Verfahrens eingesetzt werden können. Michel und Conrad (1982) nennen beispielsweise inhaltslogische Analyse und Überprüfung der inneren Konsistenz, Ermittlung der Stabilität über einen bestimmten Zeitraum oder, unter experimentell gesetzten unterschiedlichen Bedingungen, Feststellungen von Gruppenunterschieden, Korrelationen mit anderen Tests, mit Beurteilungsskalen und anderen beobachtbaren Konstruktindikatoren sowie die Faktorenanalyse. Die weiter unten aufgeführten Studien machen deutlich, dass die Konstruktvalidität des strukturierten Interviews bisher hauptsächlich mit Hilfe von Korrelationen zwischen Interview (Gesamtscore oder einzelne Dimension) und diversen externen Maßen und Kriterien (z.B. Intelligenztests, Vorgesetztenurteil, Persönlichkeitsfragebögen) überprüft wurde. Kolk (2001, S. 8) bezeichnet diese Art der Konstruktvalidierung als *externe* Konstruktvalidität, da sie mit Hilfe von Verfahren durchgeführt wird, die außerhalb des eigentlich interessierenden Messinstrumentes (z.B. AC oder Interview) stehen aber konzeptionell mit diesem zusammenhängen sollen (z.B. Selbsteinschätzung der Dominanz oder Test zur Erfassung von Job-Wissen). Von Cronbach und Meehl (1955) ist hierfür der Ausdruck nomologisches Netzwerk eingeführt worden.

Als *interne* Konstruktvalidierung bezeichnet Kolk (2001) hingegen die Untersuchung der Struktur innerhalb des interessierenden Messinstrumentes, also konvergente und diskriminante Validität innerhalb des untersuchten Verfahrens. Die Überprüfung der internen Konstruktvalidität geschieht meist mit Hilfe der Multitrait-Multimethod-Matrix (MTMM; siehe Campbell & Fiske, 1959), zu deren Auswertung seit einigen Jahren lineare Strukturgleichungsmodelle bzw. die konfirmatorische Faktorenanalyse herangezogen wird (vgl. Kleinmann & Köller, 1997; Marsh, 1989; Widaman, 1985). Die Begriffe interne und externe *Konstruktvalidität* sind von externer und interner *Validität* zu unterscheiden, bei denen es um die Generalisierbarkeit von Studien und Aussagefähigkeit von Forschungsdesigns geht (vgl. Kacmar, Ratcliff & Ferris, 1989)

1.3.2 Studien zur externen Konstruktvalidität

Im Folgenden sollen zunächst einige Vorstellungen zu den im strukturierten Interview erfassten Konstrukten dargestellt werden, bevor dann im zweiten Teil die zentralen Ergebnisse aktueller Studien diskutiert werden.

In einer Überblicksarbeit geht Harris (1999) auf vier Konstruktbereiche ein, die potentiell im Gesamtscore des strukturierten Einstellungsinterview erfasst werden. Dabei kommt er zu dem Schluss, dass übliche Assessment Center Dimensionen, Taktisches Wissen („tacit knowledge“) und Passung zwischen Person und Organisation („person-organization-fit“) prinzipiell im strukturierten Interview erfasst werden können, kognitive Fähigkeiten jedoch nicht. Auch Posthuma, Morgeson und Campion (2002) kommen in ihrem Review zu dem Ergebnis, dass das Einstellungsinterview in der Lage ist, eine Vielzahl von Konstrukten zu erfassen – abhängig davon, welche Fragen gestellt werden und wie sie bewertet werden. Sie betonen, dass es aufgrund der interpersonalen Situation möglich ist, Konstrukte zu erfassen, die mit anderen Methoden der Personalauswahl nur schwer zu messen sind (z.B. Person-Organization-Fit, Interpersonale Fähigkeiten).

Welche empirischen Ergebnisse zur externen Konstruktvalidität strukturierter Interviews können aufgeführt werden? Die berichteten Studien untersuchten die Zusammenhänge zwischen 1. Gesamtscore im Interview und allgemeiner Intelligenz (Huffcutt et al., 1996), 2. einzelnen Dimensionen mit etablierten Messinstrumenten (Mumford, Costanza, Connelly & Johnson, 1996), 3. einzelnen Interview-Dimensionen und der „Overall Job Performance“ (Huffcutt et al., 2001) sowie 4. dem Gesamtscore im Interview und verschiedenen Konstruktbereichen (Salgado &

Moscoso, 2002). Abschließend werden Studien zur externen Konstruktvalidität des MMI dargestellt.

Huffcutt et al. (1996) untersuchten metaanalytisch den Zusammenhang zwischen Gesamtscore im Einstellungsinterview und allgemeiner kognitiver Leistungsfähigkeit. Bei Zusammenfassung aller 49 Studien mit insgesamt mehr als 12000 Teilnehmern kamen sie auf eine mittlere Korrelation von r = .25. Dabei scheint der Zusammenhang für hoch-strukturierte Interviews mit r = .23 etwas niedriger zu liegen, wobei die Fragen des BI mit r = .12 deutlich unter den Fragen des SI mit r = .21 liegen. Die höchsten Zusammenhänge zwischen Gesamtscore im Interview und allgemeiner Intelligenz ergeben sich für Jobs mit geringer Komplexität (r = .36) und wenn die Interviewer die Ergebnisse der Intelligenztests vorliegen haben (r = .37). Nach Ansicht von Huffcutt et al. (1996) scheinen damit die meisten Einstellungsinterviews zumindest zum Teil auch die kognitive Leistungsfähigkeit zu erfassen. Die durch die allgemeine Intelligenz erklärte Varianz in den Interviews liegt jedoch selbst bei minderungskorrigierten Korrelationen unter 20% und ist für strukturierte Interviews nochmals deutlich geringer.

Die Konstruktion biographischer Fragen wird ausführlich bei Mumford, Costanza, Connelly und Johnson (1996) dargestellt. Außerdem finden sich dort Übersichten über 7 Feld- und 6 Laborstudien, in denen strukturierte BI für unterschiedlichste Konstrukte entwickelt wurden (z.B. Selbstregulation, Locus of Control, Integrität, Planung & Organisation, Führung). Die einzelnen Dimensionen der Interviews wurden soweit möglich mit etablierten Messinstrumenten korreliert. Die teilweise hohen Übereinstimmungen zwischen gleichen Konstrukten (z.B. Stress-Toleranz mit emotionaler Stabilität r = .66), werden von Mumford et al. (1996) als Beleg für die Konstruktvalidität der von ihnen vorgestellten Konstruktionsmethode gewertet. Leider beziehen sich die in dieser Studie präsentierten Ergebnisse lediglich auf das BI und sind auch nicht durch eine metaanalytische Vorgehensweise abgesichert.

In der Metaanalyse von Huffcutt et al. (2001) wurden die Korrelationen zwischen einzelnen Interviewdimensionen (nicht dem Gesamtscore!) und der „Overall Job Performance“ (Einschätzung durch den Vorgesetzten) untersucht. Zunächst entwickelten Huffcutt et al. (2001) aufgrund von Literaturrecherchen eine Taxonomie von Konstrukten, die potentiell in Interviews erfasst werden. Insgesamt kamen sie auf 22 Konstrukte in folgenden Oberkategorien:

1. Mentale Fähigkeiten (Allgemeine Intelligenz, angewandte mentale Fähigkeiten, Kreativität und Innovation)
2. Kenntnisse und Fähigkeiten (Job-Wissen und Fähigkeiten, Ausbildung und Training, Erfahrung und Arbeitsgeschichte)
3. Persönlichkeitsmerkmale (z.B. Extraversion, emotionale Stabilität)
4. Angewandte soziale Fertigkeiten (Kommunikation, Interpersonale Fähigkeiten, Führung, Überzeugen und Verhandeln)
5. Interessen und Vorlieben (berufsbezogen, Hobbys)
6. Passung zur Organisation (Werte und moralische Standards)
7. Physische Attribute (Generelle Attribute, jobbezogene Fähigkeiten)

Die Zuordnung der 47 Interviewstudien mit insgesamt 338 Dimensionen zu den verschiedenen Konstrukten ergab, dass in den Interviews insbesondere versucht wurde, Persönlichkeitsmerkmale (34.9%), angewandte soziale Fertigkeiten (27.8%) und Mentale Fähigkeiten (16.3%) zu erfassen. Dabei zeigte sich, dass mit strukturierten Interviews andere Konstrukte erfasst werden sollten, als mit unstrukturierten Interviews. Hoch strukturierte Interviews wurden besonders häufig zur Erfassung folgender Konstrukte konstruiert: Angewandte mentale Fähigkeiten, Job-Wissen und Fähigkeiten, angewandte soziale Fähigkeiten (insb. Kommunikation, interpersonale Fähigkeiten, Führung) und Passung zur Organisation. In einem zweiten Schritt bestimmten Huffcutt et al. (2001) metaanalytisch die Validität der verschiedenen Konstrukte. Hierzu wurden die mittleren Korrelationen zwischen den einzelnen Konstrukten und der „Overall Job Performance“ (Vorgesetztenurteil) bestimmt. Über alle Interviews hinweg betrachtet ergaben sich (unkorrigierte) Validitäten im Bereich von r = .13 bis .32. Die höchsten Korrelationen ergaben sich für folgende Konstrukte: Kreativität und Innovation (.32), Job-Wissen und Fähigkeiten (.23), Erfahrung und Arbeitsgeschichte (.27), Verträglichkeit (.28), Emotionale Stabilität (.26), Interpersonale Fähigkeiten (.21), Führung (.26) und Passung zur Organisation (.27). Alle Konstruktbereiche – außer physische Attribute sowie Interessen und Vorlieben – enthalten damit Konstrukte, die zur Vorhersage von Berufserfolg geeignet sind. Es zeigte sich weiterhin, dass die mittlere Validität strukturierter Interviews mit r = .39 deutlich höher ist, als für unstrukturierte Interviews (r = .24, beide Werte korrigiert). Insgesamt gab es zehn Konstrukte, für die mindestens vier Studien mit strukturierten Interviews zur Verfügung standen. Die

höchsten mittleren Validitäten erreichten dabei emotionale Stabilität (.31) und Passung zwischen Person und Organisation (.32), gefolgt von Extraversion, Führung, Überzeugen und Verhandlung (je .22), Gewissenhaftigkeit (.20), angewandte mentale Fähigkeiten (.19), Job-Wissen und Fähigkeiten (.18), Kommunikation (.17) und berufsbezogene Interessen (.14). Es lässt sich festhalten, dass die verschiedenen Dimensionen der Interviews größtenteils prognostische Validität besitzen. Ob aber mit den einzelnen, rational konstruierten Dimensionen auch die beabsichtigten Konstrukte erfasst werden, bleibt auch nach dieser Studie ungeklärt.

In einer weiteren aktuellen Metaanalyse untersuchen Salgado und Moscoso (2002) die Konstruktvalidität von konventionellen Interviews und strukturierten Verhaltensinterviews (meist BI und SI). Erstere beinhalten hauptsächlich Fragen zu Zeugnissen, Berufserfahrung und zur Selbstbewertung, während in letzteren hauptsächlich Fragen zur Berufserfahrung und zum Verhalten im Beruf gestellt werden (vgl. Janz, 1982). Insgesamt wurden die Zusammenhänge mit Variablen aus 11 Konstruktbereichen untersucht, nämlich allgemeine Intelligenz, Job-Wissen, Berufserfahrung, situationsbezogenes Urteil (inkl. „tacit knowledge“), Big Five Persönlichkeitsfaktoren, Notendurchschnitt und soziale Fähigkeiten. Diese Konstruktbereiche haben sich nach Salgado und Moscoso (2002) als valide Prädiktoren für Berufserfolg erwiesen und werden als relevante Konstrukte von Einstellungsinterviews diskutiert. Die Ergebnisse der Metaanalyse (alle Werte unkorrigiert) weisen darauf hin, dass konventionelle Interviews hauptsächlich soziale Fähigkeiten (mittlere Korrelation von .22), allgemeine Intelligenz (.20) und die Persönlichkeitsfaktoren emotionale Stabilität (.17) und Extraversion (.16) erfassen, während die (strukturierten) Verhaltensinterviews vor allem Berufserfahrung (.43), soziale Fähigkeiten (.34), Job-Wissen (.27) und situationsbezogenes Urteil (.22) erfragen und zu einem geringeren Ausmaß auch allgemeine Intelligenz (.14). Zusätzliche Analysen zeigen weiterhin, dass der Zusammenhang zwischen Intelligenz und SI mit .33 (korrigierte mittlere Korrelation) deutlich größer ist, als mit .19 für Intelligenz und BI (ebenfalls korrigierte mittlere Korrelation).

Die aufgeführten Studien zeigen, dass es noch an Metaanalysen fehlt, in denen die Zusammenhänge zwischen einzelnen Dimensionen des Interviews mit etablierten Verfahren zur Erfassung eben dieser Dimensionen untersucht werden.

1.3.2.1 Studien zur externen Konstruktvalidität des Multimodalen Interviews MMI

Auch die externe Konstruktvalidität des MMI war schon Gegenstand zahlreicher Untersuchungen. So entwickelten beispielsweise Schuler und Funke (1989) ein Interview zur Auswahl von Auszubildenden einer Bank. Ziel des Interviews war die Erfassung sozialer Kompetenz. Wie erwartet, ergaben sich höhere Korrelationen zwischen dem Abschneiden von 307 Bewerbern im MMI und sozialer Kompetenz (r = .60) als mit einem kognitiven Leistungstest (r = .21) sowie eine vernachlässigbare Korrelation mit Geschlecht (r = .10; leichte Bevorzugung von Frauen). Soziale Kompetenz wurde dabei auf sechs Verhaltensbeobachtungsskalen (Eigeninitiative, Argumentationsfähigkeit, Kundenorientierung, Selbstsicherheit, Kontaktfähigkeit, Teamfähigkeit) mit Hilfe von sechs Beobachtern pro Bewerber erfasst. Spätere Studien (z.B. Schuler et al., 1995) ergaben jedoch deutlich höhere Zusammenhänge zwischen MMI und kognitivem Fähigkeitstest (r = .37).

Das oben entwickelte Interview wurde von Schuler (1992) in leicht veränderter Form (zusätzlicher Block von situativen Fragen) auch in einem Bewerbungstraining mit 69 Studierenden eingesetzt. Als Kriterien zur Konstruktvalidierung dienten dabei ein AC, zwei Persönlichkeitsfragebögen, ein Test zur Leistungsmessung, ein Test zur sozialen Kompetenz, zwei Intelligenztests sowie die Schul- und Studienleistungen. Wie auch schon bei Schuler und Funke (1989) waren die Korrelationen mit den kognitiven Maßen eher niedrig. So korrelierte das MMI mit der Abiturnote zu r = -.08, mit der Vordiplomnote zu r = -.06 und mit den Intelligenztests zu r = .09 beziehungsweise r = .23. Weiterhin ergaben sich relativ niedrige Korrelationen zwischen MMI und solchen Persönlichkeitskonstrukten, die nach Schuler (1992) nicht Zielkonstrukt des Interviews waren (z.B. Lebenszufriedenheit r = .03, Soziale Orientierung r = .02), während die Korrelationen des Interviews mit den Zielkonstrukten eher hoch waren (z.B. Erfolgspotential r = .37, Dominanz r = .52). Dies galt insbesondere für das Biographische Interview. Die Korrelationen zwischen MMI und Leistungsmotivation bzw. sozialer Kompetenz waren mit r = .28 bzw. r = .56 ebenfalls hoch bis sehr hoch. In der gleichen Größenordnung waren auch die Zusammenhänge mit den verschiedenen AC-Übungen. Die höchste Korrelation ergab sich dabei zwischen MMI und strukturierter Gruppendiskussion (r = .57), etwas niedrigere Korrelationen ergaben sich für die unstrukturierte Gruppendiskussion und die Postkorbaufgabe (r = .22 und r = .27)

Eine weitere Studie (Schuler & Moser, 1995) war auf das Konstrukt Leistungsmotivation ausgerichtet. Wie erwartet ergaben sich hohe Korrelationen zwischen MMI und Leistungsstreben (r = .57) bzw. Ausdauer und Fleiß (r = .54). Anders als in der Studie von Schuler (1992), ergaben sich bei diesem Interview auch relativ hohe Korrelationen mit Abiturnoten und Vordiplom (r = .48 bzw. r = .39).
Zusammenfassend lässt sich aufgrund der obigen Studien festhalten: Je nach Zielsetzung können mit strukturierten Interviews eine Vielzahl von unterschiedlichen Konstrukten erfasst werden. Eine ähnliche Vorstellung findet sich auch schon bei Schuler (1992). Seines Erachtens existiert *das* Einstellungsinterview nicht. Vielmehr bezeichnet er das Interview als eine Kategorie diagnostischer Instrumente, die lediglich durch die Gesprächsform (Frage und Antwort) gekennzeichnet ist. Infolgedessen kommt er zu der Aussage: „Wir können das Interview als eine Art „Hülle“ auffassen, die inhaltlich mit verschiedenen Konstrukten gefüllt und formal durch verschiedene Methoden und Modi repräsentiert werden kann.“ (Schuler, 1992, S. 284).

1.3.3 Studie zur internen Konstruktvalidität

Experten im Bereich des Einstellungsinterviews (z.B. Maurer, Sue-Chan & Latham, 1999) fordern, dass konvergente und diskriminante Validität der innerhalb eines strukturierten Interviews erfassten Dimensionen mit Hilfe der Multitrait-Multimethod-(MTMM)-Matrix (vgl. Campbell & Fiske, 1959) untersucht werden sollten. Während dieser Ansatz der Konstruktvalidierung bei anderen multimodalen eignungsdiagnostischen Verfahren (z.B. Assessment Center, AC) lange Zeit als Methode der Wahl galt (vgl. Kleinmann & Köller, 1997), gibt es bei den strukturierten Interviews erst eine entsprechende Studie (Schuler, 1989; Schuler & Funke, 1989). In dieser Studie wurde ein MMI mit 307 Bewerbern für die Ausbildung zum Bankkaufmann durchgeführt. Im Rahmen der Auswertung wurde eine MTMM-Matrix mit sieben Dimensionen und zwei Methoden (situative und biographische Fragen) berechnet. Die Autoren kommen zu dem Ergebnis, dass sowohl im Biographischen als auch im Situativen Interview hohe Korrelationen zwischen den verschiedenen Dimensionen bestehen, also keine diskriminante Validität gegeben ist. Da diese heterotrait-monomethod Korrelationen außerdem höher waren als die meisten monotrait-heteromethod Korrelationen (konvergente Validität), aber niedriger als die meisten heterotrait-heteromethod Korrelationen, bezweifeln die Autoren, dass es

irgendeinen systematischen Effekt bezüglich der Zusammenhänge zwischen den Dimensionen gibt. Schuler und Funke (1989) führen drei potentielle Ursachen für diese niedrige Konstruktvalidität an: A) mangelnde Unabhängigkeit der Dimensionen bei der Konstruktion des Interviews (der Bereich Soziale Fähigkeiten ist zu eng, als dass sich sieben Subdimensionen unterscheiden lassen), B) die Beobachter sind unter Umständen einem Halo-Fehler unterlegen und C) die wahren Korrelationen zwischen den Dimensionen sind möglicherweise sehr hoch. Leider wird die empirische MTMM-Matrix weder in Schuler (1989) noch in Schuler und Funke (1989) aufgeführt, so dass eine Reanalyse der Daten nicht möglich ist.

Zusammenfassend lässt sich bis hierhin festhalten: Das Einstellungsinterview ist ein weitverbreitetes und gut akzeptiertes Verfahren, dessen strukturiert durchgeführten Vertreter (z.B. BI, SI, MMI) eine hohe prognostische Validität besitzen. Forschungsbedarf besteht jedoch hinsichtlich der Konstruktvalidität dieser Verfahren. So fehlen insbesondere MTMM-Analysen zur internen Konstruktvalidität strukturierter Interviews. Damit komme ich nun zu den Forschungsfragen der vorliegenden Arbeit.

1.4 Forschungsfragen der vorliegenden Arbeit

Wie oben dargestellt, ist die (interne) Konstruktvalidität strukturierter Interviews ein bisher vernachlässigtes Thema (Maurer et al., 1999). Dies ist sehr bedauerlich, da die Art der Konstrukte und die Qualität ihrer Erfassung direkten Einfluss auf die Kriteriumsvalidität des Interviews bei sich ändernden Rahmenbedingungen hat (vgl. Cronshaw & Wiesner, 1989). Die erste zentrale Fragestellung der Arbeit lautet daher:

1. Wie hoch ist die interne Konstruktvalidität strukturierter Interviews einzuschätzen? (Studie 1, Kapitel 2)

Wir gehen davon aus, dass die interne Konstruktvalidität strukturierter Interviews ähnlich niedrig wie beim AC ausfällt (vgl. Bycio, Alvares & Hahn, 1987; Robertson, Gratton & Sharpley, 1987; Sackett & Dreher, 1982). Erste empirische Hinweise auf ein solches Ergebnis liefert eine Studie von Schuler (1989 bzw. Schuler & Funke, 1989), welche den MTMM-Ansatz auf das MMI angewandt hat. Um die Vergleichbarkeit unserer Ergebnisse zu sichern, greifen wir daher ebenfalls auf das

MMI zurück, welches sich als prognostisch valides und akzeptiertes Verfahren bewährt hat (vgl. Schuler, 1992; Schuler & Moser, 1995).

Bei Bestätigung der obigen Annahme sollen in zwei weiteren Studien mögliche Einflussfaktoren auf die Konstruktvalidität strukturierter Interviews – wiederum am Beispiel des MMI – untersucht werden. Entsprechende Arbeiten sind bisher unbekannt, weswegen einschlägige Ergebnisse der AC-Forschung herangezogen werden. Ein solches Vorgehen erscheint plausibel, da AC und MMI trotz elementarer Unterschiede auch einige Gemeinsamkeiten aufweisen (s.a. Höft & Schuler, 2001; Schuler, 1996). So handelt es sich bei beiden Verfahrensgruppen um prognostisch valide, simulationsorientierte, standardisierte, multimodale, eignungsdiagnostische Instrumente mit interaktivem Charakter. Beide Verfahren werden aufgrund von Anforderungsanalysen entwickelt, wodurch ein hoher Bezug zum (zukünftigen) Arbeitsplatz gewährleistet ist. Die Beurteilung erfolgt aufgrund von Fremd-einschätzungen durch meist mehrere Beobachter und mit Hilfe von verhaltensverankerten Beurteilungsskalen bzw. Dimensionen (eigenschafts-heterogener Anforderungsbereich), so dass ähnliche Prozesse und Mechanismen der Informationsverarbeitung bzw. der Eindrucks- und Entscheidungsbildung zu erwarten sind.

Es gibt eine Vielzahl von Studien, welche Moderatoren der Konstruktvalidität im AC untersucht haben. Lievens (1998) ordnet die in 21 Studien ermittelten Einflussfaktoren den folgenden vier Bereichen zu: Dimensionen (z.B. geringe Anzahl von Dimensionen), Beobachter (z.B. Teilnahme von Psychologen), Übungen (z.B. Training der Rollenspieler) und Bewertungsprozess (z.B. Verwendung verhaltensverankerter Checklisten). Während die dort berichteten Einflussfaktoren meist „technischer“ Natur sind und neue Leitlinien für die Konstruktion und Durchführung von AC nach sich ziehen, weisen Arthur, Woehr und Maldegen (2000) auf eine gänzlich andere Erklärung hin. So ist es möglich, dass im AC nicht die intendierten Dimensionen (Durchsetzungsfähigkeit, Ausdruck, etc.) erfasst werden, sondern stattdessen andere (Persönlichkeits-)Konstrukte, wie z.B. Self-monitoring, Impression-management oder Role congruency. Eine ähnliche Vorstellung wird auch von Kleinmann (1993, 1997a, 1997b) vertreten. Kleinmann (1993) konnte zeigen, dass die AC-Teilnehmer unterschiedlich fähig sind, die relevanten Anforderungs-dimensionen zu erkennen, und dass das Ausmaß des Erkennens Einfluss auf die konvergente Validität hat. Die Annahmen von Kleinmann (1993, 1997b) sind auch

insofern interessant, als dass sie eine plausible Erklärung für das sogenannte Konstrukt-Kriteriumsvaliditäts-Paradox liefern (vgl. Arthur et al., 2000). Dieses Paradox bezeichnet den Sachverhalt, dass AC trotz geringer Konstruktvalidität eine nachgewiesene Kriteriumsvalidität besitzen. Der Zusammenhang zwischen Prädiktor (z.B. AC, Interview) und Kriterium (z.B. Vorgesetztenurteil) wird nun dadurch erklärt, dass die Fähigkeit, Anforderungsdimensionen zu erkennen, beide Messungen beeinflusst und so gemeinsame Varianz herstellt (Alternativerklärungen bei Klimoski & Brickner, 1987).

Damit komme ich nun zum Hauptziel der vorliegenden Arbeit, der Übertragung des Forschungs-Ansatzes von Kleinmann auf das strukturierte Einstellungsinterview. Die zentralen Fragestellungen der Studien 2 und 3 lauten daher:

2. Hat das Ausmaß, in dem Bewerber die Anforderungsdimensionen eines strukturierten Interviews erkennen, Einfluss auf die konvergente Validität und die gezeigte Leistung? (Studie 2, Kapitel 3)

3. Hat die Bekanntgabe der Anforderungsdimensionen (Transparenz) Einfluss auf die Konstrukt- und Kriteriumsvalidität eines strukturierten Interviews? (Studie 3, Kapitel 4)

Nach Bearbeitung dieser drei Fragestellungen sollen die Ergebnisse abschließend im Rahmen einer Gesamtdiskussion studienübergreifend diskutiert werden. Hierbei wird es insbesondere um die Frage gehen, ob strukturierte Interviews in der Lage sind, die aus einer Anforderungsanalyse gewonnenen Dimensionen zu messen, und wie sich die interne Konstruktvalidität strukturierter Interviews verbessern lässt.

2 Konvergente und diskriminante Validität des strukturierten Interviews: Some troubling empirical findings?

Zusammenfassung. In der vorliegenden Studie (*N* = 110) wird die Konstruktvalidität eines Multimodalen Interviews untersucht. Während die Analysen der Multitrait-Multimethod-Matrix dabei auf eine geringe interne Konstruktvalidität hinweisen (mittlere konvergente Validität von .24 und mittlere diskriminante Validität von .41), ergeben sich bei Betrachtung des nomologischen Netzwerkes positive Anhaltspunkte auf externe Konstruktvalidität. In einer weiteren Auswertung werden die Effekte auf die Multitrait-Multimethod-Matrix untersucht, welche sich dadurch ergeben, dass identische Dimensionen teilweise durch identische Beobachter und teilweise durch verschiedene Beobachter beurteilt werden (gemeinsame Beobachtervarianz). Diese stellen sich – anders als beim Assessment Center – als vernachlässigbar heraus.

2.1 Einleitung

Die prognostische Validität strukturierter Interviews war Gegenstand zahlreicher Studien und wurde in einer Vielzahl von Metaanalysen bestätigt (Huffcutt & Arthur, 1994; McDaniel et al., 1994; Schmidt & Hunter, 1998; Schmidt & Rader, 1999; Wiesner & Cronshaw, 1988; Wright, Lichtenfels & Pursell, 1989). Hingegen wurde die Konstruktvalidität strukturierter Interviews lange Zeit vernachlässigt (Harris, 1999; Maurer et al., 1999; Schuler, 1989) und erst seit kurzer Zeit gibt es zwei Metaanalysen (Huffcutt et al., 2001; Salgado & Moscoso, 2002), welche die mit Hilfe des strukturierten Interviews erfassten Konstrukte untersuchen. Dabei greifen die Autoren der dort zitierten Studien ausnahmslos auf die Idee des nomologischen Netzwerkes zurück (vgl. Cronbach & Meehl, 1955), d.h. sie untersuchen die Zusammenhänge zwischen Interview und Messinstrumenten, welche außerhalb des strukturierten Interviews stehen. Diese Art der Konstruktvalidität wird von Kolk (2001) als *externe* Konstruktvalidität bezeichnet. Nach Campbell und Fiske (1959) ist es jedoch notwendig, dass man vor der Untersuchung der Kriteriumsvalidität und des nomologischen Netzwerks zunächst Vertrauen in den Test als solchen gewinnt.

Daher empfehlen die Autoren die Untersuchung der Multitrait-Multimethod-(MTMM)-Matrix (Campbell & Fiske, 1959), welche Rückschlüsse auf die *interne* Konstruktvalidität (vgl. Kolk, 2001) zulässt und das gängige Paradigma in der AC-Forschung darstellt. Trotz wiederholter Forderungen (z.B. Maurer et al., 1999) gibt es jedoch, wie bereits erwähnt, erst eine Studie (Schuler, 1989 bzw. Schuler & Funke, 1989), welche die Konstruktvalidität eines strukturierten Interviews mit Hilfe der MTMM-Matrix untersucht. Diese Studie kommt zu einem vernichtenden Urteil bezüglich konvergenter und diskriminanter Validität. Forschungsgegenstand der vorliegenden Studie ist die interne Konstruktvalidität strukturierter Einstellungsinterviews. Zunächst muss ein dimensionsbasiertes Interview entwickelt werden, welches verschiedene Methoden (z.B. im Sinne verschiedener Fragetechniken) umfasst. Ein entsprechendes Verfahren stellt das Multimodale Interview (MMI) von Schuler (1992) dar, dessen Komponenten Selbstpräsentation (SP), Biographisches Interview (BI, Janz, 1989) und Situatives Interview (SI, Latham et al., 1980) als besonders valide gelten (Schuler & Moser, 1995).

2.1.1 Potentielle Einflussfaktoren auf die interne Konstruktvalidität

Bei der Entwicklung unseres MMI sollen mögliche Einflussfaktoren auf die interne Konstruktvalidität berücksichtigt werden. Entsprechende Faktoren wurden bisher noch nicht explizit für strukturierte Interviews untersucht. Aufgrund der Ähnlichkeiten zwischen MMI und AC (vgl. Kapitel 1.4) scheint es angemessen, auf Ergebnisse der AC-Forschung zurückzugreifen. Dort sind inzwischen eine ganze Reihe von Merkmalen beschrieben, welche sich nachweislich positiv auf die Konstruktvalidität des AC auswirken (vgl. Arthur et al., 2000; Lievens, 1998). Bei der Konstruktion unseres MMI sollen folgende Einflussfaktoren explizit beachtet werden:

1. Geringe Anzahl von Beobachtungsdimensionen (Gaugler & Thornton, 1989)
2. Gute Beobachtbarkeit und begriffliche Unabhängigkeit der Dimensionen (Kleinmann, Exler, Kuptsch & Köller, 1995)
3. Art der Beobachter (Psychologen, vgl. Sagie & Magnezy, 1997)
4. Art des Beobachtertrainings (Frame-of-Reference-Training, vgl. Woehr & Huffcutt, 1994; Schleicher, Day, Mayes & Riggio, 2002).
5. Geringes Verhältnis von Anzahl Beobachter zu Teilnehmer (Gaugler, Rosenthal, Thornton & Bentson, 1987)

6. Hohe Reliabilität der Messungen durch Verhaltensanker (vgl. Metaanalyse zum Einstellungsinterview von Conway et al., 1995)

Durch den Einbezug verschiedener potentieller Einflussfaktoren kann man die vorliegende Studie als vernünftigen Versuch auffassen, ein strukturiertes Interview zu konstruieren, welches konvergente und diskriminante Validität zeigen sollte, sofern dies eine Eigenschaft des MMI sein sollte.

2.1.2 Ziele der Studie

Damit komme ich nun zu den Zielen der vorliegenden Studie:

1. Prüfung der internen Konstruktvalidität des strukturierten Interviews mit adäquaten Auswertungsmethoden am Beispiel des MMI
2. Explorative Analyse von Korrelaten des strukturierten Interviews mit einer Vielzahl unterschiedlicher Verfahren
3. Kontrolle des Einflussfaktors „gemeinsame Beobachtervarianz“ auf konvergente und diskriminante Validität

Ad 1: Prüfung der internen Konstruktvalidität

Die von Campbell und Fiske (1959) vorgeschlagene MTMM-Matrix ist nach Marsh und Grayson (1995) die verbreitetste Methode zur Bestimmung der Konstruktvalidität. Trotzdem gibt es auch nach mehr als 40 Jahren immer noch keine einheitlichen Leitlinien bezüglich ihrer Auswertung. Im Laufe der Zeit wurden eine Vielzahl von Auswertungsmethoden vorgeschlagen (vgl. Tomás, Hontangas & Oliver, 2000). Außerdem wurde in den letzten Jahren verstärkt die Theorie der Generalisierbarkeit diskutiert (vgl. Arthur et al., 2000; Lievens, 2001). Zusammenfassend hat sich nach Becker und Cote (1994) die konfirmatorische Faktorenanalyse (CFA) als die am häufigsten eingesetzte Methode – in Alternative zu den von Campbell und Fiske (1959) vorgeschlagenen Kriterien – herausgestellt. Nach Ansicht verschiedener Autoren (z.B. Kleinmann & Köller, 1997) ist sie auch die adäquate Methode zur Auswertung von MTMM-Matrizen. Daher soll die Auswertung der internen Konstruktvalidität sowohl mit Hilfe von MTMM-Matrizen als auch mit Hilfe der konfirmatorischen Faktorenanalyse erfolgen.

Ad 2: Korrelate des strukturierten Interviews

Zur explorativen Analyse von Korrelaten des MMI kommen die folgenden Verfahren zum Einsatz, deren Verwendung ausführlich bei Hartstein (2003) begründet wird:

- ein speziell für die Untersuchung konstruiertes AC;
- der Intelligenz-Struktur-Test 2000 (IST 2000) von Amthauer, Brocke, Liepmann und Beauducel (1999);
- die Textilfabrik (PC-gestützte Simulation zur Erfassung der Problemlösefähigkeit) von Hasselmann und Strauß (1995, siehe auch Hasselmann, 1993);
- der Fragebogen zu Einstellungen und Selbsteinschätzungen (FES) von Marcus und Schuler (1998). Dieses Verfahren gehört zur Gruppe der Integritätstests und dient zur Erfassung kontraproduktiver Verhaltensweisen im Betrieb, wie z.B. Neigung zur Gewalttätigkeit, Diebstahl, Alkoholismus und Absentismus (vgl. auch Marcus, 2000);
- ein Fragebogen zur sozialen Erwünschtheit (SE) von Crowne und Marlowe (1960) in der Übersetzung von Lück und Timaeus (1969)
- sowie ein Fragebogen zum Self-Monitoring (SM) von Novak und Kammer (unveröffentlicht) in der Version von Mielke und Kilian (1990).

Ad 3: Gemeinsame Beobachtervarianz

Weiterhin soll in einer dritten Fragestellung die Art der Datenintegration bzw. -auswertung („within-exercise“ vs. „within-dimension“) als möglicher Moderator der Konstruktvalidität beachtet werden (vgl. Silverman, Dalessio, Woods & Johnson, 1986; Kolk, 2001). In der Forschung zur Konstruktvalidität des AC haben sich sowohl die geringen monotrait-heteromethod-Korrelationen (niedrige konvergente Validität) als auch die hohen heterotrait-monomethod-Korrelationen (niedrige diskriminante Validität) als problematisch herausgestellt. Diese Korrelationen führen dazu, dass bei der Analyse von MTMM-Matrizen meist starke Übungsfaktoren und schwache Dimensionsfaktoren gefunden werden, was dann als geringe Konstruktvalidität interpretiert wird.

Zur Begründung der hohen heterotrait-monomethod-Korrelationen werden verschiedene Argumentationsstränge herangezogen. Einerseits ist vorstellbar, dass es sich um „echte Übungseffekte“ handelt. So ist denkbar, dass manche Teilnehmer

grundsätzlich bei einer Präsentation besser abschneiden als bei einer führerlosen Gruppendiskussion (bzw. umgekehrt). Silverman et al. (1986) hingegen führen die hohen heterotrait-monomethod-Korrelationen auf die Art der Datenintegration zurück, welche übungsweise oder dimensionsweise erfolgen kann. So konnten Silverman et al. (1986) in einer experimentellen Studie zeigen, dass die Konstruktvalidität steigt, wenn die Beobachter die Bewertungen innerhalb der Dimensionen vornehmen, anstatt innerhalb der Übungen. Das Ergebnis dieser Studie konnte allerdings weder von Harris, Becker und Smith (1993) noch von Kleinmann, Andres, Fedtke, Godbersen und Köller (1994) repliziert werden.

Nach Robertson, Gratton und Sharpley (1987) ist nicht nur die gemeinsame Übungsvarianz für die hohen heterotrait-monomethod-Korrelationen verantwortlich. Hinzu kommt, dass die verschiedenen Dimensionen innerhalb einer Übung durch die gleichen Beobachter bewertet werden. Die erhöhten heterotrait-monomethod-Korrelationen sind daher auch durch die sogenannte gemeinsame Beobachtervarianz erklärbar, welche auf Beobachtungsfehler, wie z.B. „Halo“, zurückgeführt werden kann. Die gemeinsame Beobachtervarianz fehlt jedoch aufgrund von Beobachterrotation üblicherweise in den monotrait-heteromethod-Korrelationen, so dass die Messung der konvergenten Validität nicht von dieser artifiziellen Varianzquelle profitieren kann. Kolk (2001) hat in zwei Studien den Einfluss der gemeinsamen Beobachtervarianz auf die Konstruktvalidität des AC untersucht und bestätigt. Während Kolk (2001) in der ersten Studie den Einfluss der gemeinsamen Beobachtervarianz experimentell variierte, wurde in der zweiten Studie durch eine spezielle Auswertung der MTMM-Matrix das Ausmaß an gemeinsamer Beobachtervarianz manipuliert. Ein ähnliches Vorgehen wird hier gewählt.

Weiterhin soll in der vorliegenden Studie die Reliabilität des MMI untersucht werden. Diese stellt ebenfalls ein grundlegendes Kriterium zur Beurteilung der Qualität eines Messinstrumentes dar. So kommen beispielsweise Cronshaw und Wiesner (1989) in ihrer Metaanalyse zu dem Ergebnis, dass Reliabilität und Validität von Einstellungsinterviews zu .48 korrelieren. Aus diesem Grund werden in Studie 1 die entsprechenden Kennwerte (Beurteilerübereinstimmung und innere Konsistenz) ausführlich dargestellt. Eine ausführliche Darstellung der statistischen Analysen, wie sie im Rahmen von Fragestellung 1 und 3 durchgeführt werden, findet sich am Ende des Methodenteils.

2.2 Methode

2.2.1 Überblick

Zur Erhebung der Daten wurden zweitägige Bewerbertrainings für die Position eines Management-Trainees mit interessierten Berufstätigen und Studierenden am Ende des Studiums durchgeführt. Der erste Tag bestand aus AC-Übungen und MMI, der zweite Tag aus verschiedenen Leistungs- und Persönlichkeitstests. Die Motivation der Teilnehmer bestand darin, ein AC aus eigener Erfahrung kennen zu lernen sowie individuelles Feedback über ihr Verhalten – insbesondere ihre Stärken und Schwächen – zu erhalten. Die Anwerbung der Teilnehmer erfolgte hauptsächlich über eine eigene Internetseite (www.ac-bewerbungstraining.de), aber auch mit Hilfe von Aushängen und Handzetteln sowie Berichten und Interviews in regionalen und überregionalen Tageszeitungen. Zur Erhöhung des Commitments mussten die Teilnehmer im Vorfeld der Untersuchung eine Teilnahmegebühr von 15 € entrichten. Die im Rahmen der Studie entwickelten Messinstrumente kamen teilweise auch in Studie 2 und 3 zum Einsatz. Daher wird ihre Konstruktion etwas ausführlicher als üblich dargestellt.

2.2.2 Vorversuch: Entwicklung von Assessment Center (AC) und Multimodalem Interview (MMI)

Es wurde ein dimensionsbasiertes MMI mit den Komponenten Selbstdarstellung, biographische und situative Fragen entwickelt. Diese Elemente haben sich nach Schuler und Moser (1995) als besonders kriteriumsvalide erwiesen. Außerdem wurde zur externen Konstruktvalidierung ein AC entworfen, das auf den gleichen Dimensionen wie das MMI basiert. Bei der Entwicklung unseres Interviews konnten wir auf andere Arbeiten zurückgreifen, die ebenfalls situative und biographische Fragen aufgrund von Jobanalysen entwickelt hatten.

Zentral ist die Arbeit von Borchert (2001), welche in Zusammenarbeit mit Praktikern aus dem Bereich der Personalauswahl ein MMI entwickelt hat. Hierzu wurde eine Jobanalyse durchgeführt und typische Alltagssituationen aus dem Arbeitsleben eines Management-Trainees bzw. eines Abteilungsleiters gesammelt. Die zur erfolgreichen Bewältigung dieser Situationen relevanten Verhaltensweisen wurden anschließend zu den vier Dimensionen Arbeitsorganisation/Planung, Führungsverhalten,

Informationsverhalten und Kooperation zusammengefasst. Dimensionen mit ähnlichem Bedeutungsgehalt werden häufig in der Praxis verwendet (vgl. Jeserich; 1981).

Weitere situative Fragen wurden den Arbeiten von Deller (1991) und Klehe (2000) entnommen. Das Situative Interview bei Deller (1991) umfasst in seiner Vorform 31 Fragen. Es wurde für einen Druckmaschinenhersteller entwickelt und zur Auswahl von Hochschulabgängern für die Position des Länderreferenten genutzt. Klehe (2000) konnte in zwei Vorversuchen die meisten der Fragen den vier Dimensionen Verhandlungsgeschick/Überzeugungsfähigkeit, Kunden- & Serviceorientierung, Kooperation und Problemlöseorientierung zuordnen.

Weitere Fragen für unser biographisches Interview und die Selbstvorstellung beruhen auf Arbeiten von Schuler und Kollegen (Deutscher Sparkassen- und Giroverband, 1988). Die dort aufgeführten Items wurden für die Auswahl von Auszubildenden im Bankenbereich entwickelt und mussten daher auf die neue Zielpopulation Hochschulabgänger angepasst werden.

Aufgrund der beschriebenen Voruntersuchungen können die verwendeten Fragen als relevant für die Auswahl von Management-Trainees angenommen werden.

2.2.2.1 Workshop1: Auswahl der Dimensionen (Prüfung von Beobachtbarkeit und Unabhängigkeit)

Neben der Voraussetzung einer geringen Anzahl zu beobachtender Anforderungsdimensionen (Gaugler & Thornton, 1989) wurden auch die Beobachtbarkeit und Unabhängigkeit der Dimensionen als Einflussfaktoren auf die Konstruktvalidität identifiziert (Kleinmann et al., 1995, Haaland & Christiansen, 2002). Im Vorfeld des Workshops wurden durch drei Experten der Personalauswahl neun verhaltensverankerte Beobachtungsdimensionen zusammengestellt. Diese sollten sowohl gut beobachtbar als auch relativ unabhängig voneinander sein. Außerdem sollten sich die Dimensionen an den situativen und biographischen Fragen orientieren, die zur Entwicklung des MMI zur Verfügung standen (s.o.). Der eigentliche Workshop wurde dann mit zehn erfahrenen studentischen Beobachtern analog zu Kleinmann et al. (1996) bzw. Kleinmann (1997b) durchgeführt.

Zunächst wurden die Dimensionen und Verhaltensanker den Beobachtern ausführlich vorgestellt und anhand verschiedener AC-Übungen im realitätsnahen Einsatz auf ihre prinzipielle Beobachtbarkeit hin bewertet. Um die Ergebnisse quantifizierbar zu machen, vergaben die Beobachter Ratings von 1 (schlecht

beobachtbar) bis 10 (sehr gut beobachtbar). Es zeigte sich, dass die Mittelwerte der Ratings von 6.1 bis 9.8 reichten. Damit können die Dimensionen als ausreichend gut beobachtbar gelten. In einem zweiten Schritt wurden die Dimensionen durch die Beobachter auf ihre Unabhängigkeit hin bewertet. Zunächst ordneten die Beobachter ihre verhaltensnah formulierten Beobachtungen (z.B. löst einen Konflikt zwischen zwei Teilnehmern) aus dem ersten Teil des Workshops den verschiedenen Dimensionen zu (z.B. Soziale Kompetenz und Steuerung sozialer Prozesse). Anschließend wurden alle 36 möglichen Paare von Dimensionen gebildet und die Beobachter schätzten anhand eines 10-stufigen Ratings die Unabhängigkeit der Dimensionen ein (1 = unabhängig, 4 = eher unabhängig, 7 = eher abhängig, 10 = abhängig). Als unabhängig wurden zwei Dimensionen bezeichnet, wenn die beobachteten Verhaltensweisen eindeutig einer einzigen der beiden Dimensionen zugeordnet werden konnten. Als abhängig hingegen wurden zwei Dimensionen bezeichnet, wenn sie eine Vielzahl gemeinsamer Verhaltensweisen enthielten. Anschließend wurde durch Mittelung der Ratings die Unabhängigkeit der Dimensionen für jedes Dimensionspaar bestimmt. Bei den 36 Paaren ergaben sich Werte im Bereich 1.7 bis 6.4. Höhere Werte ergaben sich insbesondere für Paare, welche die Dimension Soziale Kompetenz enthielten, welche daher eliminiert wurde. Die verbleibenden 28 Dimensionspaare wurden hingegen als ausreichend unabhängig angesehen. Lediglich in 4 Fällen zeigten sich mittlere Unabhängigkeitsratings mit einem Wert größer als 5. Als Ergebnis des Workshops ergaben sich somit die folgenden acht gut beobachtbaren und relativ unabhängigen Dimensionen und zugehörigen Verhaltensanker:

1. Systematisches Denken & Handeln (SDH; strukturiert komplexe Sachverhalte; erkennt das Wesentliche; erkennt Zusammenhänge; formuliert Ziele; setzt Prioritäten),
2. Zusammenarbeit (ZU; geht Kompromisse ein; schafft „win win“-Situationen; berücksichtigt Bedürfnisse / Wünsche anderer; bezieht andere aktiv mit ein; geht mit anderen fair um),
3. Steuerung sozialer Prozesse (SSP; weist Aufgaben zu / delegiert; steuert die Diskussion/ einzelne Gesprächssequenzen; vertritt den eigenen Standpunkt; übernimmt Verantwortung; kontrolliert Ergebnisse und Prozesse),

4. Umgang mit Informationen (UI; nutzt möglichst immer mehrere, unabhängige Infoquellen; verschafft sich selbständig nötige Informationen; versorgt andere mit Informationen; hält keine wichtigen Informationen zurück; sorgt dafür, dass Informationen an die richtigen Stellen gelangen),
5. Wirtschaftliches Denken (WD; berücksichtigt Konsequenzen und Langzeitfolgen seiner Handlungen für das Unternehmen; behält den wirtschaftlichen Erfolg im Blick; ist dem Unternehmen gegenüber loyal; berücksichtigt Unternehmensziele bei seinem Vorgehen; argumentiert unter Berücksichtigung von wirtschaftlichen Gesichtspunkte),
6. Ausdruck/mündliche Formulierung (AF; ist akustisch zu verstehen; formuliert flüssig und ansprechend / plastisch; Sätze sind übersichtlich und klar strukturiert; benutzt verständlichen Wortschatz; setzt Gestik / Mimik angemessen ein),
7. Engagement/Initiative (EI; Beteiligung ist konstant hoch; meldet sich häufig zu Wort; setzt sich motiviert mit der Thematik auseinander; arbeitet während der ganzen Übung konzentriert mit; zeigt sich interessiert),
8. Fachwissen (FW; analysiert schnell die Fakten und kann sie anhand eines theoretischen Fachhintergrundes effektiv verwerten; benutzt Fachausdrücke; beteiligt sich an fachspezifischen Diskussionen; zeigt interdisziplinäres Wissen; ist fachlich kompetent).

In Studie 1 gingen nur die ersten drei Dimensionen ein, da für diese bereits eine Vielzahl von situativen und biographischen Fragen vorhanden waren. Die weiteren Dimensionen wurden später für eine Folgestudie (Kapitel 3) benötigt.

2.2.2.2 Workshop 2: Auswahl und Überarbeitung der Interviewfragen

Im Vorfeld des zweiten Workshops wurden durch zwei Experten der Personalauswahl insgesamt 34 situative und biographische Fragen gesammelt, die zur Erfassung der drei Dimensionen – Systematisches Denken & Handeln (SDH), Zusammenarbeit (ZU) und Steuerung sozialer Prozesse (SSP) – brauchbar erschienen. Als Grundlage dienten die oben beschriebenen Arbeiten von Borchert (2001), Deller (1991), Klehe (2000) und Schuler und Kollegen (Deutscher Sparkassen- und Giroverband, 1988). Im eigentlichen Workshop wurde dann mit Hilfe von zehn erfahrenen studentischen Beobachtern (andere Personen als im Workshop 1) die Verständlichkeit der Interviewfragen und der verhaltensverankerten Beobachtungsskalen überprüft sowie die Zuordnung von Fragen zu Dimensionen

vorgenommen. Die Verständlichkeit der Interviewfragen wurde überprüft, indem die Fragen im Plenum vorgelesen und diskutiert wurden. Nur in seltenen Fällen waren die Fragen unklar bzw. missverständlich und mussten verändert werden. Anschließend wurden die Beobachter – ähnlich wie im ersten Workshop – mit den acht Dimensionen vertraut gemacht. Soweit notwendig wurden die Verhaltensanker umgearbeitet, so dass sie im Kontext Interview sinnvoll erschienen (beispielsweise „sagt, dass er planvoll vorgeht“ anstatt „geht planvoll vor“). Nun ordneten die Beobachter in Einzelarbeit die Interviewerfragen den Dimensionen zu. Hierbei verwendeten die Beobachter Ratings von 1 (erfasst Dimension X nicht) bis 10 (erfasst Dimension X sehr gut).

Die letztendliche Formulierung und Auswahl der Fragen erfolgte durch zwei Experten der Personalauswahl. Grundlage waren einerseits die qualitativen und quantitativen Ergebnisse aus dem zweiten Workshop (Ratings bezüglich der Zuordnung der Fragen zu Dimensionen, Diskussion optimaler Antworten). Zum anderen konnten wir auf statistische Informationen aus den oben genannten Arbeiten zurückgreifen (insbesondere Verteilung der Antworten, Itemschwierigkeit, Trennschärfe und Beobachter-Übereinstimmung). Wichtigstes Kriterium bei der endgültigen Auswahl der Fragen war jedoch die eindeutige Zuordnung der Frage zu einer einzigen Dimension (vgl. Workshop 2). Als Ergebnis erhielten wir zwölf situative und zwölf biographische Fragen, welche mit jeweils vier Fragen die drei Dimensionen – Systematisches Denken & Handeln (SDH), Zusammenarbeit (ZU) und Steuerung sozialer Prozesse (SSP) – erfassten. Die Instruktion für die Selbstpräsentation wurde in Anlehnung an Schuler und Kollegen (Deutscher Sparkassen- und Giroverband, 1988) formuliert. Die Teilnehmer wurden gebeten sich selbst kurz vorzustellen. Dabei sollten sie insbesondere darüber berichten, was sie bislang gemacht haben, wo ihre Stärken und Schwächen liegen und was sie erreichen möchten.

2.2.2.3 Workshop 3: Auswahl der Assessment Center (AC) Übungen

Im Vorfeld des dritten Workshops wurden durch drei Experten der Personalauswahl insgesamt acht AC-Übungen (vier Gruppendiskussionen, zwei Rollenspiele und zwei Postkörbe) herausgesucht, die zur Erfassung der insgesamt acht Dimensionen brauchbar erschienen, wobei jedoch insbesondere die drei relevanten Dimensionen (SDH, ZU und SSP) beobachtbar sein sollten. Im eigentlichen Workshop nahmen dann zwölf erfahrene studentische Beobachter (andere Personen als im ersten bzw. zweiten Workshop) teil. Ziel des Workshops war es, die Relevanz der Dimensionen

für die verschiedenen AC-Übungen zu prüfen. Hierzu wurden die Beobachter zunächst mit den acht Dimensionen vertraut gemacht. Anschließend wurden alle AC-Übungen gemeinsam durchgespielt, wobei jeweils ein Teil der Beobachter als Teilnehmer fungierte. Nach jeder Übung brachten die Beobachter die insgesamt acht Dimensionen in eine Rangreihe, so wie sie für das erfolgreiche Abschneiden in dieser Übung als wichtig angesehen wurden. Aufgrund dieser Ergebnisse wurden zwei Übungen (führerlose Gruppendiskussionen, einmal mit und einmal ohne Rollenvorgabe) ausgesucht, bei denen die drei relevanten Dimensionen – Systematisches Denken & Handeln (SDH), Zusammenarbeit (ZU) und Steuerung sozialer Prozesse (SSP) – übereinstimmend am relevantesten eingeschätzt wurden sowie je drei Dimensionen, die zwar beobachtbar, aber weniger relevant für den Erfolg bei dieser Übung waren (diese Dimensionen wurden für eine Folgestudie benötigt).

Abschließend wurden die verschiedenen Auswertungshinweise in den Postkörben soweit möglich mit Hilfe der Beobachter den verschiedenen Dimensionen zugeordnet (z.B. „Teilnehmer erkennt Kollision von Termin A und B“ wurde der Dimension Systematisches Denken & Handeln zugeordnet). Hierdurch war es möglich, im Postkorb nicht nur einen Gesamtpunktwert zu erhalten, sondern auch noch Punktwerte auf den verschiedenen Dimensionen.

2.2.3 Hauptversuch

Die Bewerbungstrainings wurden im Sommersemester 2001 an den Universitäten in Marburg und Gießen[2] durchgeführt. An jedem Training nahmen zwei Moderatoren, acht Beobachter und maximal acht Teilnehmer teil. Zwischen Vormittag (AC-Übungen) und Nachmittag (MMI) wechselte die Zuordnung von Teilnehmern und Beobachtern. Auf diese Weise wurde sichergestellt, dass die Beurteilung der Teilnehmer im MMI unabhängig vom gezeigten Verhalten im AC bewertet wurde.

2.2.3.1 Teilnehmer

Insgesamt wurden 15 zweitägige Trainings mit 110 Teilnehmern und Teilnehmerinnen durchgeführt (53 Frauen und 57 Männer); für 108 Personen liegen die vollständigen Daten vor. Das Alter der Teilnehmer lag zwischen 21 und 36 Jahren mit einem Mittelwert von 26.94 Jahren (*SD* = 2.89). Die meisten Teilnehmer kamen

[2] Wir danken Herrn Prof. Dr. Michael Frese für seine freundliche Unterstützung.

aus den Wirtschaftswissenschaften (42.1%) bzw. aus den Naturwissenschaften (21.5%). Weitere Teilnehmer stammten aus den Bereichen Jura (6.5%), Psychologie (5.5%), Theologie (3.6%) und sonstigen Naturwissenschaften (11.8%). Die mittlere Studiendauer über alle Teilnehmer hinweg betrug 9.71 Semester (*SD* = 3.77). Insgesamt 52 Teilnehmer (47.3%) verfügten über erste Berufserfahrungen, wobei jedoch erst 31 Teilnehmer (28.2%) ihr Studium beendet hatten. Nur 7 Teilnehmer (6.3%) hatten bereits Erfahrungen mit AC.

2.2.3.2 Beobachter

Es wurden insgesamt 32 Beobachter in eintägigen Trainings geschult. Die Beobachter waren zum größten Teil Studierende der Psychologie im Hauptstudium mit dem Schwerpunkt Arbeits- und Organisationspsychologie und Kenntnissen im Bereich AC, die auf diese Weise ihre fachliche Qualifikation erhöhen wollten. Die Beobachter erhielten während der gesamten Datenerhebung keine Informationen über die der Studie zugrunde liegenden Hypothesen.

2.2.3.3 Beobachtertraining

Die Beobachterschulungen wurden im Sinne eines „Frame-of-Reference-Trainings“ (FOR) durchgeführt, da sich dieses Training als besonders effektiv herausgestellt hat (vgl. Woehr & Huffcutt, 1994). Im FOR-Ansatz werden die Besonderheiten der einzelnen Dimensionen betont, ein einheitliches Verständnis der definierten Dimensionen hergestellt, die Nutzung von Verhaltensankern geübt und praktische Erfahrungen im Umgang mit den verschiedenen Übungen gewonnen. Eine ausführliche Darstellung eines FOR-Trainings findet sich bei Arthur et al. (2000).

Zunächst wurden die Beobachter mit den drei aus dem Vorversuch hervorgegangenen Anforderungsdimensionen (Systematisches Denken & Handeln, Zusammenarbeit, Steuerung sozialer Prozesse) und den Übungen vertraut gemacht.. Dies wurde unter anderem dadurch erreicht, dass die Beobachter Verhaltensbeispiele erhielten und diese den verschiedenen Dimensionen zuordnen mussten. Anschließend wurde genaues Beobachten geübt. Hierzu wurden einerseits Videos vorgeführt, zum anderen die Übungen (teilweise) live durchgespielt. Mehrmals wurden die Beobachter auf denkbare Beobachtungsfehler hingewiesen und Möglichkeiten ihrer Vermeidung diskutiert. Insbesondere wurde betont, das Beobachtung und Bewertung des Teilnehmerverhaltens in zwei getrennten Schritten erfolgen soll. Nach der individuellen Bewertung der beobachteten Verhaltensweisen diskutierten die Beobachter in wechselnd zusammengesetzten Kleingruppen ihre

Beobachtungen und Beurteilungen um die Möglichkeit zu erhalten zu einem gemeinsamen Bezugsrahmen („frame of reference“) zu gelangen. Weiterhin führte jeder Beobachter mindestens ein strukturiertes Interview im Rahmen des Trainings durch. Gegen Ende des Trainings erhielten die Teilnehmer die Möglichkeit einen vollständig bearbeiteten Postkorb auszuwerten und übrig gebliebene Fragen zu klären. Außerdem erhielten die Beobachter zum Abschluss ein kurzes Training im Führen von Feedbackgesprächen, welches ebenfalls ein Rollenspiel umfasste.

2.2.3.4 Beobachterkonferenz

Die Beobachter wurden dazu angehalten, sowohl im AC als auch im MMI möglichst viele relevante Verhaltensweisen auf ihren Beobachtungsbögen zu notieren. Die Bewertung der einzelnen verhaltensverankert definierten Dimensionen sowie eine Gesamteinschätzung der Leistung erfolgte dann direkt im Anschluss an jede Übung mit Hilfe einer 5-stufigen Skala (1 = erfüllt die Anforderungen vollständig, 3 = leichte Veränderungen wünschenswert, 5 = starke Veränderungen wünschenswert); hierbei fand jedoch kein Austausch zwischen den Beobachtern statt. Nach Durchführung der drei AC-Übungen kamen die Beobachter zur ersten Beobachterkonferenz zusammen. Die Leistungsbeurteilungen der Teilnehmer wurden nacheinander diskutiert, wobei übungsweise und nicht dimensionsweise vorgegangen wurde. Eine Einigung der Beobachter musste nicht erfolgen. Am Nachmittag wurden die Beobachter dann neuen Teilnehmern zugeordnet, die sie außer in einer kurzen Begrüßungsrunde bisher noch nicht gesehen hatten. Nach Durchführung des MMI kamen die Beobachter in vier Gruppen zu einer zweiten Konferenz zusammen. Die endgültige Leistungsbewertung der Teilnehmer erfolgte durch Mittelwertbildung der entsprechenden Beobachterurteile.

2.2.4 Statistische Analysen

2.2.4.1 Bestimmung der interne Konstruktvalidität

Die Auswertung der internen Konstruktvalidität erfolgt sowohl mit Hilfe von MTMM-Matrizen als auch mit Hilfe der konfirmatorischen Faktorenanalyse. Für letztere existieren verschiedene Modellvorstellungen, denen unterschiedliche Annahmen zugrunde liegen. Am populärsten sind die Ansätze mit korrelierenden Dimensionen und Methoden („correlated traits correlated methods“; CTCM-Ansatz bzw. traditionelle konfirmatorische Faktorenanalyse; vgl. Jöreskog, 1971 nach Tomás et

al., 2000) und mit korrelierenden Dimensionen und korrelierten Fehlern („correlated trait correlated uniqueness"; CTCU oder CU-Ansatz; vgl. Marsh, 1989). Beide Ansätze sind jedoch mit einer Reihe von Stärken und Schwächen verbunden. Nach Durchsicht der relevanten Literatur haben wir uns daher zu folgendem Vorgehen entschlossen:

1. Es sollen verschiedene Methoden zur Auswertung der MTMM-Matrix eingesetzt werden. So weist beispielsweise Conway (1996) darauf hin, dass unterschiedliche Methoden zu uneinheitlichen Ergebnissen kommen. Er empfiehlt, verschiedene Methoden bei der Auswertung einzusetzen, da einheitliche Ergebnisse, die durch verschiedene Auswertungstechniken gewonnen werden, die Glaubwürdigkeit und Aussagekraft der Resultate erhöhen (als Beispiel siehe Arthur et al., 2000; Silverman et al., 1986).

2. Auswertung nach den Kriterien von Campbell und Fiske (1959). Hiernach ist konvergente Validität gegeben, wenn die monotrait-heteromethod-Korrelationen signifikant größer als Null sind und „sufficiently large to encourage further examination" (Campbell und Fiske, 1959, S. 82). Um von diskriminanter Validität sprechen zu können sollen die folgenden drei Kriterien erfüllt sein: 1. Die monotrait-heteromethod-Korrelationen sind höher als die heterotrait-heteromethod-Korrelationen. 2. Die monotrait-heteromethod-Korrelationen sind größer als die heterotrait-monomethod-Korrelationen. 3. Die heterotrait-monomethod-Korrelationen weisen das gleiche Muster auf, wie die heterotrait-heteromethod-Korrelationen. Trotz verschiedener Kritikpunkte (z.B. wann ist ein Korrelationsmuster als gleich zu bezeichnen?) erlauben diese Kriterien immer noch, einen schnellen und intuitiv verständlichen Überblick über konvergente und diskriminante Validität zu erlangen. Außerdem wird hierdurch die Vergleichbarkeit der Ergebnisse mit anderen Studien erleichtert, da in den meisten Studien mittlere monotrait-heteromethod-Korrelationen (konvergente Validität) und mittlere heterotrait-monomethod-Korrelationen (diskriminante Validität) angegeben werden (vgl. Kolk, Born & Van der Flier, 2001).

3. Auswertung mit dem CTCM-Ansatz (traditionelle konfirmatorische Faktorenanalyse). Es wird das Vorgehen von Widaman (1985), modifiziert nach Byrne (1994), gewählt. Hierbei werden verschiedene hierarchisch abhängige Modelle berechnet, die anschließend mit Hilfe von χ^2-Differenzen-Tests paarweise auf Unterschiede im Modell-Fit getestet werden können. Dieses Vorgehen erlaubt die Bestimmung von konvergenter und diskriminanter Validität. Um Konvergenzprobleme

bei der Auswertung zu vermeiden, werden nur Personen mit vollständigen Daten zur Berechnung der Kovarianzmatrix herangezogen.

Bei Konvergenzproblemen empfiehlt Conway (1996) neue Startwerte zu probieren. Diese Startwerte können entweder zufällig gewählt werden oder mit Hilfe der „unweighted least Squares“ (ULS)-Schätzung oder der „generalized least squares“ (GLS)-Schätzung bestimmt werden. Außerdem ist es nach Marsh (1989) möglich, bei nicht konvergierenden Modellen verschiedene Parametrisierungen zu prüfen. Als Standardparametrisierung wird der Ansatz der fixierten Faktorvarianz („fixed factor variance“) gewählt, bei welcher die Faktorenvarianz auf 1 fixiert wird. Bei Konvergenzproblemen können dann fixierte Faktorladungen („fixed factor loadings“) oder die Rindskopf-Parametrisierung zu einer Lösung führen (vgl. Marsh, 1989). Weitere Möglichkeiten, wie beispielsweise die Begrenzung der Parameterschätzbereiche („range restriction“) oder die Elimination einzelner Methodenfaktoren (Maruyama, 1998), sollen nicht durchgeführt werden, da die entstehenden Lösungen als schwer interpretierbar und instabil angesehen werden.

4. Auswertung mit dem CU-Ansatz (correlated uniqueness). Hierbei wird das Vorgehen von Marsh (1989), modifiziert nach Sagie und Magnezy (1997), gewählt. Analog zum Ansatz von Widaman (1985) werden verschiedene hierarchisch voneinander abhängige Modelle berechnet, wobei konvergente und diskriminante Validität wiederum mit Hilfe von χ^2-Differenzen-Tests abgeschätzt werden. Der große Vorteil des CU-Ansatzes sind die eher selten auftretenden Konvergenzprobleme (Becker & Cote, 1994; Conway, 1996; Kleinmann & Köller, 1997). Aufgrund von theoretischen Schwächen und Interpretationsproblemen (vgl. Lance, Noble & Scullen, 2002) empfehlen verschiedene Autoren (z.B. Becker & Cote, 1994; Conway, 1996), immer auch die Modelle der traditionellen konfirmatorischen Faktorenanalyse zu rechnen.

2.2.4.2 Einfluss der gemeinsamen Beobachtervarianz

Kolk (2001) hat in zwei Studien den Einfluss der gemeinsamen Beobachtervarianz auf die Konstruktvalidität des AC untersucht und bestätigt. In der vorliegenden Studie wird ein analoges Vorgehen gewählt. Hierzu wird jeder Teilnehmer zufällig zwei Beobachtern zugeordnet, die mit ihm das komplette MMI durchführten. Üblicherweise werden nun die einzelnen Elemente der MTMM-Matrix dadurch bestimmt, dass man die Mittelwerte der Ratings von verschiedenen Beobachtern in einer Übung bezüglich der gleichen Dimension berechnet und anschließend miteinander korreliert. Auf diese

Weise werden jedoch die heterotrait-monomethod-Korrelationen durch gemeinsame Beobachtervarianz künstlich erhöht. Wenn – wie im MMI üblich – die gleichen Beobachter auch noch in verschiedenen Übungen eingesetzt werden, können auch monotrait-heteromethod- und heterotrait-heteromethod-Korrelationen künstlich erhöht werden. Um den Einfluss der gemeinsamen Beobachtervarianz im MMI abschätzen zu können, wurde deshalb eine MTMM-Matrix berechnet, in der die Dimensionsratings jedes einzelnen Beobachters miteinander korreliert wurden. Jedes Element einer konventionellen MTMM-Matrix wird daher in vier Korrelationen aufgeteilt. Jeweils zwei Korrelationen basieren auf Ratings des gleichen Beobachters (gemeinsame Beobachtervarianz), während die anderen beiden Korrelationen auf den Ratings von zwei verschiedenen Beobachtern beruhen (d.h. ohne gemeinsame Beobachtervarianz). Nach Fischer-Z-Transformation lassen sich nun die üblichen gemittelten monotrait-heteromethod-, heterotrait-monomethod- und heterotrait-heteromethod-Korrelationen berechnen, und zwar getrennt für die Bedingung gemeinsame Beobachtervarianz bzw. nicht gemeinsame Varianz.

2.3 Ergebnisse

Im Folgenden werden zunächst die Ergebnisse zur Reliabilität und zur internen Konstruktvalidität des MMI wiedergegeben, bevor dann die Korrelationen mit diversen externen Messinstrumenten berichtet werden. Abschließend wird der Einfluss der gemeinsamen Beobachtervarianz auf die interne Konstruktvalidität untersucht.

2.3.1 Reliabilität

In der vorliegenden Studie wurde die Reliabilität als Beurteilerübereinstimmung (Objektivität) und innere Konsistenz (Cronbach´s Alpha) bestimmt. Die Ergebnisse für die einzelnen Komponenten bzw. Dimensionen des MMI finden sich in Tabelle 2.

Tabelle 2: Beurteilerübereinstimmung und innere Konsistenz (α) des Multimodalen Interviews (MMI) und seiner Komponenten bzw. einzelnen Dimensionen

	Anzahl Items	Interrater-Reliabilität	α	α [a]
Selbstpräsentation (SP)				
Gesamtscore	3	.82	.74	.92
Systematisches Denken & Handeln	1	.66	-	-
Zusammenarbeit	1	.72	-	-
Steuerung sozialer Prozesse	1	.79	-	-
Biographisches Interview (BI)				
Gesamtscore	12	.83	.68	.68
Systematisches Denken & Handeln	4	.79	.50	.75
Zusammenarbeit	4	.68	.29	.56
Steuerung sozialer Prozesse	4	.88	.59	.81
Situatives Interview (SI)[b]				
Gesamtscore	12	.84	.68	.68
Systematisches Denken & Handeln	4	.81	.54	.78
Zusammenarbeit	4	.81	.40	.67
Steuerung sozialer Prozesse	4	.79	.46	.72
Multimodales Interview (MMI)[b]				
Gesamtscore	27	.88	.80	.64
Systematisches Denken & Handeln	9	.81	.61	.68
Zusammenarbeit	9	.79	.55	.62
Steuerung sozialer Prozesse	9	.86	.66	.72

Anmerkungen: *N* = 110.

[a] Zur besseren Vergleichbarkeit wurde Cronbach´s Alpha mit Hilfe der Spearman-Brown-Formel auf die gemeinsame Itemzahl von 12 Items umgerechnet. [b] *N* = 108.

Die *Beurteilerübereinstimmungen* für die Gesamtscores im BI, SI und MMI liegen mit .83, .84 und .88 im mittleren Bereich der üblicherweise für strukturierte Interviews angegeben wird. Die Selbstpräsentation erreicht hingegen mit einer Beurteiler-übereinstimmung von .82 einen sehr hohen Wert. Bezüglich der einzelnen Dimensionen im Gesamtinterview fallen unsere Beurteilerübereinstimmungen mit .79, .81 und .86 ebenfalls eher hoch aus.

Zum Vergleich: Conway et al. (1995) berichten in ihrer Metaanalyse eine mittlere Interrater-Reliabilität von .65 bis .92 für hochstrukturierte Interviews und in Latham (1989) finden sich acht Studien zum SI, deren Beurteilerübereinstimmungen im Bereich von .76 bis .96. liegen. Die Selbstpräsentation wird bei Schuler (1992) bzw. Schuler (1989) mit Werten im Bereich von .40 bis .45 angegeben und Schuler und

Moser (1995) berichten für die drei in einem MMI erhobenen Dimensionen Innovation, Kooperation mit dem Vorgesetzten und Führung Interrater-Reliabilitäten von .70 bis .76.

Auch die Werte für die *inneren Konsistenzen* der einzelnen Komponenten liegen mit .68, .68 und .74 im Bereich der üblicherweise erreicht wird. Die innere Konsistenz des Gesamtinterviews beträgt .80 und fällt damit eher niedrig aus. Auch die Werte für die einzelnen Dimensionen des Gesamtinterviews fallen mit .55, .61 und .66 eher niedrig aus.

Zum Vergleich: Schuler und Moser (1995) berichten von drei Studien, in denen die inneren Konsistenzen für die einzelnen Komponenten Werte im Bereich .58 bis .72 (Studie 2), .62 bis .83 (Studie 3) beziehungsweise .70 bis .77 (Studie 4) annehmen. Für das Gesamtinterview werden in diesen Studien Werte von .82, .83 und .87 angegeben. Außerdem berichten Schuler und Moser (1995) in Studie 3 innere Konsistenzen für die einzelnen Dimensionen des Gesamtinterviews. Diese liegen im Bereich von .70 bis .77 (8 bis 11 Items pro Dimension).

In der Literatur werden verschiedene potentielle Einflussfaktoren auf die innere Konsistenz des Einstellungsinterviews genannt. So konnten Conway et al. (1995) in ihrer Metaanalyse zeigen, dass die innere Konsistenz strukturierter Interviews niedriger ausfällt, als bei unstrukturierten Interviews. Dieser Effekt wird von den Autoren durch Halo-Fehler erklärt, welche die Beurteilungen im unstrukturierten Interview stark beeinflussen. Weiterhin konnte gezeigt werden, dass Anzahl der Interviewfragen und Höhe der inneren Konsistenz positiv zusammenhängen. Latham (1989) stellt außerdem die empirisch noch nicht bestätigte Vermutung auf, dass Interviews, welche lediglich eine Dimension erfassen, eine höhere innere Konsistenz erreichen, als Interviews, welche eine Vielzahl verschiedener Anforderungen messen. Entsprechend sollte die innere Konsistenz des Gesamtinterviews (bzw. der einzelnen Interviewkomponenten BI und SI) niedriger ausfallen, als die innere Konsistenz der zugehörigen Interviewdimensionen. Da die Höhe der inneren Konsistenz durch die Anzahl der Items beeinflusst wird, wurden zur Überprüfung dieser Annahme alle empirisch gewonnenen inneren Konsistenzen mit Hilfe der Spearman-Brown-Formel einheitlich auf 12 Fragen umgerechnet. Es zeigte sich, dass immerhin sechs der neun Dimensionen in BI, SI und MMI höhere innere Konsistenzen erreichen als die zugehörigen Gesamtwerte. Somit kann die Annahme von Latham (1989) tendenziell durch unsere Daten bestätigt werden.

2.3.2 Interne Konstruktvalidität

Tabelle 3 zeigt die Mittelwerte, Standardabweichungen und Interkorrelationen der Dimensionen in den drei Interviewteilen des MMI. Es ergeben sich mittlere monotrait-heteromethod-Korrelationen von .24 (konvergente Validität) und mittlere heterotrait-heteromethod-Korrelationen von .22 bzw. mittlere heterotrait-monomethod-Korrelationen von .41 (diskriminante Validität). Wenn man die vier Leitlinien von Campbell und Fiske (1959) auf die vorliegende MTMM-Matrix anwendet, kommt man zu dem Ergebnis, dass das MMI keine diskriminante Validität besitzt und die konvergente Validität zwar formal gegeben ist, aber eher gering ausfällt (nur 6 von 9 Korrelationen signifikant).

Tabelle 3: Multitrait-Multimethod-Matrix

	M	*SD*	1	2	3	4	5	6	7	8	9
SP											
1. SDH	2.53	.81	-								
2. ZU	2.51	.98	.46**	-							
3. SSP	3.26	1.06	.52**	.50**	-						
BI											
4. SDH	2.64	.68	.03	.24*	.11	-					
5. ZU	2.73	.59	.11	.32**	.23*	.45**	-				
6. SSP	2.82	.85	.29**	.44**	.41**	.28**	.37**	-			
SI											
7. SDH	2.66	.69	.22*	.24*	.17	.31**	.33**	.38**	-		
8. ZU	2.55	.65	.04	.15	-.15	.22*	.32**	.09	.27**	-	
9. SSP	2.69	.57	.28**	.28**	.12	.10	.37**	.31**	.43**	.42**	-

Anmerkungen: *N* = 108. SP = Selbstpräsentation. BI = Biographisches Interview. SI = Situatives Interview. SDH = Systematisches Denken & Handeln. ZU = Zusammenarbeit. SSP = Steuerung sozialer Prozesse.

*p < .05. **p < .01 (jeweils 2-seitig).

Wie oben beschrieben wurden zur Auswertung der MTMM-Matrix anschließend konfirmatorische Faktorenanalysen gerechnet (vgl. Tabelle 4). Die Angemessenheit der verschiedenen Modelle wurde anhand zweier Gruppen von Kriterien überprüft. Erstens sollten angemessene Modelle konvergieren und zulässige Lösungen (z.B. positive Fehler-Varianz) ergeben. Zweitens wurden verschiedene Fit-Indices (*RMSEA*, *CFI*) berechnet. Die Beurteilung der Fit-Indices erfolgte nach den üblichen

Kriterien (vgl. Loehlin, 1998). Hiernach soll der *RMSEA* kleiner gleich einem Wert von .05 sein und der *CFI* Werte größer gleich .95 erreichen.

Tabelle 4: Goodness-of-Fit Statistiken aus LISREL-Analysen für verschiedene faktorenanalytische Modelle

Modellbezeichnung und Erläuterung (nach Marsh, 1989)		*df*	χ^2	*p*	*RMSEA*	*CFI*	Zulässig?
traditionelle CFA							
4D	3 korrelierende Dimensionen 3 korrelierende Methoden	12	20.39	.06	.08	.96	Nein[b, d]
2D	1 Dimension, 3 korrelierende Methoden	15	21.62	.12	.06	.97	Nein[a, b, c]
1D	3 korrelierende Methoden	24	57.67	.00	.12	.86	Ja
CU-Ansatz							
4E	3 korrelierende Dimensionen, 3 korrelierende Fehler	15	32.94	.01	.11	.91	Ja
2E	1 Dimension, 3 korrelierende Fehler	18	35.45	.01	.09	.92	Ja
1E	3 korrelierende Fehler	27	59.58	.00	.13	.72	Ja

Anmerkung: *N* = 108. CFA = konfirmatorische Faktorenanalyse. CU = correlated uniqueness. *RMSEA* = Root Mean Square Error of Approximation. *CFI* = Comparative Fit Index. Zulässig? = sind alle geschätzten Modellparameter innerhalb zulässiger Bereiche.

[a]Konvergiert nicht. [b]Fehler-Varianz ist negativ. [c]Theta-Delta ist nicht positiv definit. [d]Korrelation zwischen latenten Dimensions-Faktoren ist größer als 1.

Wie erwartet (vgl. Becker & Cote, 1994; Kleinmann & Köller, 1997) ergab sich bei dem klassischen correlated-trait-correlated-method-Ansatz (Modellreihe D) nur ein konvergierendes Modell mit zulässigen Lösungen (Modell 1D). Daher wurde diese Modellreihe in der folgenden Auswertung nicht weiter betrachtet. Die Schätzung der Modelle nach dem CU-Ansatz (Modelle 4E bis 1E) verlief hingegen problemlos. Die Betrachtung der Fit-Indices (*RMSEA*, *CFI*) zeigt, dass ein Modell mit nur einer Dimension (2E) knapp am besten passt. Doch auch dieser Fit ist eher unbefriedigend (*RMSEA* > .05 und *CFI* < .95). Damit weist auch diese Analyse auf eine geringe interne Konstruktvalidität des MMI hin.

Zur Abschätzung von konvergenter und diskriminanter Validität wurden weiterhin χ^2-Differenzen-Tests berechnet (vgl. Tabelle 5). Hierbei ergaben sich Hinweise auf konvergente Validität ($\Delta\chi^2 = 26.64$; $p < .01$), während die diskriminante Validität nicht gegeben ist. Auch diese Ergebnisse weisen darauf hin, dass ein Modell mit nur einer Dimension die angemessene Lösung zur Erklärung der MTMM-Daten darstellt.

Tabelle 5: Modellvergleiche (χ^2-Differenzen-Test) zur Abschätzung von konvergenter und diskriminanter Validität

Test auf	Modellvergleich	$\Delta\chi^2$	Δdf	$\Delta\chi^2_{krit.05}$	$\Delta\chi^2_{krit.01}$
diskriminante Validität	4E vs. 2E	2.51	3	7.81	11.34
konvergente Validität	4E vs. 1E	26.64	12	21.02	26.22

Anmerkungen: $N = 108$. $\Delta\chi^2$ = Differenz in χ^2. Δdf = Differenz der Freiheitsgrade. $\Delta\chi^2_{krit.05}$ = kritische χ^2-Differenz auf 5% Niveau bei entsprechenden Freiheitsgraden. 4E = Modell mit 3 korrelierenden Dimensionen und 3 korrelierenden Fehlern. 2E = Modell mit einer Dimension und 3 korrelierenden Fehlern. 1E = Modell mit 3 korrelierenden Fehlern.

Nach der Analyse der Gesamtmatrix wurden die einzelnen Parameter von Modell 4E zur Abschätzung von konvergenter und diskriminanter Validität untersucht. In Tabelle 6 sind die durch die Dimensionsfaktoren erklärten Varianzanteile und in Tabelle 7 die Korrelationen zwischen den Dimensionsfaktoren dargestellt.

Es zeigt sich, dass die Dimensionsfaktoren knapp 30% der Varianz erklären. Dieser Wert ist als ausreichend einzuschätzen und weist auf eine gewisse konvergente Validität hin. Zum Vergleich: Kleinmann und Köller (1997) berichten für ein AC, welches ebenfalls mit Modell 4E ausgewertet wurde, eine erklärte Varianz von 44%; in Lievens und Keer (2001) wird ein Wert von 36% angegeben. Andere Autoren gelangten bei Auswertung mit Modell 4D auf durch die Dimensionen erklärte Varianzanteile von 4% (Bycio et al., 1987) bzw. 6% (Kudisch, Ladd & Dobbins, 1997). Die vorhandene konvergente Validität zeigt sich auch darin, dass die Ladungen sämtlicher Dimensionsfaktoren signifikant größer Null sind.

Tabelle 6: Durch die Dimensionsfaktoren erklärte Varianzanteile (Modell 4E)

Systematisches Denken & Handeln (SDH)	
Selbstpräsentation (SP)	8.1%
Biographisches Interview (BI)	26.2%
Situatives Interview (SI)	32.1%
Zusammenarbeit (ZU)	
Selbstpräsentation (SP)	24.1%
Biographisches Interview (BI)	53.7%
Situatives Interview (SI)	6.3%
Steuerung sozialer Prozesse (SSP)	
Selbstpräsentation (SP)	16.4%
Biographisches Interview (BI)	77.0%
Situatives Interview (SI)	17.6%
Durchschnitt	29.1%

Anmerkung: 4E = Modell mit 3 korrelierenden Dimensionen und 3 korrelierenden Fehlern.

Bei der Betrachtung der Korrelationen zwischen den Dimensionsfaktoren ergaben sich sehr hohe Werte im Bereich von .81 bis .92. Damit sind die erfassten Dimensionen nur schwerlich als unabhängig zu bezeichnen. Auch dieses Ergebnis spricht für die eindimensionale Lösung (Modell 2E) und damit für eine eher geringe diskriminante Validität.

Tabelle 7: Korrelationen zwischen den Dimensionsfaktoren (Modell 4E)

	1. SDH	2. ZU	3. SSP
1. Systematisches Denken & Handeln (SDH)	-		
2. Zusammenarbeit (ZU)	.82**	-	
3. Steuerung sozialer Prozesse (SSP)	.81**	.92**	-

Anmerkung: 4E = Modell mit 3 korrelierenden Dimensionen und 3 korrelierenden Fehlern.

**p < .01 (2-seitig).

Zur Abschätzung des Methodeneinflusses wurden im Modell 4E die mittleren Korrelationen zwischen den Fehlervarianzen berechnet. Für die Selbstpräsentation ergab sich eine mittlere Korrelation der Fehlervarianzen von .36, während das BI und das SI Werte von -.06 bzw. .24 erreichten. Der Einfluss der verschiedenen Interviewmethoden (BI und SI) auf die Bewertungen kann damit als deutlich geringer eingeschätzt werden, als der Einfluss der Methode Selbstpräsentation.

Diese Ergebnisse stehen im Einklang mit Lievens und Keer (2001), welche ebenfalls die Konstruktvalidität eines AC mit Hilfe von Modell 4E analysiert haben. Dort zeigte

sich für die Methoden Präsentation, Rollenspiel und Postkorb ein hoher Methodeneinfluss (mittlere Korrelationen der Fehlervarianzen im Bereich .49 bis .70), während die Fehlervarianzen des BI im Mittel lediglich zu .18 korrelierten.

2.3.3 Korrelate des Multimodalen Interviews (MMI)

Die Zusammenhänge zwischen dem MMI und den verschiedenen Messinstrumenten sind in Tabelle 8 wiedergegeben. Wie auch schon bei Schuler (1992) korrelieren die Gesamtwerte von MMI und AC mit r = .45 ($p < .001$) recht hoch miteinander. Dabei zeigte sich auch in unseren Daten, dass das MMI höher mit den Gruppendiskussionen zusammenhing, als mit der Postkorbaufgabe. Die Korrelationen zwischen den einzelnen Komponenten des MMI (SP, BI, SI) und dem Gesamtwert bzw. den einzelnen Dimensionen im AC sind ebenfalls durchgehend signifikant. Betrachtet man nun die neun Korrelationen zwischen den einzelnen Dimensionen in AC und MMI – welche Hinweise auf die externe Konstruktvalidität der Verfahren liefern – so fallen diese zwischen identischen Dimensionen etwas höher aus, als zwischen verschiedenen Dimensionen. Dies gilt insbesondere für die Dimension Zusammenarbeit (ZU).

Schmidt und Hunter (1998) vermuten, dass zwischen Verhaltensinterviews und Integritätstests ebenfalls Zusammenhänge bestehen. In unserer Untersuchung zeigten sich in der Tat signifikante Korrelationen zwischen dem eingesetzten Integritätstest (FES) und dem Gesamtwert im MMI. Diese Zusammenhänge waren insbesondere auf die Dimensionen Systematisches Denken & Handeln und Zusammenarbeit zurückzuführen. Da sich sowohl Integritätstests, als auch Verhaltensinterviews in den Metaanalysen von Ones, Viswesvaran und Schmidt (1993) bzw. Schmidt und Hunter (1998) als Prädiktoren für die allgemeine berufliche Leistung herausgestellt haben, entspricht dieses Ergebnis den Erwartungen. Marcus, Funke und Schuler (1997) gelangen zu dem Fazit, dass „kontraproduktives Verhalten über die gesamte Lebensspanne mit großer Stabilität auftritt bzw. ausbleibt“ (S. 5) und auch Marcus (2000) betont, dass sich kontraproduktives Verhalten am besten aus vergangenem Verhalten vorhersagen lässt. Diese Aussagen stimmen mit unserem Ergebnis überein, dass das vergangenheitsorientierte Interview (BI) am höchsten mit dem FES korreliert. Eine genauere Analyse der Zusammenhänge zwischen den 9 Subskalen des FES und dem Gesamtwert im MMI zeigt, dass die gefundenen Korrelationen auf die Unterskalen „Vertrauen/Glaube an das Gute“ (*r* =

.29, $p < .001$) und „Gelassenheit“ ($r = .23$, $p < .01$) zurückzuführen sind. Während die erste Subskala den einstellungsorientierten Integritätstests zugeordnet wird, gehört die zweite Subskala zu den eigenschaftsorientierten Tests.

Tabelle 8: Korrelationen zwischen dem Multimodalem Interview (MMI) bzw. seinen Komponenten und verschiedenen externen Maßen

Verfahren	MMI				SP	BI	SI
	Gesamt	SDH	ZU	SSP	Gesamt	Gesamt	Gesamt
AC							
Gesamtwert	.45**	.41**	.35**	.37**	.31**	.36**	.38**
SDH	.36**	.37**	.22*	.32**	.29**	.25**	.29**
ZU	.43**	.36**	.39**	.33**	.25**	.40**	.37**
SSP	.42**	.39**	.34**	.34**	.29**	.33**	.38**
GD1	.46**	.41**	.35**	.40**	.37**	.32**	.36**
PK	.19*	.14†	.27**	.08	.01	.24**	.25**
GD2	.35**	.36**	.21*	.32**	.27**	.28**	.27**
FES							
Gesamtskalenwert	.23**	.24**	.26**	.11	.09	.31**	.18*
Einstellungen	.21*	.22*	.22*	.10	.08	.26**	.18*
Eigenschaft	.17*	.16*	.21*	.08	.07	.25**	.10
IST 2000							
verbal[a]	.17*	.19*	.03	.19*	.10	.08	.22*
numerisch[b]	.39**	.36**	.22†	.32*	.35**	.26*	.19†
figural[a]	.04	-.03	.01	.10	.07	-.04	.03
Self-Monitoring[c]							
Gesamtwert	.13†	.06	.16†	.11	.11	.22*	.04
Skala soz. Fertigkeiten	.17*	.12	.17†	.15†	.07	.25**	.08
Skala soz. Vergleichsprozesse	.07	.06	.13†	.01	.07	.08	.00
Skala Inkonsistenz	-.00	-.08	.030	.04	.07	.07	-.19*
soziale Erwünschtheit	.18*	.20*	.16†	.12	.12	.23**	.08
Textilfabrik							
Gesamtkapital	.20*	.25**	.11	.16†	.09	.16†	.25**

Anmerkungen: $N = 108$. MMI = Multimodales Interview. SP = Selbstpräsentation. BI = Biographisches Interview. SI = Situatives Interview. SDH = Systematisches Denken & Handeln. ZU = Zusammenarbeit. SSP = Steuerung sozialer Prozesse. GD1 = Gruppendiskussion ohne Rollenvorgabe. PK = Postkorb. GD2 = Gruppendiskussion mit Rollenvorgabe. AC = Assessment Center. FES = Fragebogen zu Einstellungen und Selbsteinschätzungen. IST 2000 = Intelligenz-Struktur-Test 2000.

[a] $N = 106$. [b] $N = 54$. [c] $N = 107$.

†$p < .10$, *$p < .05$, **$p < .01$ (jeweils 1-seitig).

Auch die Zusammenhänge zwischen MMI und dem Faktor verbale Intelligenz (IST 2000) entsprechen den Erwartungen. So berichten Huffcutt et al. (1996) in ihrer

Metaanalyse eine mittlere Korrelation von r = .23 zwischen strukturierten Interviews und sprachlicher Intelligenz, wobei BI-Fragen mit r = .12 deutlich unter den SI-Fragen mit r = .21 liegen. In der vorliegenden Untersuchung fällt der Zusammenhang zwischen dem Faktor verbale Intelligenz und dem BI mit r = .08 (*ns*) bzw. der Selbstpräsentation mit r = .10 (*ns*) ebenfalls gering aus, während sich mit dem SI eine signifikante Korrelation von r = .22 (p < .05) ergibt. Plausibel ist auch, dass Zusammenarbeit – als einzige Dimension des Gesamtinterviews – nicht signifikant mit verbaler Intelligenz korreliert (r = .03, *ns*). Die weiteren Ergebnisse sind hingegen eher schwierig in den bestehenden Forschungsstand einzuordnen. So ergeben sich für den Faktor figurale Intelligenz und dem MMI lediglich zufällige Zusammenhänge, während der Faktor numerische Intelligenz wiederum substantielle Korrelationen mit dem MMI aufweist. Letztere lassen sich vor allem auf die Dimensionen Systematisches Denken & Handeln sowie Steuerung sozialer Prozesse bzw. auf die Selbstpräsentation zurückführen. Bei der Interpretation der Ergebnisse sollte jedoch beachtet werden, dass der Faktor numerische Intelligenz lediglich bei 54 Personen erhoben wurde. Eine genauere Analyse der Zusammenhänge zwischen dem MMI und den neun Unterskalen des IST 2000 zeigt, dass insbesondere die Aufgabe „Zahlenreihen“ hoch mit den verschiedenen Komponenten des MMI korreliert.

Die Zusammenhänge zwischen Self-Monitoring und MMI sind eher gering und ohne eindeutiges Muster. Ähnlich uneinheitliche Ergebnisse werden auch zwischen Self-Monitoring und dem Abschneiden im AC berichtet. Während Diemand und Schuler (1991) substantielle Korrelationen zwischen der Selbstüberwachungstendenz von Teilnehmern und der Gesamtleistung im AC finden konnten, wurde dieses Ergebnis in einer späteren Untersuchung (Diemand & Schuler, 1998) nicht bestätigt. Letztere vermuten, dass in hochstandardisierten Verfahren – wie es auch das MMI eines darstellt – sogar Personen mit extrem hoher Tendenz zur Selbstüberwachung keine ausreichende Gelegenheit zur gezielten Selbstdarstellung bzw. Steuerung ihrer Außenwirkung haben. Daher sollte sich Self-Monitoring bei diesen Verfahren nicht auf die gezeigte Leistung auswirken. Ähnlich wie in bisherigen Untersuchungen (Diemand & Schuler, 1991; Moser, Diemand & Schuler, 1996) scheint am ehesten die Subskala Soziale Fertigkeiten mit der Leistung in einem interaktiven Personalauswahlverfahren zu korrelieren.

Auch zwischen sozialer Erwünschtheit und dem MMI zeigen sich nur geringe Zusammenhänge. Dies ist plausibel, da die Möglichkeiten zur gezielten

Selbstdarstellung in einem hochstrukturierten Verfahren begrenzt sind (s.o.), so dass auch Personen mit hoher sozialer Erwünschtheit nicht besser abschneiden. Am ehesten scheint sich soziale Erwünschtheit positiv auf das BI auszuwirken. Da die Bewerber ganz allgemein nach Situationen in der Vergangenheit gefragt werden, werden sich Personen mit hoher sozialer Erwünschtheit besonders vorteilhafte Begebenheiten aussuchen.

Ein letzter Hinweis zu den im MMI erfassten Konstrukten liefern die Korrelationen mit der Textilfabrik, einem Verfahren zur Erfassung der Problemlösefähigkeit. Insbesondere das SI bzw. die Dimension Systematisches Denken & Handeln – beide erfassen Aspekte der verbalen und numerischen Intelligenz – korrelieren hoch mit der Textilfabrik.

2.3.4 Einfluss der gemeinsamen Beobachtervarianz auf die interne Konstruktvalidität

Zur Bearbeitung der letzten Fragestellung wurde eine 18 x 18 MTMM-Matrix (3 Übungen, 3 Dimensionen, 2 Beobachter) getrennt für „Beobachter 1“ und „Beobachter 2“ berechnet (vgl. Kolk, Born & Van der Flier, in press-a). Die entsprechenden Ergebnisse sind in Tabelle 9 dargestellt.

Tabelle 9: Multitrait-Multimethod-Korrelationsmatrix mit getrennten Ratings für Beobachter 1 und 2

			SP						BI						SI					
			SDH		ZU		SSP		SDH		ZU		SSP		SDH		ZU		SSP	
			Bb1	Bb2	Bb1	Bb2	Bb1	Bb2	Bb1	Bb2	Bb1	Bb2	Bb1	Bb2	Bb1	Bb2	Bb1	Bb2	Bb1	Bb2
	SDH	Bb1																		
		Bb2																		
SP	ZU	Bb1	.43	.41																
		Bb2	.31	.40																
	SSP	Bb1	.44	.43	.47	.36														
		Bb2	.42	.49	.50	.43														
	SDH	Bb1	.01	.09	.24	.18	.20	.10												
		Bb2	-.09	.04	.22	.19	.08	.01												
BI	ZU	Bb1	.15	.18	.27	.26	.26	.23	.48	.35										
		Bb2	-.02	.06	.27	.28	.19	.15	.37	.35										
	SSP	Bb1	.28	.20	.44	.34	.40	.35	.31	.27	.41	.24								
		Bb2	.27	.27	.44	.37	.37	.37	.23	.22	.37	.31								
	SDH	Bb1	.15	.22	.28	.22	.17	.13	.36	.31	.34	.26	.39	.42						
		Bb2	.19	.21	.19	.15	.15	.16	.27	.17	.35	.21	.29	.32						
SI	ZU	Bb1	.01	.09	.13	.14	-.17	-.18	.20	.18	.29	.22	.12	.11	.28	.25				
		Bb2	-.01	.05	.13	.12	-.05	-.14	.21	.20	.26	.33	.05	.04	.19	.26				
	SSP	Bb1	.25	.27	.26	.25	.06	.10	.06	.03	.30	.22	.28	.32	.45	.38	.47	.33		
		Bb2	.15	.25	.23	.21	.11	.13	.15	.10	.43	.32	.25	.29	.37	.36	.40	.31		

Anmerkungen: *N* = 108. SP = Selbstpräsentation. BI = Biographisches Interview. SI = Situatives Interview. SDH = Systematisches Denken & Handeln. ZU = Zusammenarbeit. SSP = Steuerung sozialer Prozesse. Bb1 = „Beobachter 1“. Bb2 = „Beobachter 2“. Hellgrau = heterotrait-monomethod-Korrelationen. Dunkelgrau = monotrait-heteromethod-Korrelationen

Um den Einfluss der gemeinsamen Beobachtervarianz auf die Konstruktvalidität abzuschätzen, wurden die Korrelationen aus Tabelle 9 Fischer-Z transformiert (siehe Bortz, 1999) und die mittleren monotrait-heteromethod-, heterotrait-monomethod- und heterotrait-heteromethod-Korrelationen für gleiche bzw. unterschiedliche Beobachter berechnet (Tabelle 10).

Tabelle 10: Mittlere konvergente und diskriminante Validität mit und ohne gemeinsame Beobachtervarianz

mittlere Korrelationen	mit gemeinsamer Beobachtervarianz	ohne gemeinsame Beobachtervarianz
monotrait-heteromethod (konvergente Validität)	.22	.21
heterotrait-monomethod (diskriminante Validität)	.39	.35
heterotrait-heteromethod	.19	.19

Wie erwartet sind die Korrelationen – welche gemeinsame Beobachtervarianz enthalten – deskriptiv etwas höher, als die Korrelationen ohne gemeinsame Beobachtervarianz. Wie auch schon bei Kolk et al. (in press-a) ergibt sich der stärkste Effekt für die diskriminante Validität. Insgesamt gesehen sind die Auswirkungen der gemeinsamen Beobachtervarianz jedoch sehr gering. Zur statistischen Absicherung der Effekte wurden die Korrelationen wie bei Silverman et al. (1986) als Datenpunkte betrachtet und nach Fishers Z-Transformation t-Tests (1-seitig) für unabhängige Stichproben gerechnet. Hierbei zeigte sich, dass die Korrelationen unter der Bedingung gemeinsame Beobachtervarianz nicht signifikant größer sind, als die Korrelationen ohne gemeinsame Beobachtervarianz. Lediglich für die heterotrait-monomethod-Korrelationen ergab sich ein tendenzieller Effekt (p = .082).

2.4 Diskussion

Die Untersuchungen zur Reliabilität des MMI (Beurteilerübereinstimmung und innere Konsistenz) fielen weitgehend zufriedenstellend aus. So ergaben sich auch bei uns die üblicherweise berichteten hohen Beurteilerübereinstimmungen. Diese sind nach Hunter und Hunter (1984) primär auf die Verwendung von verhaltensverankerten Bewertungsskalen zurückzuführen. Hingegen fiel die innere Konsistenz eher niedrig aus. Aufgrund der Metaanalyse von Conway et al. (1995) konnte dieses Ergebnis für unser hochstrukturiertes Interview erwartet werden. Wir schätzen die etwas geringere Homogenität unserer Interviewfragen jedoch eher positiv ein, da mit den verschiedenen Fragen auch drei unterschiedliche Beurteilungsdimensionen erfasst werden sollen. So waren dann auch – wie von Latham (1989) vermutet – die mit Hilfe der Spearman-Brown-Formel auf vergleichbare Länge gerechneten inneren Konsistenzen für die Dimensionen größtenteils höher, als für das Gesamtinterview. Damit konnten in der vorliegenden Studie die einschlägigen Befunde zur Reliabilität des strukturierten Interviews repliziert werden, womit die m.E. wichtigsten Voraussetzungen zur Überprüfung der nachfolgenden Fragestellungen gegeben sind.

2.4.1 Die interne Konstruktvalidität des Multimodalen Interviews (MMI)

Ziel der ersten Fragestellung war die Überprüfung der internen Konstruktvalidität des MMI. Hierbei war zu beachten, dass unterschiedliche Methoden teilweise zu

unterschiedlichen Ergebnissen gelangen (vgl. Conway, 1996). Um die verlässliche Interpretation der Resultate zu ermöglichen, wurden daher unterschiedliche Auswertungsstrategien eingesetzt.

Zunächst wurden nach Fischer-Z-Transformation die mittleren konvergenten und diskriminanten Korrelationen berechnet. Diese lagen im Wertebereich aktueller Studien zur internen Konstruktvalidität des AC (Donahue, Truxillo, Cornwell & Gerrity, 1997; Kudisch, Ladd & Dobbins, 1997). Dort ergaben sich für die monotrait-heteromethod-Korrelationen (konvergente Validität) Werte von .2 bis .3, für die heterotrait-monomethod-Korrelationen (diskriminante Validität) Werte von .4 bis .5 und für die heterotrait-heteromethod-Korrelationen Werte im Bereich von .1 bis .2. Weitere Ergebnisse zur konvergenten und diskriminanten Validität des AC finden sich in der Metaanalyse von Kolk et al. (2001).

Weiterhin analysierten wir die interne Konstruktvalidität des MMI mit Hilfe der konfirmatorischen Faktorenanalyse. Hierbei wurden nach dem Vorschlag von Sagie und Magnezy (1997) bzw. Lievens und Keer (2001) sowohl traditionelle Modelle als auch der „correlated uniqueness“ (CU)-Ansatz berechnet. Aufgrund von Konvergenzproblemen beim traditionellen Ansatz wurden in der weiteren Auswertung nur die Ergebnisse des CU-Ansatzes beachtet. Dabei zeigte das Modell mit nur einer Dimension den besten Gesamt-Fit. Modellvergleiche mit Hilfe von χ^2-Differenzen-Tests ergaben Hinweise auf konvergente, aber nicht auf diskriminante Validität. Weiterhin zeigte sich, dass ein zufriedenstellend großer Anteil der Varianz in den Leistungsbewertungen durch die Dimensionsfaktoren erklärt wurde, was ebenfalls auf konvergente Validität hinweist. Hingegen waren die Korrelationen zwischen den Dimensionsfaktoren sehr hoch, was gegen diskriminante Validität spricht.

Zusammenfassend lässt sich festhalten: Die verschiedenen Auswertungen zur Analyse der internen Konstruktvalidität des strukturierten Interviews ergeben ein einheitliches Bild. Anders als bei Schuler (1989) bzw. Schuler und Funke (1989), welche sowohl bezüglich konvergenter als auch diskriminanter Validität zu einem vernichtenden Fazit gelangen, finden wir in unserer Untersuchung zumindest einige Anhaltspunkte auf konvergente Validität. Wir führen die leichte Verbesserung der konvergenten Validität auf die sorgfältige Konstruktion des Interviews und die umfassende Beachtung potentieller Einflussfaktoren zurück (z.B. Anzahl der Dimensionen, Beobachtbarkeit und Unabhängigkeit der Dimensionen, Beobachtertraining). Für Praktiker in Wirtschaftsunternehmen könnten diese

Ergebnisse daher Ansporn sein, bei der Entwicklung eigener Interviews ebenfalls entsprechende Faktoren zu beachten. Dies gilt insbesondere dann, wenn die Interviews im Rahmen von Personalentwicklungsmaßnahmen durchgeführt werden und somit darauf abzielen, unterschiedliche Dimensionen zu erfassen. Insgesamt weisen die Ergebnisse der vorgenommenen Auswertungen jedoch immer noch auf eine eher geringe interne Konstruktvalidität des Interviews hin.

2.4.2 Korrelate des Multimodalen Interviews (MMI)

In einer explorativen Analyse wurden im Rahmen der zweiten Fragestellung die Korrelationen zwischen einzelnen Interviewkomponenten bzw. Dimensionen und verschiedenen externen Maßen untersucht. Hierbei ergaben sich plausible Zusammenhänge, die teilweise auch schon in anderen Untersuchungen (z.B. Schuler & Moser, 1995) berichtet wurden. So korrelierte das Gesamtinterview relativ hoch mit den verschiedenen AC-Übungen und den einzelnen AC-Dimensionen, während die Zusammenhänge mit dem Integritätstest, sozialer Erwünschtheit und der Textilfabrik nur eine mittlere Höhe erreichten.

Bei den einzelnen Interviewkomponenten korrelierte das BI besonders hoch mit dem Integritätstest, numerischer Intelligenz, Self-Monitoring und sozialer Erwünschtheit. Hingegen ergaben sich für das SI hohe Zusammenhänge mit der Textilfabrik und verbaler Intelligenz. Beide Verfahren korrelierten ähnlich hoch mit dem AC, während die Selbstpräsentation dort etwas niedrigere Werte erreichte und weiterhin lediglich mit numerischer Intelligenz signifikante Zusammenhänge aufwies.

Auch für die einzelnen Dimensionen des MMI zeigten sich solche differentiellen Zusammenhangsmuster. Während die Dimension Zusammenarbeit vor allem mit dem Integritätstest korrelierte und tendenziell mit Self-Monitoring und sozialer Erwünschtheit zusammenhing, ergaben sich für die Dimension Steuerung sozialer Prozesse signifikante Korrelationen mit verbaler und numerischer Intelligenz. Die Dimension Systematisches Denken & Handeln wiederum korrelierte mit dem Integritätstest und sozialer Erwünschtheit, vor allem aber auch mit verbaler und numerischer Intelligenz sowie der Problemlöseaufgabe (Textilfabrik). Ein letzter Hinweis darauf, dass mit den Interviewdimensionen die beabsichtigten Konstrukte erfasst wurden, lieferten die Korrelationen mit den AC-Dimensionen, welche zwischen identischen Dimensionen deskriptiv etwas höher ausfielen, als zwischen unterschiedlichen Dimensionen.

2.4.3 Einfluss der gemeinsamen Beobachtervarianz

In der letzten Fragestellung wurde der Einfluss der gemeinsamen Beobachtervarianz auf die interne Konstruktvalidität untersucht. Analog zu Kolk et al. (in press, Studie 2) berechneten wir zunächst eine MTMM-Matrix, in welcher die Dimensionsratings der *einzelnen* Beobachter korreliert wurden – und nicht wie üblich die Mittelwerte der Beobachterurteile. Nach Fischer-Z-Transformation konnten nun die mittleren konvergenten und diskriminanten Korrelationen mit und ohne gemeinsame Beobachtervarianz berechnet werden. Anders als in der AC-Forschung zeigten sich hierbei keine signifikanten Unterschiede zwischen den Korrelationen mit bzw. ohne gemeinsamer Beobachtervarianz.

Vorhandene Unterschiede (d.h. höhere Korrelationen zwischen den Bewertungen gleicher Beobachter) werden üblicherweise auf Beobachterfehler (Halo-Fehler) zurückgeführt. Die geringen Unterschiede dieser Untersuchung weisen somit darauf hin, dass der Halo-Effekt im strukturierten Interview als klein einzuschätzen ist. Dies ist plausibel, denn anders als im AC, wo die Beobachter gleichzeitig ein Urteil über verschiedene Dimensionen fällen sollen, wird im Multimodalen Interview jede Frage einzeln bewertet, und zwar mit Hilfe von verhaltensverankerten Beobachtungsskalen. Die Bewertungen der einzelnen Dimensionen erfolgt daher deutlich reliabler als im AC. Die gemeinsame Übungsvarianz in den Bewertungen als Erklärung für niedrige diskriminante Validität (hohe Korrelationen zwischen verschiedenen Dimensionen) konnte natürlich auch in dieser Studie nicht ausgeschaltet werden. Gegen eine substantielle Übungsvarianz sprechen jedoch die relativ niedrigen Korrelationen der Fehlervarianzen im BI und SI (vgl. Ergebnisse zur internen Konstruktvalidität), welche auf einen geringen Methodeneinfluss hinweisen. Da weder gemeinsame Beobachtervarianz noch gemeinsame Übungsvarianz eine plausible Erklärung für die geringe diskriminante Validität liefern, können wir annehmen, dass die untersuchten Dimensionen trotz der aufwendigen Entwicklung konzeptionell immer noch sehr nah zusammenhängen und somit die hohen Interkorrelationen der Dimensionen wahre Varianz widerspiegeln.

Insgesamt lässt sich jedenfalls festhalten, dass in der vorliegenden Studie die gemeinsame Beobachtervarianz einen vernachlässigbaren Einfluss auf die interne Konstruktvalidität des strukturierten Interviews besitzt. Es bleibt jedoch wünschenswert, dieses Ergebnis mit Hilfe einer experimentellen Manipulation zu replizieren (vgl. Kolk et al., in press-a, Studie 1).

2.4.4 Ausblick auf die Studien 2 und 3

Die vorliegende Studie unterstreicht die Analogie der Ergebnisse von AC-Forschung und Forschung zum strukturierten Interview. Obwohl sich beide Verfahren als kriteriumsvalide herausgestellt haben und Facetten der externen Konstruktvalidität geklärt sind (vgl. Scholz and Schuler 1993; Salgado & Moscoso, 2002), so ist doch die interne Konstruktvalidität als ungenügend zu bezeichnen. Nach Arthur et al. (2000) gibt es beim AC zwei Gruppen von Erklärungen für dieses Phänomen der mangelhaften Konstruktvalidität bei gleichzeitiger Kriteriumsvalidität:

1. Es werden andere Konstrukte als die intendierten beobachtet (z.B. soziale Erwünschtheit, Self-Monitoring). Diese Konstrukte erklären die gemeinsame Varianz zwischen Prädiktor und Kriterium.
2. Starke Messfehler verhindern den Nachweis von konvergenter und diskriminanter Validität. Methodische Verbesserung bei der Entwicklung der Verfahren (Unabhängigkeit und Beobachtbarkeit der Dimensionen etc.) minimieren diese Fehler und ermöglichen so den Nachweis der Konstruktvalidität.

Für das AC konnten solche methodischen Einflussfaktoren nachgewiesen werden (vgl. die Metaanalyse von Kolk et al., 2001 oder die Überblicksarbeit von Lievens, 1998). Für das strukturierte Interview (am Beispiel des MMI) könnten wir jedoch trotz Beachtung einer Vielzahl potentieller Einflussfaktoren keine interne Konstruktvalidität feststellen. Auch der Vergleich von Urteilen mit bzw. ohne gemeinsamer Beobachtervarianz erbrachte keine nennenswerte Unterschiede in der internen Konstruktvalidität. Unseres Erachtens ist es daher sinnvoll, bei der Untersuchung der internen Konstruktvalidität des strukturierten Interviews verstärkt personenbezogene Konstrukte und somit den ersten Erklärungsansatz zu beachten. In einer Untersuchung zum AC konnte nachgewiesen werden, dass die Fähigkeit, Anforderungsdimensionen zu erkennen, Einfluss auf Konstrukt- und Kriteriumsvalidität besitzt (Kleinmann, 1997b). In den beiden folgenden Studien soll nun dieser Ansatz auf das strukturierte Interview übertragen werden.

3 Die Fähigkeit, Anforderungsdimensionen zu erkennen: Einfluss auf konvergente Validität und Leistung im Multimodalen Interview

Zusammenfassung. Welche Folgen hat das Erkennen von Anforderungsdimensionen auf die Leistungsbeurteilung im strukturierten Interview? Kleinmann (1993) konnte zeigen, dass Personen im Assessment Center (AC) besser abschneiden, wenn sie die zugrunde liegenden Anforderungsdimensionen erkennen. Außerdem wies er nach, dass das Ausmaß des Erkennens die konvergente Validität beeinflusst. In der vorliegenden Studie (N = 95) wurden die wesentlichen Überlegungen und Ergebnisse dieser Untersuchung auf das strukturierte Interview übertragen und repliziert. Weiterhin wurden die Zusammenhänge zwischen Beurteilung im AC bzw. Interview und Erkennensleistung im AC bzw. Interview berechnet. Die Ergebnisse weisen darauf hin, dass es sich beim Erkennen der Anforderungsdimensionen um eine Fähigkeit bzw. Fertigkeit handelt, die relativ unabhängig vom zugrunde liegenden Instrument erfasst werden kann.

3.1 Einleitung

Eignungsdiagnostische Verfahren können dahingehend unterschieden werden, ob das gewünschte Zielverhalten eindeutig aus der Aufgabe hervorgeht oder nicht. Im ersten Fall, wie er beispielsweise bei Intelligenztests oder Konzentrationstests häufig zu finden ist, haben die zu testenden Personen eine einzige Aufgabe: Die optimale Bearbeitung, d.h. richtige Beantwortung der Items. Anders sieht es beispielsweise bei situativen AC-Übungen, verschiedenen Interviewtechniken und Persönlichkeits-fragebögen aus. Die Kriterien, nach denen diese Verfahren bewertet werden, sind nicht eindeutig formuliert bzw. den Teilnehmern kommuniziert; daher spricht man in diesem Fall auch von intransparenten Verfahren. So sind etwa in einem Verhandlungsrollenspiel sowohl Durchsetzungsfähigkeit als auch Kooperation zwei prinzipiell plausible Dimensionen, nach denen das Verhalten der Teilnehmer bewertet wird. Es ist anzunehmen, dass, vor allem in Bewerbungssituationen, die Teilnehmer eine möglichst positive Beurteilung erhalten möchten. Wenn sie gezielt

ihr Verhalten beeinflussen wollen und dabei mit intransparenten Verfahren getestet werden, haben sie mindestens zwei Aufgaben zu leisten. Einerseits müssen sie die relevanten Bewertungsdimensionen *erkennen*, andererseits sich in den Interviews und Übungen entsprechend der Kriterien *verhalten*. Konkret bedeutet dies beispielsweise, dass ein Bewerber zunächst einmal erkennen muss, dass in einem spezifischen Interview kooperatives Verhalten erfasst werden soll. Wenn ihm das gelingt, dann hat er die Möglichkeit während des Interviews gezielt Situationen zu benennen, in denen er kooperatives Verhalten gezeigt hat. Die Leistung des Teilnehmers hängt demnach nicht nur von seiner wahren Fähigkeit bzgl. einer bestimmten Dimension ab, sondern auch davon, welche Dimension er als relevant erkannt hat. Um die Kriterien zu erkennen, müssen die vorhandenen Hinweisreize („demand characteristics“, vgl. Orne, 1962) richtig interpretiert werden. Dabei kann davon ausgegangen werden, dass es den Bewerbern in unterschiedlichem Ausmaß gelingt, die relevanten Dimensionen zu entschlüsseln. Im Kontext der Personalauswahl kommen als Quellen dafür u.a. Vorerfahrungen mit dem Verfahren, Informationen über das Unternehmen und die Stellenanzeige, Übungsmaterial und Verhalten der Beobachter in Betracht (vgl. Kleinmann, 1993).

Einige der obigen Annahmen wurden ausführlich am Beispiel des Assessment Centers (AC) untersucht (z.B. Bungard, 1987; Kleinmann, 1991, 1993, 1997b). In diesen Studien konnte gezeigt werden, dass die Bewertungskriterien eines AC für die Teilnehmer größtenteils intransparent sind. Aufgrund von qualitativen Daten kommt Bungard (1987) zu dem Schluss, dass sich die Art der individuellen Hypothese in dem konkret gezeigten Verhalten der Teilnehmer widerspiegelt. Weiterhin konnte Kleinmann (1993, 1997b) zeigen, dass die Teilnehmer unterschiedlich gut im Erkennen der relevanten Dimensionen sind und dass das Ausmaß des Erkennens mit der Leistungsbewertung kovariiert. Diese Fähigkeit, Anforderungsdimensionen zu erkennen, wird im Folgenden „capability to discern dimensions“ (CDD) genannt (Richter & Kleinmann, 2002).

Für die vorliegende Arbeit, die sich insbesondere mit der Konstruktvalidität des strukturierten Interviews beschäftigt, ist ein weiteres Ergebnis interessant. So berichtet Kleinmann (1993), dass das Ausmaß des Erkennens bzw. die Transparenz der Bewertungsdimensionen Einfluss auf die konvergente Validität im AC hat. Dahinter steckt folgende Überlegung: Die bewertete Leistung hängt mit dem Ausmaß des Erkennens zusammen. Wenn ein Teilnehmer zwei identische Dimensionen in

zwei verschiedenen Übungen einmal erkennt und einmal nicht, so hat dies Leistungsunterschiede zur Folge, die sich dann wiederum auf die Korrelation der beiden Dimensionen und damit auf die konvergente Validität auswirken. Zur Erläuterung ein Beispiel: Ein Teilnehmer erkennt bei einer Übung, dass es um die Dimension Kooperation geht, bei einer anderen Übung jedoch nicht. Dieser Teilnehmer wird in den beiden Übungen ein unterschiedliches Verhalten zeigen (z.B. einmal kooperativ und einmal kompetitiv). Die Bewertungen in beiden Übungen bezüglich der Dimension Kooperation fallen daher unterschiedlich aus, was eine niedrige Korrelation bzw. konvergente Validität zur Folge hat. Wenn der Teilnehmer die Anforderungsdimension Kooperation jedoch in beiden Übungen konsistent erkennt, dann resultiert eine eher einheitliche Bewertung, was sich wiederum positiv auf die konvergente Validität auswirkt.

Hauptziel der vorliegende Studie ist es, die oben diskutierten Annahmen, die für das AC bestätigt werden konnten, nun auf ein weiteres intransparentes eignungsdiagnostischen Verfahren zu übertragen, das Multimodale Interview (MMI). Insbesondere sollen folgende Hypothesen überprüft werden (vgl. Kleinmann, 1993):

Hypothese 1: Über die bewerteten Anforderungsdimensionen in den einzelnen Interviewteilen entwickeln Personen interindividuell unterschiedliche Annahmen, die sich im Ausmaß, in dem sie den tatsächlichen Anforderungsdimensionen entsprechen, unterscheiden. Personen sind über die verschiedenen Interviewteile hinweg unterschiedlich fähig, die relevanten Anforderungen zu erkennen.

Hypothese 2: Personen werden im Einstellungsinterview um so besser bewertet, je genauer sie die Anforderungsdimensionen erkannt haben.

Hypothese 3: Personen schneiden auf den Dimensionen, die sie erkannt haben, intraindividuell besser ab als auf denen, die sie nicht erkannt haben.

Hypothese 4: Wenn identische Anforderungen in zwei Interviewteilen gleichermaßen erkannt werden, werden sie in den beiden Interviewteilen ähnlicher bewertet (höhere konvergente Validität), als wenn nur in einer der beiden Interviewteile die Anforderungsdimension erkannt wird.

Weiterhin gehen wir davon aus, das es sich bei CDD um eine personenbezogene Fähigkeit handelt, die relativ unabhängig vom jeweiligen eignungsdiagnostischen Instrument erfasst werden kann. Damit sollte CDD unabhängig vom zugrunde liegenden Instrument zur Vorhersage von Leistungsbeurteilungen eingesetzt werden können. Diese Annahme führt zu den drei folgenden Hypothesen:

Hypothese 5: Die in einem AC ermittelten CDD-Werte korrelieren mit den im MMI ermittelten CDD-Werten.

Hypothese 6: Die in einem AC ermittelten CDD-Werte korrelieren mit der Bewertung im MMI.

Hypothese 7: Die im MMI ermittelten CDD-Werte korrelieren mit der Bewertung im AC.

Sollten die obigen Hypothesen bestätigt werden können, werden die folgenden Überlegungen interessant. Beim MMI handelt es sich um ein Verfahren mit hoher prognostischer Validität (z.B. Huffcutt et al., 2002; Schuler & Moser, 1995; Taylor & Small, 2000), dessen Konstruktvalidität bei Auswertung mit der MTMM-Matrix eher gering ausfällt (vgl. Schuler, 1989 und Kapitel 2). Dieses sogenannte Validitätsparadox ist schon seit längerem Gegenstand kontroverser Diskussionen in der AC-Forschung. Nach Arthur et al. (2000) lassen sich zwei Arten von Erklärungen unterscheiden. Zum einen wird vermutet, dass die Beobachterurteile stark messfehlerbehaftet sind, so dass keine Konstruktvalidität nachgewiesen werden kann. Daher sind methodische Verbesserungen bei Auswahl der Dimensionen, Art der AC-Implementierung etc. notwendig (eine Übersicht findet sich bei Lievens, 1998). Andere Autoren halten es für möglich, dass im AC andere als die intendierten Konstrukte erfasst werden (siehe Klimoski & Brickner, 1987). Nach Kleinmann (1993) könnte CDD ein solches Konstrukt darstellen, das für die gemeinsame Varianz zwischen Prädiktor (z.B. AC) und Kriterium (z.B. Vorgesetztenbeurteilung) verantwortlich ist. Eine erste empirische Bestätigung erfuhr diese Annahme durch Kleinmann (1997b). Die abschließende Hypothese lautet damit:

Hypothese 8: Ein Teil der gemeinsamen Varianz von MMI und AC kann auf CDD zurückgeführt werden.

3.2 Methode

3.2.1 Überblick

Es wurde ein eintägiges Bewerbungstraining für Studierende am Ende des Studiums und interessierte Berufstätige konzipiert. Der Vormittag bestand aus drei typischen AC-Übungen (Gruppendiskussion mit/ohne Rollenvorgabe, Postkorb; kurz GD1, GD2 und PK), am Nachmittag wurden die wichtigsten Komponenten des MMI durchgeführt (Selbstpräsentation, Biographisches und Situatives Interview; kurz SP, BI und SI). Die Beobachter wurden in einem Frame-of-Reference-Training (vgl. Arthur et al., 2000) auf ihre Aufgabe vorbereitet. In einer Vorstudie mit erfahrenen Beobachtern waren potentiell relevante Dimensionen für die einzelnen Übungen erarbeitet worden. Im Bewerbungstraining wurden die Teilnehmer dann auf einigen dieser Dimensionen bewertet. Außerdem wurden die Hypothesen der Teilnehmer über die bewerteten Dimensionen erfasst sowie die Korrektheit dieser Annahmen bewertet. Insgesamt ist das Vorgehen als analog zu Kleinmann (1993) anzusehen.

3.2.2 Vorversuch

Zur Entwicklung des Bewerbungstrainings waren verschiedene Vorarbeiten notwendig, die ausführlich in Studie 1 (Kapitel 2) beschrieben sind. In einem ersten Workshop mit zehn erfahrenen studentischen Beobachtern wurden zunächst acht gut beobachtbare und hinreichend unabhängige Beobachtungsdimensionen entwickelt. Im Einzelnen waren dies Systematisches Denken & Handeln (SDH), Zusammenarbeit (ZU), Steuerung sozialer Prozesse (SSP), Umgang mit Informationen (UI), Wirtschaftliches Denken (WD), Ausdruck/mündliche Formulierung (AF), Engagement/Initiative (EI) und Fachwissen (FW).

Im Vorfeld des zweiten Workshops wurden durch die drei Experten verschiedene AC-Übungen ausgesucht. Diese wurden mit zehn anderen studentischen Beobachtern durchgespielt und die Relevanz der oben entwickelten Dimensionen für das erfolgreiche Abschneiden in den einzelnen Übungen bestimmt. Es ergaben sich zwei Übungen (führerlose Gruppendiskussionen, einmal mit und einmal ohne Rollenvorgabe), bei denen die drei Dimensionen – Systematisches Denken & Handeln (SDH), Zusammenarbeit (ZU) und Steuerung sozialer Prozesse (SSP) – übereinstimmend am relevantesten eingeschätzt wurden sowie je drei Dimensionen,

die zwar beobachtbar, aber weniger relevant für den Erfolg bei diesen Übung waren (sogenannte Distraktoren). Außerdem wurde ein Postkorb ausgewählt, bei dem die verschiedenen Auswertungshinweise den drei relevanten Dimensionen zugeordnet werden konnten (z.B. „Teilnehmer erkennt Kollision von Termin A und B“ entspricht der Dimension Systematisches Denken & Handeln). Hierdurch war es möglich, im Postkorb nicht nur einen Gesamtpunktwert zu erhalten, sondern auch noch Punktwerte auf den verschiedenen Dimensionen.

Im Vorfeld des dritten Workshops wählten zwei Experten der Personalauswahl insgesamt 34 situative und biographische Fragen aus, die zur Erfassung der drei Anforderungsdimensionen brauchbar erschienen. Als Grundlage dienten dabei die Arbeiten von Borchert (2001), Deller (1991), Klehe (2000) sowie von Schuler und Kollegen (Deutscher Sparkassen- und Giroverband, 1988). Im eigentlichen Workshop wurden dann mit Hilfe von zehn erfahrenen studentischen Beobachtern die Verständlichkeit der situativen und biographischen Fragen überprüft sowie die Zuordnung von Fragen zu Dimensionen vorgenommen. Für jede Frage lagen damit eine Vielzahl von qualitativen und quantitativen Daten vor (z.B. Beobachterübereinstimmung, Verteilung der Bewertungen, Beobachterhinweise zu den verhaltensverankerten Beobachtungsskalen), die von den zwei Experten zur endgültigen Auswahl der Fragen herangezogen werden konnten. Wichtigstes Kriterium war dabei ihre eindeutige Zuordnung zu einer einzigen Dimension. Insgesamt ergaben sich so 12 situative und 12 biographische Fragen, so dass jede der drei Dimensionen (Systematisches Denken & Handeln, Zusammenarbeit und Steuerung sozialer Prozesse) mit jeweils vier Fragen erfasst werden konnte.

3.2.3 Teilnehmer

Es wurden insgesamt 15 Bewerbungstrainings durchgeführt, wobei die Daten von 95 Teilnehmern und Teilnehmerinnen erhoben werden konnten (48 Frauen und 47 Männer). Das Alter der Teilnehmer lag zwischen 21 und 36 Jahren mit einem Mittelwert von 26.87 Jahren (*SD* = 2.95). Die meisten Teilnehmer kamen aus dem Bereich Wirtschaftswissenschaften (41.3%) bzw. aus den Naturwissenschaften (21.7%) und hatten im Durchschnitt 9.68 Semester studiert (*SD* = 3.63). Insgesamt 26 Teilnehmer (27.4%) hatten ihr Studium bereits beendet. Nur 5 Teilnehmer (5.3%) verfügten über Erfahrungen mit AC. Die Anwerbung der Teilnehmer erfolgte hauptsächlich über eine eigens entworfene Internetseite und mit Hilfe von

Aushängen und Handzetteln. Die Teilnehmer wurden darüber informiert, dass die Trainings im Rahmen eines wissenschaftlichen Forschungsprojektes stattfinden. Ziel der Untersuchung sei die Verbesserung von AC-Übungen. Zur Erhöhung des Commitments mussten die Teilnehmer im Vorfeld des Trainings eine Teilnahmegebühr von 15 € entrichten.

3.2.4 Beobachter und Beobachtertraining

Die Beobachter waren größtenteils Studierende der Psychologie im Hauptstudium mit dem Schwerpunkt Arbeits- und Organisationspsychologie, die mit der Teilnahme an der Studie ihre fachliche Qualifikation erhöhen wollten und während der gesamten Datenerhebung keine Informationen über die der Studie zugrunde liegenden Hypothesen erhielten. Die Beobachterschulung wurde im Sinne eines „Frame-of-Reference-Trainings“ (vgl. Arthur et al., 2000 bzw. Kapitel 2.2.3.3) durchgeführt. Hierzu wurden die Beobachter in einem eintägigen Training ausführlich mit den drei Dimensionen, möglichen Beobachterfehlern, Beobachterunterlagen, Ablauf des Bewerbungstrainings und den einzelnen Übungen vertraut gemacht.

3.2.5 Hauptstudie

An jedem Bewerbungstraining nahmen zwei Moderatoren sowie acht Beobachter und maximal acht Teilnehmer teil. Aufgrund von Fragestellungen, die in einer Folgestudie bearbeitet werden (vgl. Kapitel 4), fanden die AC-Übungen vormittags statt und das MMI am Nachmittag. Außerdem wechselte die Zuordnung von Teilnehmern und Beobachtern, so dass die Teilnehmer im AC durch andere Beobachter bewertet wurden als im MMI.

Da im Rahmen der Studie CDD (capability to discern dimensions) untersucht werden sollte, wurden den Teilnehmern verschiedene Hinweise bezüglich der Dimensionen gegeben, wie sie auch für reale Bewerbungssituationen typisch sind. Zunächst erstellten die Teilnehmer im Vorfeld Bewerbungsunterlagen auf eine mit der Teilnahmebestätigung verschickte Stellenanzeige, in welcher jede der drei relevanten Anforderungsdimensionen (Systematisches Denken & Handeln, Zusammenarbeit und Steuerung sozialer Prozesse) durch jeweils zwei Hinweise verschlüsselt mitgeteilt wurde (z.B. „...wenn Sie zudem jemand sind, der mit ständig wachsender Komplexität umgehen kann“ für die Dimension Systematisches Denken & Handeln). Zum Beginn des eigentlichen Bewerbungstrainings erhielten die Teilnehmer

außerdem einen fiktiven, schriftlichen Erfahrungsbericht eines Mitarbeiters, in dem ebenfalls in kodierter Form die drei kritischen Dimensionen benannt wurden. Zudem waren die Übungen und Interviewfragen passend zu den Dimensionen ausgewählt worden (vgl. Vorversuch). Weitere Quellen, aufgrund derer die Teilnehmer Hinweise auf die relevanten Dimensionen bekamen (z.B. Verhalten anderer Teilnehmer und/oder der Beobachter, vgl. Kleinmann, 1993), konnten nicht planmäßig kontrolliert werden. Die Hypothesen bzw. Annahmen der Teilnehmer über die erfassten Dimensionen wurden analog zu Kleinmann (1993) erfasst. Das Vorgehen wird ausführlich in den folgenden Abschnitten beschrieben.

Der Trainingstag begann mit der Begrüßung der Teilnehmer durch die Moderatoren und dem Einsammeln der Bewerbungsunterlagen. Die Teilnehmer wurden über den Ablauf des Tages informiert und darauf hingewiesen, dass sie nach jeder Übung mit Hilfe eines kurzen Fragebogens – dem sogenannten Reflexionsbogen – über ihre Wahrnehmung bezüglich der Übung befragt werden würden. Dieser Fragebogen diene Forschungszwecken und werde nicht in die Bewertung des AC eingehen. Dann wurde die erste Übung durchgeführt (führerlose Gruppendiskussion ohne Rollenvorgabe). Die Teilnehmer wurden hierzu in zwei Gruppen aufgeteilt und in die entsprechenden Räume geschickt. Nach der Übung kamen die Teilnehmer wieder im Hauptraum zusammen und füllten ihre Reflexionsbögen aus. Auf der ersten Seite wurde erklärt, dass der Fragebogen dazu diene, die Annahmen und Gedanken der Teilnehmer über die bewerteten Beobachtungskriterien zu erfassen und nicht in die Bewertung des AC eingehen werde. Die zweite Seite war mit folgendem Satz überschrieben: „Bei der Durchführung der Übung hatte ich die Hypothese(n), dass es auf die folgenden Verhaltensweisen ankommt“. In diesem Fragebogen konnten die Teilnehmer bis zu maximal sechs Hypothesen aufschreiben und in bezug auf die Wichtigkeit für ihr tatsächliches, während der Übung gezeigtes Verhalten in eine Rangreihe bringen. Dieses Vorgehen, d.h. Übung bzw. Interviewkomponente mit anschließendem Reflexionsbogen, wurde den ganzen Tag über hinweg beibehalten. Im Unterschied zu den AC-Übungen und zur Selbstpräsentation befanden sich auf den Reflexionsbögen zum BI und SI jeweils alle Interviewfragen im Originalwortlaut und die Teilnehmer konnten zu jeder einzelnen Frage bis zu zwei Hypothesen aufschreiben.

Am Ende des Tages wurden die Teilnehmer mit den im Vorversuch entwickelten Dimensionen vertraut gemacht und bekamen eine Liste mit den entsprechenden

Verhaltensankern ausgeteilt. Ferner bekamen die Teilnehmer die im Laufe des Tages ausgefüllten Reflexionsbögen mit den von ihnen notierten Hypothesen zurück. Ihre Aufgabe bestand nun darin, jede der von ihnen im Laufe des Tages aufgestellten Hypothesen einer einzigen von sechs vorgegebenen Dimensionen (die drei relevanten Dimensionen plus die drei Distraktoren, vgl. Kapitel 3.2.2) zuzuordnen, und zwar derjenigen Dimension, auf welche die entsprechende Hypothese am besten passte. Das Ausmaß der Passung von Hypothese und Dimension wurde, anders als bei Kleinmann (1993), durch ein Rating von 1 (passt etwas) bis 4 (passt vollständig) durch die Teilnehmer selbst nochmals beurteilt. Diese geänderte Art der Operationalisierung von CDD hatte sich in einem Vorversuch, der im Rahmen eines anderen Forschungsprojekts durchgeführt worden war (vgl. Mucha & Waldeyer, 2002), als reliabler gegenüber dem bisherigen Vorgehen erwiesen. Wiederum wurden die Teilnehmer darauf hingewiesen, dass die Fragebögen lediglich zu Forschungszwecken dienten und nicht in die Gesamtbewertung eingingen. Das Ausmaß des Erkennens (CDD) wurde dann berechnet, indem die Ratings der Hypothesen addiert wurden, welche den korrekten, d.h. relevanten Dimensionen zugeordnet waren. Hierbei wurden in jeder Übung drei Ratings genommen, und zwar das jeweils höchste Rating für jede der drei relevanten Dimensionen.

Während die Teilnehmer ihre Reflexionsbögen ausfüllten, bewerteten die Beobachter getrennt voneinander die Teilnehmerleistungen in der zuletzt durchgeführten Übung. Grundlage hierfür waren ihre handschriftlichen Notizen und die verhaltensverankert definierten Dimensionen. Nach Durchführung der drei AC-Übungen kamen die Beobachter zur ersten Beobachterkonferenz zusammen. Sie wurden dazu angehalten, insbesondere solche Bewertungen zu diskutieren, in denen Beurteilungsunterschiede von zwei oder mehr Punkten (5-stufige Beurteilungsskala) vorkamen. Auf diese Weise konnten die unterschiedlichen Beurteilungen meist nach kurzer Debatte geklärt werden (z.B. wenn ein Beobachter bestimmte Verhaltensweisen übersehen hatte). Eine Einigung der Beobachter musste jedoch nicht erfolgen. Die endgültige Leistungsbewertung der Teilnehmer erfolgte durch Mittelwertbildung der verschiedenen Beobachterurteile. Am Nachmittag wurden die Beobachter dann in Zweierteams aufgeteilt und zwei neuen Teilnehmern zugeordnet, die sie außer in einer kurzen Begrüßungsrunde bisher noch nicht gesehen hatten. Nach Durchführung aller Komponenten des MMI führten diese Zweierteams eine weitere Beobachterkonferenz in der oben beschriebenen Art durch.

3.3 Ergebnisse

3.3.1 Hypothese 1

In Hypothese 1 wurde postuliert, dass sich die Teilnehmer darin unterscheiden, in welchem Ausmaß sie die tatsächlichen Anforderungsdimensionen erkennen. Die Erhebung und Berechnung der CDD-Werte wurden bereits im Methodenteil ausführlich dargestellt. In Tabelle 11 folgt eine Übersicht über die theoretisch möglichen CDD-Werte und wichtige empirische Verteilungsparameter (Minimum, Maximum usw.) sowie ein Test auf Normalverteilung.

Tabelle 11: Verteilung der capability-to-discern-dimensions-(CDD)-Werte

	Min/Max Theorie	Min/Max Empirie	*M*	*SD*	Test auf Normal-verteilung[a]	*p* (2-seitig)
SP	0-12	0-10	2.59	2.05	1.35	.05
BI	0-48	9-46	27.55	8.09	.72	.68
SI	0-48	3-43	22.61	8.55	.93	.36
Interv. (BI & SI)	0-96	16-86	50.16	14.92	.67	.76
MMI	0-108	16-90	52.74	15.81	.47	.98
MMI(z)[b]	z-Werte	-1.24-1.21	0.00	0.55	.52	.95

Anmerkung: *N* = 95. Min/Max = Minimum und Maximum der CDD-Werte. SP = Selbstpräsentation. BI = Biographisches Interview. SI = Situatives Interview. Interv. = Interview aus BI und SI. MMI = Multimodales Interview.

[a] Kolmogorov-Smirnov-Statistik. [b] Erläuterung siehe Text.

Da es in der Selbstpräsentation deutlich weniger CDD-Punkte zu erreichen gab als im BI bzw. SI, wurden die CDD-Gesamtwerte auf zwei unterschiedliche Arten berechnet. In der ersten Variante wurden die Punktwerte in den einzelnen Komponenten einfach addiert. Hierdurch hat die Selbstpräsentation einen relativ geringen Einfluss auf den Gesamtwert. In der zweiten Variante wurden die Punktzahlen in den einzelnen Komponenten zunächst z-transformiert und anschließend addiert, so dass die einzelnen Interviewteile den gleichen Beitrag zum Gesamtergebnis leisteten. Auf die Ergebnisse in den weiteren Hypothesen hatte diese Unterscheidung keinen Einfluss, so dass sie im Folgenden nicht berichtet wird.

Wie auch schon bei Kleinmann (1993) sind die empirischen Verteilungen der CDD-Werte angenähert normalverteilt und nutzen dabei fast die gesamte Spannweite der theoretisch möglichen Werte aus. Die Teilnehmer unterscheiden sich somit deutlich darin, in welchem Ausmaß sie die zugrunde liegenden Anforderungsdimensionen erkennen. Im Folgenden wird nun auf die Frage eingegangen, ob diese Unterschiede auf eine Fähigkeit zurückzuführen sind. Analog zu Kleinmann (1993) berechneten wir daher zunächst die interne Konsistenz (Cronbach´s Alpha) der CDD-Werte. Die entsprechenden Werte sind in Tabelle 12 dargestellt.

Tabelle 12: Reliabilität der capability-to-discern-dimensions-(CDD)-Werte im Multimodalen Interview (MMI) und Assessment Center (AC)

	Alpha	Anzahl und Art der Items
Selbstpräsentation (SP)	.14	3 Dimensionen
Biographisches Interviews (BI)	.65	12 Items
Situatives Interview (SI)	.65	12 Items
Interview (BI & SI)	.78	24 Items
MMI	.73	9 Dimensionen
AC	.59	9 Dimensionen

Anmerkung: *N* = 95.

Wie erwartet zeigten sich, außer für die Selbstpräsentation, zufriedenstellend hohe interne Konsistenzen. Cronbach´s Alpha wurde mit Hilfe einer Varianzanalyse auf Signifikanz getestet (vgl. Hoyt, 1941). Hierbei zeigte sich, dass die inneren Konsistenzen für alle berechneten Skalen signifikant waren. Das Erkennen der Dimensionen variiert also nicht zufällig von Aufgabe zu Aufgabe, was als Hinweis darauf gewertet kann, dass es eine Fähigkeit, Anforderungsdimensionen zu erkennen (CDD), gibt, und dass sich die Teilnehmer in der Ausprägung dieser Fähigkeit unterscheiden. Zum Vergleich sind in Tabelle 12 auch die Ergebnisse für das AC aufgeführt. Hierbei ergab wie bei Kleinmann (1993) ein Cronbach´s Alpha von .59.

Kleinmann (1993) weist darauf hin, dass die Teilnehmer bei der Zuordnung von Hypothesen und Dimensionen möglicherweise raten und dass daher – ähnlich wie in einem Multiple-Choice-Test – Teilnehmer mit vielen Hypothesen auch mehr richtige Dimensionen treffen und damit höhere CDD-Werte erhalten. Ein positiver

Zusammenhang zwischen Anzahl der Hypothesen und Anzahl der erkannten Dimensionen kann jedoch nicht unbedingt als Beleg für ein Raten der Teilnehmer herangezogen werden. Stattdessen ist es auch möglich, dass Personen mit einer besseren Wahrnehmungsfähigkeit entsprechend mehr Hinweisreize verarbeiten und auf diese Weise mehr Hypothesen aufstellen. Wir wollen uns daher nun überlegen, was für ein Korrelationsmuster sich ergeben müsste, wenn die Teilnehmer die Zuordnung von Hypothese und Dimension aufgrund von Zufallseffekten (Raten) vornehmen würden. Wenn man davon ausgeht, dass die Wahrscheinlich eine relevante Dimension zu erraten einen bestimmten Wert annimmt, dann sollte mit steigender Anzahl an Hypothesen sowohl die Anzahl „falscher“ Hypothesen (d.h. Hypothesen, die einer Distraktor-Dimension zugeordnet werden) als auch die Anzahl „richtiger“ Hypothesen (d.h. Hypothesen, die einer „tatsächlichen“ Dimension zugeordnet werden) ansteigen. Zur Veranschaulichung ein Beispiel: Bei einer Ratewahrscheinlichkeit von 60% sollte ein Teilnehmer mit 100 Hypothesen 60 richtige und 40 falsche Hypothesen haben, ein Teilnehmer mit 110 Hypothesen 66 richtige und 44 falsche usw. Als Folge des Ratens sollte sich also eine hohe positive Korrelation zwischen Anzahl falscher und Anzahl richtiger Hypothesen ergeben. Empirisch ergibt sich jedoch eine negative Korrelation von r = -.19, *ns*. Das gefundene Korrelationsmuster widerspricht somit der „Rate-Hypothese“, wodurch die Erklärung gestärkt wird, dass sich in den CDD-Werten die tatsächliche Erkennensleistung widerspiegelt.

3.3.2 Hypothese 2

In der zweiten Hypothese wird die Annahme überprüft, dass Personen, die in größerem Ausmaß die zugrunde liegenden Anforderungsdimensionen erkannt haben, auch besser bewertet werden. Um diese Hypothese zu prüfen, wurden Korrelationen zwischen den CDD-Werten und zwei unterschiedlichen Leistungsmaßen berechnet. Dies sind zum einen der Mittelwert der dimensionsbezogenen Ratings in der entsprechenden Übung und zum anderen die Gesamtbewertung in der Übung. Letztere sollte von den Beobachtern unabhängig von den Dimensionen eingeschätzt werden. Die Ergebnisse sind in Tabelle 13 dargestellt.

Tabelle 13: Produkt-Moment-Korrelation von capability-to-discern-dimensions-(CDD)-Werten und Leistungsmaßen

	Korrelation CDD mit	
	Mittelwert Dimensionen	Gesamtbewertung
Selbstpräsentation (SP)	.09	.06
Biographisches Interviews (BI)	.31**	.21**
Situatives Interview (SI)	.38**	.31**
Interview (BI & SI)	.35**	.27**
Multimodales Interview (MMI)	.26**	.19*
Gruppendiskussion 1	.26**	.26**
Gruppendiskussion 2	.33**	.30**
Postkorb	.19*	.30**
Assessment Center (AC)	.39**	.44**

Anmerkung: $N = 95$.

*p < .05. **p < .01 (jeweils 1-seitig).

Wie erwartet ergaben sich sowohl für das Interview als auch für das AC signifikante Zusammenhänge zwischen den CDD-Werten und den beiden jeweils zugehörigen Leistungsmaßen. Lediglich die Erkennenswerte in der Selbstpräsentation korrelierten nicht mit den zugehörigen Fremdbeurteilungen. Hypothese 2 kann damit – von der Selbstpräsentation abgesehen – eindeutig bestätigt werden.

Ein weiteres interessantes Ergebnis ergibt sich, wenn man die Beurteilungen auf den einzelnen Dimensionen mit den zugehörigen Erkennenswerten korreliert. Während im MMI die Korrelationen für alle drei Dimensionen eine ähnliche Höhe aufweisen, gibt es im AC lediglich für die beiden Dimensionen Zusammenarbeit und Steuerung sozialer Prozesse signifikante Korrelationen zwischen Erkennensleistung und Beobachterurteil. Erkennen die Teilnehmer hingegen im AC, dass die Dimension Systematisches Denken & Handeln erfasst werden soll, so schneiden sie nicht unbedingt besser auf ihr ab.

3.3.3 Hypothese 3

Analog zu Kleinmann (1993) wird in Hypothese 3 postuliert, dass Personen in den Interviewteilen besser abschneiden, deren zugrunde liegende

Anforderungsdimension sie vollständig erkannt haben. Anders formuliert: Die intraindividuelle Variation der Leistungsbeurteilungen wird zumindest zum Teil dadurch erklärt, ob der Teilnehmer die Anforderungsdimension erkannt hat oder nicht. Eine Dimension bzw. Interviewfrage wird dabei als vollständig erkannt bezeichnet, wenn ein Teilnehmer seine Hypothese der korrekten Dimension zuordnet und die Passung von Hypothese und Dimension ein Rating von 4 (passt vollständig) bekommen hat. Eine Dimension bzw. Interviewfrage wird hingegen als nicht erkannt bezeichnet, wenn der Teilnehmer seine Hypothese einer falschen Dimension, d.h. einem Distraktor, zugeordnet hat. Anschließend wurde pro Person und Bedingung (erkannt vs. nicht erkannt) die mittlere Leistungsbeurteilung berechnet. In den einzelnen Interviewkomponenten Selbstpräsentation, BI bzw. SI ergaben sich auf diese Weise 27, 77 bzw. 78 Personen, die mindestens eine Dimension vollständig erkannt hatten bzw. nicht erkannt hatten. Für das Gesamtinterview waren es dann entsprechend mehr, und zwar 83 Personen. Zum Schluss wurden die Unterschiede in den mittleren Bewertungen mit Hilfe von t-Tests für abhängige Gruppen auf Signifikanz getestet (vgl. Tabelle 14).

Tabelle 14: T-Test für abhängige Gruppen. Mittlere Leistung der Teilnehmer in den vollständig erkannten bzw. nicht erkannten Fragen

		Dimension nicht erkannt		Dimension erkannt			
	N	*M*	*SD*	*M*	*SD*	*df*	*t*
Selbstpräsentation (SP)	27	2.75	1.11	2.45	.76	26	1.94 *
Biographisches Interviews (BI)	77	2.88	.64	2.54	.83	76	3.41 ***
Situatives Interview (SI)	78	2.89	.59	2.46	.64	77	4.67 ***
Multimodales Interview (MMI)	83	2.88	.45	2.57	.65	82	4.10 ***

Anmerkung: Beurteilung der Leistung von 1 = erfüllt die Anforderungen vollständig bis 5 = starke Veränderungen wünschenswert.

*p < .05. ***p < .001 (jeweils 1-seitig).

Die Unterschiede in allen Verfahren waren signifikant. Wie erwartet wurden die Teilnehmer besser bewertet, wenn sie die Anforderungsdimensionen erkannt hatten.

3.3.4 Hypothese 4

In der vierten Hypothese wird die Vermutung überprüft, dass sich das Ausmaß, in dem die Anforderungsdimensionen erkannt werden, auf die konvergente Validität des MMI auswirkt. Da jede Dimension im BI und SI durch vier Fragen erfasst wird und drei Dimensionen beurteilt wurden, gibt es für jeden Teilnehmer 72 verschiedene Wertepaare, in denen beide Bewertungen durch die Beobachter auf der gleichen Dimension vorgenommen wurden. Jedes Wertepaar wurde nun einer von drei Gruppen zugeordnet (vgl. Kleinmann, 1993):

Gruppe 1: Keine Dimension des Itempaares wurde erkannt.
Gruppe 2: Eine Dimensionen des Itempaares wurden erkannt.
Gruppe 3: Beide Dimensionen des Itempaares wurden erkannt.

Analog zu Hypothese 3 wurde eine Dimension als „erkannt“ bezeichnet, wenn ein Teilnehmer seine Hypothese der korrekten Dimension zuordnete und die Passung von Hypothese und Dimension ein Rating von 4 (passt vollständig) bekommen hatte. Im Gegensatz dazu wurde eine Dimension als „nicht erkannt“ bezeichnet, wenn der Teilnehmer seine Hypothese einer falschen Dimension, d.h. einem Distraktor, zugeordnet hatte. Da die Teilnehmer unterschiedlich gut im Erkennen der Anforderungsdimensionen waren, gab es bestimmte Teilnehmer, die häufiger bzw. seltener in den drei Gruppen zu finden sind als andere. Um diese Konfundierung von Teilnehmern und Gruppenzugehörigkeit aufzulösen, wurden nur diejenigen Teilnehmer ausgewählt, welche in allen drei Gruppen zu finden waren. Es wurde dann per Zufallszahl ein Wertepaar für jede der drei Gruppen ausgewählt. Nach Hypothese 4 sollten die Korrelationen in Gruppe 1 bzw. 3 und damit die konvergente Validität höher sein als in Gruppe 2. Die Ergebnisse finden sich in Tabelle 15.

Tabelle 15: Produkt-Moment-Korrelationen in den Gruppen 1, 2 und 3 als Maß für die konvergente Validität

	Gruppe 1	Gruppe 2	Gruppe 3
Korrelation	.08	-.04	.35
p (2-seitig)	.50	.71	.00

Anmerkung: *N* = 81. Gruppe 1 = keine Dimension erkannt. Gruppe 2 = eine Dimension erkannt. Gruppe 3 = 2 Dimensionen erkannt (Erläuterung im Text).

Es zeigte sich wie erwartet eine relativ hohe Korrelation in Gruppe 3. Diese war signifikant größer ist als in Gruppe 2 (z = 2.55; p < .01) bzw. Gruppe 1 (z = 1.82; p < .05). Die Korrelationen in den beiden letztgenannten Gruppen unterschieden sich jedoch nicht signifikant voneinander (z = 0.73; p = .23; zur Berechnung siehe Bortz, 1999).

3.3.5 Hypothese 5 bis 7

Hypothese 5 bis 7 basieren auf der Annahme, dass es sich bei CDD um eine personenbezogene Fähigkeit handelt, die relativ unabhängig vom diagnostischen Instrument (hier AC bzw. MMI) erfasst werden kann. Daher wurde zunächst in Hypothese 5 ein Zusammenhang zwischen den im AC bzw. MMI ermittelten CDD-Werten postuliert. Wie erwartet zeigte sich eine signifikante Korrelation von r = .46 (N = 95, p < .001). Da die Reliabilität der CDD-Werte (vgl. Hypothese 1) mit einem Cronbach´s Alpha von .73 bzw. .59 nur mittelhoch ist, wurde eine Minderungskorrektur durchgeführt. Für die messfehlerbereinigte Korrelation ergab sich auf diese Weise ein Wert von r = .69, was auf eine bedeutsame gemeinsame Varianz der CDD-Werte im Interview und AC schließen lässt.
Ein weiterer Hinweis darauf, dass die CDD-Werte in den einzelnen Übungen etwas ähnliches erfassen zeigt sich, wenn man die einzelnen CDD-Werte miteinander korreliert (vgl. Tabelle 16).

Tabelle 16: Produkt-Moment-Korrelationen zwischen den verschiedenen capability-to-discern-dimensions-(CDD)-Werten

CDD-Wert	1.	2.	3.	4.
1. in der Selbstpräsentation (SP)	-			
2. im Biographischen Interview (BI)	.30**	-		
3. im Situativen Interview (SI)	.38**	.61**	-	
4. im Assessment Center (AC)	.42**	.36**	.35**	-

Anmerkung: N = 95.

**p < .01 (1-seitig).

Es zeigte sich wie erwartet, dass die erfassten CDD-Werte positiv zusammenhingen. Dabei korrelierten die CDD-Werte der Selbstpräsentation am höchsten mit den CDD-Werten des AC (beides Übungen), während die CDD-Werte des BI am höchsten mit

den CDD-Werten des SI (beides Interviews) korrelierten. Das Korrelationsmuster der CDD-Werte könnte dabei sowohl auf einen Methodeneffekt hinweisen (Übungen vs. Interviews), als auch auf ein methodisches Artefakt (leichte Unterschiede in der Erfassung der CDD-Werte, vgl. Methodenteil).

In Hypothese 6 und 7 wurden dann die Hypothesen geprüft, dass die im AC ermittelten CDD-Werte mit den Leistungsbeurteilungen im Interview zusammenhängen bzw. die im Interview ermittelten CDD-Werte mit den gezeigten Leistungen im AC. Wiederum ergaben sich signifikante Korrelationen von r = .26 (N = 95, p < .01) bzw. r = 35 (N = 95, p < .001), so dass auch diese beiden Hypothesen als bestätigt angesehen werden können.

3.3.6 Hypothese 8

In der letzten Hypothese wurde dann die Annahme überprüft, dass die Korrelation zwischen MMI und AC zumindest zum Teil auf CDD zurückzuführen ist. Zur Prüfung der Hypothese wurden zunächst die Leistungswerte von MMI und AC korreliert. Hierbei ergab sich ein Wert von r = .41 (df = 94, p < .001). Anschließend wurden sowohl im MMI als auch AC die zugehörigen CDD-Werte aus den Leistungsbeurteilungen herauspartialisiert und erneut eine Korrelation zwischen den Interview und AC berechnet. Nun ergab sich eine Korrelation von r = .32 (df = 91, p < .01). Letztere ist deskriptiv gesehen niedriger als die zunächst berechnete Korrelation, jedoch immer noch signifikant von Null verschieden. In Ermangelung eines geeigneten Testverfahrens konnten die beiden Korrelationen leider nicht direkt auf einen signifikanten Unterschied getestet werden.

3.4 Diskussion

In Hypothese 1 wurde die Annahme überprüft, dass sich die Teilnehmer darin unterscheiden, in welchem Ausmaß sie die Anforderungsdimensionen im strukturierten Interview erkennen. Analog zu Kleinmann (1993) wurden daher Verteilung und innere Konsistenz (Cronbach´s Alpha) der CDD-Werte untersucht. Die Analyse der CDD-Skala ergab für das Gesamtinterview (MMI) eine zufriedenstellende innere Konsistenz. Bei Betrachtung der einzelnen Komponenten zeigte sich für die Selbstpräsentation ein ungenügender Wert für Cronbach´s Alpha. Für das AC sowie das BI und SI ergaben sich hingegen ausreichende bis

befriedigende innere Konsistenzen, die in ähnlicher Höhe auch schon bei Kleinmann (1993) berichtet werden. Weiterhin waren die CDD-Werte angenähert normalverteilt und umfassten beinahe die gesamte Spannweite der theoretisch möglichen Werte.

Im Zusammenhang mit der ersten Hypothese wurde auch die Frage untersucht, ob die CDD-Werte durch Rate-Effekte der Teilnehmer beeinflusst werden. Wie oben ausgeführt, müsste sich bei einer zufälligen Zuordnung von Hypothesen und Dimensionen eine hohe positive Korrelation zwischen Anzahl der den richtigen Dimensionen zugeordneten Hypothesen und Anzahl der den Distraktoren zugeordneten Hypothesen ergeben. Empirisch zeigte sich jedoch ein tendenziell *negativer* Zusammenhang, was gegen eine zufällige Zuordnung spricht. Doch auch aus inhaltlichen Gründen ist die Rate-Hypothese unplausibel. Anders als in einem Multiple-Choice-Test können die Teilnehmer verschiedene Hypothesen auch der gleichen Dimension zuordnen. Ein Mehr an Hypothesen führt daher nicht automatisch zu einem Mehr an richtigen Dimensionen bzw. einem höheren Ausmaß des Erkennens.

Trotzdem könnte jedoch kritisch angemerkt werden, dass Hypothesen, die den falschen Dimensionen zugeordnet wurden, keine negative Gewichtung erhielten. Dies ist unseres Erachtens jedoch auch nicht sinnvoll. So wurden – insbesondere in den AC-Übungen – die Distraktoren ja gezielt danach ausgesucht, dass sie ebenfalls inhaltliche Plausibilität besitzen. Die Distraktoren sollten lediglich weniger bedeutsam für den Erfolg in der Übung sein, als die „wahren“ Dimensionen d.h. die Dimensionen, auf welchen die Teilnehmer letztendlich beurteilt wurden. Eine Person mit sehr hoher Erkennensfähigkeit, die neben den drei „echten“ Dimensionen auch noch zwei weitere plausible Distraktoren erkennt, würde daher bei einer negativen Gewichtung ungerechtfertigterweise einen Malus bekommen.

Zusammenfassend lässt sich festhalten, dass eine künstliche Erhöhung der CDD-Werte aufgrund von Rate-Effekten unwahrscheinlich ist. Vielmehr kann man davon ausgehen, dass es die höhere Erkennensfähigkeit ist, die zu einer höheren Anzahl an Hypothesen führt. Insbesondere liegen inzwischen erste Ergebnisse zur Konstruktaufklärung der CDD-Werte vor (vgl. Hartstein & Kleinmann, 2002). Dort konnten Zusammenhänge zu verschiedenen konstruktnahen Verfahren (z.B. verbale Intelligenz, soziale Wahrnehmung, Impression Management) aufgezeigt werden.

In Hypothese 2 wurde der Zusammenhang zwischen der Leistungsbeurteilung und dem Ausmaß des Erkennens untersucht.

Wie erwartet zeigten sich signifikante Korrelationen zwischen CDD-Werten und den zugehörigen Leistungsmaßen. Diese lagen im gleichen Größenbereich wie sie auch bei Kleinmann (1993, 1997b) berichtet werden. Lediglich für die Selbstpräsentation konnte die Hypothese nicht bestätigt werden. Dies könnte durch die geringe Reliabilität der CDD-Werte in der Selbstpräsentation (vgl. Hypothese 1) erklärt werden. Eine genauere Betrachtung der Ergebnisse zeigt, dass die Korrelationen zwischen CDD-Werten und dem Mittel der Dimensionen etwas höher lagen, als die Korrelationen mit der Gesamtbewertung. Wir vermuten, dass dies auf die höhere Reliabilität des erstgenannten Leistungsmaßes zurückzuführen ist. Eine Alternativerklärung würde hingegen die Konstruktvalidität der CDD-Werte stützen. So wurden die Beobachter explizit dazu aufgefordert, in die Gesamtbewertung weitere Kriterien – wie z.B. verbaler Ausdruck – zur Beurteilung heranzuziehen. Im CDD-Wert ist jedoch nur das Ausmaß des Erkennens für die drei Anforderungs-dimensionen (SDH, ZU und SSP) erfasst, so dass die höheren Korrelationen mit dem Mittelwert der drei Dimensionen plausibel sind.

Eine wichtige Einschränkung unserer Studie betrifft die Interpretation der Kausalrichtung unserer Korrelationen. So ist unklar, ob die Teilnehmer gut abschneiden, weil sie die Dimensionen erkannt haben oder ob Teilnehmer auf bestimmten Dimensionen gut sind und diese deswegen erkennen. Die Schwierigkeit einer eindeutigen Interpretation der Korrelationen liegt vor allem darin begründet, dass im Interview letztendlich nur drei verschiedene, relativ unabhängige Dimensionen erfasst wurden. Es ist daher möglich, dass ein besonders teamstarker Teilnehmer automatisch in jeder Übung die Hypothese aufstellt, dass Zusammenarbeit wichtig sei und auf diese Weise einen hohen CDD-Wert erreicht. Beabsichtigt ist jedoch, dass Teilnehmer erkennen, dass in einer spezifischen Übung Zusammenarbeit wichtig ist und sich deshalb versuchen teamfähig zu geben (unabhängig davon, welche maximale Teamfähigkeit sie haben...). In einer zukünftigen Studie sollte daher ein Interview zum Einsatz kommen, welches eine Vielzahl von unterschiedlichen Dimensionen erfassen kann. Ein ähnliches Vorgehen hat auch Kleinmann (1993) gewählt. Sein AC bestand aus fünf Übungen, in denen jeweils vier Dimensionen erfasst wurden. Insgesamt kamen 12 verschiedene Dimensionen zum Einsatz, die z.T. durch gegensätzliche Verhaltensweisen definiert wurden (z.B. Durchsetzungsfähigkeit und Teamfähigkeit). Bei einem solchen Design sind Korrelationen zwischen CDD-Wert und Leistungsbeurteilung durch die

Beobachter eher auf die Erkennensfähigkeit zurückzuführen, während die Alternativerklärung (Teilnehmer ist auf allen zwölf! Dimensionen gut und erkennt sie deshalb) geringe Plausibilität besitzt.

In Hypothese 3 wurde postuliert, dass die Erkennensfähigkeit neben interindividuellen Leistungsunterschieden (vgl. Hypothese 2) auch intraindividuelle Unterschiede in den Bewertungen erklärt. Hierzu wurden für jeden Teilnehmer eine mittlere Bewertung für die Fragen berechnet, deren Dimension erkannt wurde bzw. nicht erkannt wurde und die Unterschiede in den Bewertungen mit Hilfe eines t-Tests für abhängige Gruppen auf signifikante Unterschiede getestet. Wie erwartet erzielten die Teilnehmer in den Fragen, deren Dimension sie erkannt hatten, signifikant bessere Bewertungen als in den Interviewfragen, deren Dimension sie nicht erkannt hatten. Doch auch dieses Ergebnis kann nicht ohne Einschränkung interpretiert werden. Denn auch wenn durch das gewählte Vorgehen die personenvermittelte Abhängigkeit von Erkennensfähigkeit und Leistung behoben werden konnte, so bleibt der fragenspezifische Zusammenhang von Erkennensfähigkeit und Beurteilung bestehen. So ist es beispielsweise möglich, dass die Fragen mit den besonders leicht zu durchschauenden Dimensionen auch die Fragen mit den „leichtesten“ Antwortankern waren, während die Fragen mit den schlechten Bewertungen auch fast nie erkannt worden. Wäre dies der Fall, dann hätten wir letztendlich schwere Fragen mit leichten Fragen verglichen und nicht wie gewünscht Fragen, deren Dimension erkannt wurde und Fragen, deren Dimension nicht erkannt wurde. Eine weitere Spezifizierung der Auswertung war jedoch nicht möglich, da sonst die Fallzahlen zu klein geworden wären.

In Hypothese 4 wurde der Einfluss der Erkennensfähigkeit auf die konvergente Validität des MMI untersucht. Hierzu wurden drei Gruppen von Beurteilungspaaren gebildet, in denen die Teilnehmer entweder keine, eine oder beide Anforderungs-dimensionen erkannt hatten. Wie erwartet ergaben sich in Gruppe 3 die höchsten Zusammenhänge. Die Teilnehmer in dieser Gruppe hatten bezüglich beider Fragen erkannt, welche Anforderungsdimensionen erfasst werden sollen. Daher war ein konsistentes Antwortverhalten möglich, welches zu einer entsprechend hohen Korrelation zwischen den Bewertungen in identischen Dimensionen führte. Anders als bei Kleinmann (1993) ergab sich auch in der Gruppe 1 (keine Anforderungsdimension erkannt) nur ein geringer Zusammenhang zwischen den Bewertungen. Diese Gruppe unterscheidet sich von Gruppe 2 und 3 insofern, als das

bei letzteren die Annahme über ein inkonsistentes bzw. konsistentes Antwortverhalten plausibel ist. Daher ist für Gruppe 2 und 3 die Vorhersage über eine niedrige bzw. hohe konvergente Validität möglich. Für Gruppe 1 weiß man hingegen nur, dass in beiden Fragen die Anforderungsdimension (z.B. Zusammenarbeit) nicht erkannt wurden. Ob das Antwortverhalten folglich konsistent oder inkonsistent war, lässt sich nur schwer plausibel begründen. Entsprechend hatten wir daher in dieser Gruppe keine Hypothese über die Höhe der konvergenten Korrelation aufgestellt.

In Hypothese 5 bis 7 wurden verschiedene Annahmen bezüglich der Zusammenhänge zwischen der im AC bzw. MMI ermittelten CDD-Werte sowie den verschiedenen Leistungsbeurteilungen überprüft. Wie erwartet ergaben sich signifikante Korrelationen zwischen den CDD-Werten aus AC und den CDD-Werten im Interview (korrigiert r = .69) sowie signifikante Zusammenhänge zwischen den CDD-Werten im Interview und den Leistungswerten im AC bzw. den CDD-Werten im AC und den Beurteilungen im Interview. Damit wird die Annahme gestützt, dass die Erkennensfähigkeit relativ unabhängig vom zugrunde liegenden Testverfahren erhoben werden kann. Unterstützung erfährt diese Annahme insbesondere dadurch, dass die Leistungsbeurteilung in einem eignungsdiagnostischen Verfahren (AC bzw. MMI) durch die Erkennensfähigkeit (CDD-Werte) aus einem *anderen* Verfahren vorhergesagt werden kann und dass diese Vorhersage ähnlich gut ausfällt, wie wenn man die zum Verfahren gehörigen CDD-Werte verwendet. Einschränkend muss jedoch festgehalten werden, dass im AC und Interview die gleichen Dimensionen benutzt wurden und daher ähnlich wie bei Hypothese 2 die Frage nach der Interpretation der Kausalrichtung der korrelativen Zusammenhänge gestellt werden muss. D.h. schneiden die Teilnehmer gut ab, weil sie gut im Erkennen der Anforderungen sind und daher die Dimensionen erkannt haben oder erkennen sie die Dimensionen, weil sie auf diesen sehr leistungsfähig sind? Um diese Frage besser beantworten zu können, sollte die Studie in modifizierter Form repliziert werden, wobei in Zukunft unterschiedliche Dimensionen im AC und Interview beurteilt werden sollten. Sollten die hier gefundenen Ergebnisse repliziert werden können, so würden unsere Hypothesen massive Unterstützung erfahren.

In der letzten Hypothese wurde der Zusammenhang zwischen den Leistungsbeurteilungen in AC und MMI um die Erkennensfähigkeit bereinigt. Wie erwartet kam es deskriptiv gesehen zu einem Absinken der Korrelation, wobei in Ermangelung eines geeigneten Testverfahrens kein Signifikanztest durchgeführt

werden konnte. Auch wenn der Effekt eher klein zu sein scheint, so erhält die Annahme, dass die gemeinsame Varianz zwischen Prädiktor und Kriterium nicht auf die erfassten Dimensionen zurückzuführen ist, sondern zumindest zum Teil durch andere Konstrukte erklärt werden kann (vgl. Arthur et al., 2000; Kleinmann, 1997b; Klimoski & Brickner, 1987), weitere empirische Unterstützung. Daher soll in einer weiteren Studie (vgl. Kapitel 4) mit Hilfe einer experimentellen Manipulation die Auswirkung der Erkennensfähigkeit auf die Kriteriumsvalidität des MMI untersucht werden.

Zusammenfassend können wir festhalten, dass durch die berichteten Ergebnisse die Bedeutsamkeit und Validität der CDD-Werte unterstützt wird. Insbesondere werden die Ergebnisse von Kleinmann (1993, 1997b) dahingehend erweitert, dass das Ausmaß des Erkennens nicht nur im AC, sondern auch im strukturierten Interview einen positiven Einfluss auf die gezeigte Leistung und die konvergente Validität besitzt. In Folgestudien soll die Konstruktaufklärung der CDD-Werte vorangetrieben werden. Insbesondere wird zu überprüfen sein, ob neben verbaler Intelligenz, sozialer Wahrnehmung und bestimmten Aspekten von Impression Management (vgl. Hartstein & Kleinmann, 2002) Zusammenhänge zu weiteren Konstrukten aufgezeigt werden können.

4 Transparenz der Anforderungsdimensionen. Ein Moderator der internen Konstruktvalidität des Multimodalen Interviews? [3]

Zusammenfassung. Welche Auswirkungen hat die Bekanntgabe der Anforderungsdimensionen (Transparenz) im Multimodalen Interview (MMI)? Dieser Frage wurde in zwei unabhängigen Experimentalstudien (Studie A mit *N* = 123 bzw. B mit *N* = 176) nachgegangen. Dabei ergaben sich für beide Studien übereinstimmende Ergebnisse. So kam es unter Transparenz zu einem Leistungsanstieg der Bewerber. Dies konnte aufgrund von Assessment Center (AC) Studien nicht erwartet werden. Erwartet und bestätigt werden konnte jedoch die höhere Beurteilerübereinstimmung in der Transparenz-Bedingung. Hypothesenkonform kam es außerdem zu einem Anstieg der internen Konstruktvalidität des MMI unter Transparenz. Weiterhin waren in Studie B die korrelativen Beziehungen zwischen MMI und einem Kriterium (AC) wie erwartet unter Transparenz tendenziell niedriger als unter Intransparenz. Implikationen dieser Ergebnisse für die Praxis und Vorschläge für weitere Forschung werden diskutiert.

4.1 Einleitung

Das Einstellungsinterview ist nach der Analyse der Bewerbungsunterlagen das am häufigsten verwendete Verfahren zur Auswahl von Bewerbern (Schuler et al., 1993; Schulz et al., 1985). Inzwischen existieren viele Belege, dass vor allem strukturierte[4] Interviews eine gute prognostische Validität besitzen (z.B. Huffcutt & Arthur, 1994; Schmidt & Hunter, 1998; Schmidt & Rader, 1999; Wiesner & Cronshaw, 1988; Wright et al., 1989). Strukturierte Interviews sind damit ein geeignetes Mittel, um Personen auszuwählen, welche später im Beruf – beispielsweise von ihren Vorgesetzten – als erfolgreich eingeschätzt werden. Offen bleibt, was mit Hilfe von Interviews gemessen wird bzw. warum strukturierte Interviews prognostisch valide Verfahren sind. Forschung zu dieser Frage fehlt und ist nach Ansicht verschiedener Autoren dringend notwendig (z.B. Harris, 1999; McDaniel et al., 1994). In dieser Studie wird

[3] Teilergebnisse dieser Studie wurden bereits auf dem DGPs Kongress 2002 in Berlin vorgestellt.

[4] Zum Begriff „strukturiert“ siehe Campion, Palmer & Campion, 1997.

untersucht, ob die Konstruktvalidität strukturierter Interviews durch die Bekanntgabe der Anforderungsdimension (Transparenz) gesteigert werden kann.

Die traditionelle Erklärung für die prognostische Validität eignungsdiagnostischer Verfahren ist nach Kleinmann (1997b), „daß es [...] möglich ist, die vorgegebenen berufsrelevanten Fähigkeiten/Anforderungsdimensionen der einzelnen Kandidaten richtig zu beurteilen; und daß es eben diese Fähigkeiten/Anforderungsdimensionen sind, die den späteren Berufserfolg ausmachen“ (S. 171). Diese traditionelle Deutung ist jedoch für das Interview unplausibel, denn analog zur AC-Forschung gibt es inzwischen auch bezüglich des strukturierten Interviews Untersuchungen, die bei Auswertung der Multitrait-Multimethod-(MTMM)-Matrix (vgl. Campbell & Fiske, 1959; Kleinmann & Köller, 1997) eine mangelhafte Konstruktvalidität festgestellt haben (siehe Lober, Kleinmann, Borchert & Richter, 2002; Schuler & Funke, 1989; Kapitel 2). In diesen Interview-Studien ergeben sich konsistent höhere heterotrait-monomethod-Korrelationen (diskriminante Validität) als monotrait-heteromethod-Korrelationen (konvergente Validität). Es ist daher zweifelhaft, dass im strukturierten Interview die intendierten, vorgegebenen Konstrukte erfasst werden.

Betrachtet man die verschiedenen Arten der Validität unter dem Gesichtspunkt einer einheitlichen Theorie der Validität (Binning & Barrett, 1989), so ergibt sich für das strukturierte Interview das gleiche Validitätsparadox, wie für das AC (vgl. Arthur et al., 2000; Kolk, 2001). In der Literatur zur AC-Forschung werden für dieses Validitätsparadox unterschiedlichste Gründe diskutiert (z.B. Klimoski & Brickner, 1987), und es gibt eine Vielzahl von Versuchen, die Konstruktvalidität des AC zu verbessern (eine Übersicht findet sich bei Lievens, 1998). Eine der diskutierten Erklärungen für die geringe Konstruktvalidität des AC ist die „Intransparenz der Anforderungsdimensionen“ (Bungard, 1987), welche in mehreren Studien von Kleinmann empirische Unterstützung erhielt (Kleinmann, 1991, 1993, 1997b). Intransparenz heißt dabei, dass die Teilnehmer nicht konkret erfahren, anhand welcher Kriterien sie beurteilt werden, d.h. welche Verhaltensweisen, Selbstaussagen etc. positiv bzw. negativ bewertet werden.

Da strukturierte Einstellungs-Interviews und AC einige Gemeinsamkeiten besitzen (simulationsorientierter Ansatz, Anforderungsbezug, Fremdeinschätzung durch mehrere Beobachter, intransparente Durchführung, mangelhafte Konstruktvalidität bei hoher prognostischer Validität, siehe auch Kapitel 1.4), wurden einige Auswirkungen der Intransparenz auch schon für das Multimodale Interview (MMI)

untersucht (vgl. Borchert, 2001; Kapitel 3). Hierbei konnten die Ergebnisse der AC-Forschung größtenteils repliziert werden.

Welche Folgen ergeben sich nun im Einzelnen aus der Intransparenz eines Auswahlverfahrens? Zunächst einmal kann man davon ausgehen, dass die Kandidaten in einer Auswahlsituation positiv bewertet werden wollen. D.h. sie werden versuchen, aufgrund von diversen Hinweisreizen („demand characteristics“, z.B. Art der Instruktion, Verhalten der Interviewer, Image der Firma, vgl. Orne, 1962) subjektive Hypothesen über das richtige bzw. angemessene Verhalten aufzustellen (Bungard, 1987). Diese Hypothesen können richtig oder falsch sein (Kleinmann, 1993). Daher ist zu erwarten, dass die Kandidaten relativ viel irrelevantes Verhalten und nur wenig für die Beurteilung notwendiges Verhalten zeigen, so dass die Beobachter Schwierigkeiten bekommen, eine eindeutige Bewertung der Anforderungsdimensionen vorzunehmen.

Verschiedene Ergebnisse weisen darauf hin, dass die gezeigte und bewertete Leistung der Teilnehmer nicht nur von ihrem Können abhängt, sondern auch davon, ob sie die Anforderungskriterien richtig erkannt haben („capability to discern dimensions“, CDD, vgl. Hartstein & Kleinmann, 2002). Dieser Zusammenhang zwischen Erkennen und Leistung konnte sowohl für das AC (Kleinmann, 1991, 1993) als auch für das strukturierte Interview gezeigt werden (vgl. Lober et al., 2002; Kapitel 3). Daher ist es denkbar, dass Personen in einem intransparent durchgeführten Verfahren auf bestimmten Dimensionen schlecht abschneiden, obwohl sie eigentlich eine gute Leistung hätten zeigen können.

Zentral für die vorliegende Arbeit sind auch die folgenden Überlegungen: Wenn Bewerber eine Anforderungsdimension wie z.B. Kooperation in einer Übung als relevant identifizieren und in einer anderen Übung nicht – obwohl die Beobachter in beiden Übungen die selben Dimensionen beurteilen – dann führt dies zu einer niedrigeren konvergenten Validität, als wenn die Bewerber in beiden Übungen die Dimension „Kooperation“ erkannt haben. Diese negativen Auswirkungen der Intransparenz der Anforderungsdimensionen auf die konvergente Validität konnte ebenfalls sowohl für das AC (Kleinmann, 1991, 1993) als auch für das strukturierte Interview (Lober et al., 2002; Kapitel 3) empirisch bestätigt werden.

Anknüpfend an die oben berichteten Ergebnisse wurden inzwischen auch transparente AC (Kleinmann et al., 1996; Kleinmann, 1997b; Kolk, Born & Van der Flier, 2000; Kolk, Born & Van der Flier, in press-b) und transparente situative

Übungen (Smith-Jentsch, Salas & Brannick, 2001) untersucht. Transparenz bedeutet dabei die Bekanntgabe der Anforderungsdimensionen und der zugehörigen Verhaltensanker – beispielsweise im Rahmen einer 15minütigen Einführungsveranstaltung. Welche Folgen und Auswirkungen ergeben sich durch diese Intervention?

Zunächst einmal werden die Kandidaten durch die Vorstellung der Beurteilungskriterien auf ein ähnliches Wissensniveau gebracht (Sackett, 1987). Dies ist sinnvoll, da die unterschiedlichen Kenntnisse aufgrund mehr oder weniger systematischer Einflüsse zustande kommen können (z.B. Bücher, Freunde, Kollegen etc.), d.h. mehr oder weniger zufällig sind. Ansonsten kann es passieren, dass die Teilnehmer einer Gruppendiskussion bei der Debatte um die Einführung eines Tempolimits die Hypothese aufstellen, dass hiermit ihre politische Einstellung erfasst werden soll (Bungard, 1987).

Weiterhin ist zu vermuten, dass die Bekanntgabe der Anforderungsdimensionen den Teilnehmern ermöglicht, ihr Maximalverhalten zu zeigen, es also zu einer Leistungssteigerung der Teilnehmer kommt (vgl. Smith-Jentsch et al., 2001). Aus diesem Grund wird beispielsweise im Bereich der Pädagogischen Psychologie (Schulleistungen) schon lange eine umfassende Bekanntgabe der zu erbringenden Leistung gefordert (z.B. Gage & Berliner, 1996). Dieser Leistungsanstieg sollte gerade für das AC unproblematisch sein, da man hier ja relativ stabile und überdauernde Personenmerkmale (z.B. Durchsetzungsfähigkeit oder Teamfähigkeit) erfassen will. Bereits Sackett (1987) wies darauf hin, dass Leistungsveränderungen, die durch eine lediglich 30 Minuten dauernde Kurzberatung zustande kommen, keinesfalls als echte Varianz in den als relativ stabil angesehenen Personenmerkmalen zu interpretieren sind.

In einer Untersuchung von Kolk et al. (in press-b) konnten jedoch entgegen dieser Hypothese keine Leistungsverbesserungen beim Vergleich von transparenten und intransparenten AC festgestellt werden. Problematisch an dieser Studie ist jedoch, dass vor dem sogenannten intransparenten AC die Teilnehmer bereits mit zehn potentiellen Anforderungsdimensionen vertraut gemacht worden waren, welche ausführlich mit den entsprechenden Verhaltensankern dargestellt wurden und auch die „echten“ Dimensionen umfassten. Daher kann man in dieser Studie nicht von einer echten Intransparenz-Bedingung sprechen.

Ferner kann vermutet werden, dass die Bekanntgabe der Verhaltensanker die Darstellung dimensionsrelevanter Verhaltensweisen erleichtert (Kleinmann et al., 1995), und somit die Beobachtbarkeit der Kriterien verbessert wird. Ob sich diese verbesserte Beobachtbarkeit beispielsweise in einer Erhöhung der Beobachterübereinstimmung niederschlägt, wurde unseres Wissens bisher noch nicht untersucht.

Hingegen wurde in mehreren Studien die interne Konstruktvalidität des AC unter Transparenz untersucht. Dabei konnte in drei Studien (Kleinmann et al., 1996; Kleinmann, 1997b; Kolk et al., in press-b, Studie 2) eine Verbesserung der Konstruktvalidität festgestellt werden (Auswertung erfolgte mit Hilfe von linearen Strukturgleichungsmodellen). Lediglich in der ersten Studie von Kolk et al. (in press-b) ergab sich keine Verbesserung der internen Konstruktvalidität. Dies ist eventuell darauf zurückzuführen, dass auch die Teilnehmer im intransparenten AC mit den zehn potentiellen Anforderungsdimensionen vertraut gemacht worden waren (s.o.). Kolk et al. (in press-b) selbst weisen darauf hin, dass die Art der Übungen (Einzelübungen vs. Gruppenübungen bei Kleinmann) und die Art der Teilnehmerstichprobe (Studierende vs. echte Jobbewerber) ebenfalls zu einer Schwächung der Manipulation geführt haben könnte.

Eine Zusammenfassung der Ergebnisse zur konvergenten und diskriminanten Validität des AC unter Intransparenz bzw. Transparenz wird in Tabelle 17 dargestellt.

Tabelle 17: Studien zur konvergenten und diskriminanten Validität unter Intransparenz- und Transparenz-Bedingung

Studie	Validität	Intransparenz	Transparenz
Kleinmann, Kuptsch und Koeller (1996)	konvergent[a]	.30	.35
	diskriminant[b]	.60	.61
Kolk, Born und van der Flier (in press-b);	konvergent[a]	.31	.28
Studie 1	diskriminant[b]	.40	.49
Kolk, Born und van der Flier(in press-b);	konvergent[a]	.22	.39
Studie 2	diskriminant[b]	.52	.50

Anmerkung: Konvergent = mittlere monotrait-heteromethod Korrelation. Diskriminant = mittlere heterotrait-monomethod Korrelation.

Insgesamt scheint es so zu sein, dass die Bekanntgabe der Anforderungs-dimensionen eine (lediglich) geringe Erhöhung der konvergenten Validität zur Folge

hat, während die diskriminante Validität nahezu unverändert bleibt. Dieses Ergebnis ist konsistent mit den Erläuterungen von Kleinmann (1997b), der in seiner Argumentation ebenfalls die Auswirkungen der Transparenz auf die konvergente Validität betont.

Negative Auswirkungen der Transparenz werden von einigen Autoren bezüglich der prognostischen Validität erwartet. So vermuten beispielsweise Sackett und Dreher (1982) und auch Kleinmann (1997b), dass die Übereinstimmung der Prädiktor- und Kriteriumsvarianz beim AC nicht – wie postuliert – durch Erfassung der beabsichtigten Dimensionen gegeben ist. Vielmehr soll „ein Teil der gemeinsamen Varianz zwischen Prädiktor und Kriterium durch das gemeinsame Erkennen der relevanten Anforderungen durch [die] Teilnehmer bedingt sein“ (Kleinmann, 1997b, S. 172). Diese These erscheint plausibel, da die Kriterien für eine Beförderung in der Praxis ebenfalls von einem hohen Maß an Intransparenz gezeichnet sind (vgl. Bungard, 1987). Empirische Unterstützung erhält diese Hypothese durch zwei Studien von Kleinmann (1997b) und Smith-Jentsch et al. (2001). Diese konnten zeigen, dass die Korrelationen zwischen intransparenter Leistungssituation (AC bzw. situative Übung) und intransparent erhobenem Kriterium (AC bzw. Selbstbeurteilung) höher ausfallen, als zwischen *transparenter* Leistungssituation und intransparent erhobenem Kriterium. Kolk et al. (in press-b) sehen diese empirischen Hinweise jedoch noch nicht als überzeugend an. In ihrer eigenen Untersuchung berichten sie eine Korrelation von r = .38 zwischen verbaler Intelligenz und der im transparenten AC erhobenen Dimension „Zuverlässigkeit/Hartnäckigkeit“ („tenacity“); wohingegen der Zusammenhang im intransparenten AC mit r = .19 signifikant niedriger ausfiel.

In ihrer Metaanalyse kommen Cronshaw und Wiesner (1989) zu dem Schluss, dass nur in den unstrukturierten Interviews die komplette Varianz der Validitätskoeffizienten durch statistische Artefakte erklärt werden kann, in strukturierten Interviews hingegen nicht. Die Suche nach Moderatoren der prognostischen Validität des strukturierten Interviews erscheint also trotz der widersprüchlichen Befunde zur Manipulation „Transparenz/Intransparenz“ weiterhin sinnvoll.

Trotz der eventuell geringeren prognostischen Validität des transparent durchgeführten AC gibt es jedoch auch einige Vorteile (z.B. höhere Konstruktvalidität oder höhere Akzeptanz), die bei bestimmten Zielsetzungen, z.B. im Rahmen der Personalentwicklung, relevant sind. Dies könnte erklären, dass laut einer neueren

Erhebung immerhin fast 30% der AC transparent durchgeführt werden (Spychalski, Quinones, Gaugler & Pohley, 1997).
Hauptziel der beiden vorliegenden Studien ist es, die oben diskutierten Annahmen, die für das AC bestätigt werden konnten, nun auf ein weiteres intransparentes eignungsdiagnostischen Verfahren zu übertragen, das Multimodale Interview (MMI).

4.1.1 Hypothesen

In den beiden Studien A und B sollen folgende Hypothesen überprüft werden (vgl. Kleinmann, 1997b; Kleinmann et al., 1996; Kolk et al., in press-b):

Hypothese 1: Teilnehmer, die im Multimodalen Interview über die Anforderungsdimensionen informiert werden (transparente Bedingung), erhalten bessere Bewertungen als Teilnehmer, die keine Informationen über die bewerteten Anforderungsdimensionen erhalten (intransparente Bedingung).

Hypothese 2a: Werden die Anforderungsdimensionen den Teilnehmern nicht bekannt gegeben, wird die Varianz der Beobachterratings weitestgehend durch die Methoden (Interviewtechniken) erklärt.

Hypothese 2b: Werden die Anforderungsdimensionen bekannt gegeben, wird die Varianz der Beobachterratings sowohl durch die verschiedenen Methoden (Interviewtechniken) als auch durch die einzelnen Personenmerkmale erklärt.

Hypothese 3: Werden die Anforderungsdimensionen im strukturierten Einstellungsinterview bekannt gegeben (Transparenz-Bedingung), ergibt sich eine höhere Übereinstimmung zwischen den Beobachtern, als wenn die zu beobachtenden Anforderungsdimensionen den Teilnehmern nicht mitgeteilt werden (Intransparenz-Bedingung).

Hypothese 4: Werden die Anforderungsdimensionen im strukturierten Einstellungsinterview bekannt gegeben (Transparenz-Bedingung), so ist die Übereinstimmung zwischen den Interviewergebnissen und den Bewertungen auf einem intransparenten Kriterium geringer, als wenn die zu beobachtenden Anforderungsdimensionen den Teilnehmern nicht mitgeteilt werden (Intransparenz-Bedingung). Ein entsprechendes Kriterium (AC) wird im Rahmen von Studie B erhoben.

Zur Prüfung der Hypothesen werden zwei experimentelle Laborstudien durchgeführt, wobei unter der einen Bedingung (Intransparenz) ein übliches Multimodales Interview durchgeführt wird, während unter der jeweils anderen Bedingung (Transparenz) die Anforderungskriterien im Vorfeld des Interviews bekannt gegeben werden.
Dabei unterscheiden sich die beiden Transparenz-Bedingungen folgendermaßen voneinander: In Studie A bekommen die Teilnehmer zu Beginn des Interviews lediglich eine genaue Erläuterung der zugrunde liegenden Anforderungsdimensionen und zugehörigen Verhaltensanker. Zur Verstärkung der Intervention bekommen die Teilnehmer in Studie B zusätzlich während des Interviews vor jeder einzelnen Interviewfrage einen kurzen Hinweis darauf, welche Dimension mit der folgenden Frage erfasst werden soll.
In Studie B kann weiterhin eine Vermutung von Kolk et al. (in press-b) untersucht werden. Danach sollten Jobbewerber mit Arbeitserfahrung eher in der Lage sein, ihr Verhalten aufgrund von Hinweisreizen zu verändern. Wir erwarten daher für Teilnehmer mit Arbeitserfahrung einen stärkeren Anstieg der Leistung unter Transparenz als für Personen ohne Berufserfahrung. Die Diskussion beider Studien erfolgt gemeinsam am Ende des Kapitels.

4.2 Studie A

4.2.1 Methode

In Studie A nahmen 123 Hochschüler und Hochschulabsolventen teil, davon 59 Männer und 64 Frauen. Mehr als 39% der Teilnehmer kamen aus dem Bereich der Wirtschaftswissenschaften. Weiterhin gab es noch eine größere Gruppe von Juristen (6.5%) und Psychologen (4.1%). Die durchschnittliche Semesterzahl der Teilnehmer lag bei 9.12 Semestern (*SD* = 4.33), und das mittlere Alter betrug 27.8 Jahre (*SD* = 5.37). 52 Teilnehmer (42.2%) hatten ihr Studium bereits abgeschlossen, und acht Teilnehmer hatten schon früher einmal an einem Bewerbungstraining teilgenommen.
Die Motivation zur Teilnahme an dieser Studie bestand darin, den „Ernstfall Einstellungsinterview“ zu proben, um dann aufgrund von qualifiziertem Feedback besser auf zukünftige Bewerbungssituationen vorbereitet zu sein. Zur Erhöhung des Commitments mussten die Teilnehmer vor Beginn des Trainings eine geringe

Teilnahmegebühr überweisen. Sowohl die Teilnehmer als auch die Beobachter waren „blind“ bezüglich der experimentellen Manipulation und bezüglich der wahren Zielsetzung der Studie (Doppelblindstudie). Um dies zu erreichen wurden die Beobachter nur in einer der beiden Experimental-Bedingungen eingesetzt. Unter Umständen könnte dieses Vorgehen zu systematischen Unterschieden in den beiden Beobachtergruppen führen (z.B. eine Beobachtergruppe vergibt konsistent bessere Beurteilungen). Da die Bewertung der Konstruktvalidität auf der Basis von Korrelations-Matrizen erfolgt, wählten wir aber ein eher konservatives Vorgehen. Der Einsatz der gleichen Beobachter unter beiden Experimentalbedingungen hätte außerdem unweigerlich zur Entdeckung der experimentellen Manipulation geführt. Weiterhin wurde zunächst die Intransparenz-Bedingung (N = 64) komplett durchgeführt und erst anschließend die Transparenz-Bedingung (N = 59). Auf diese Weise sollte verhindert werden, dass die Beobachtungsdimensionen allgemein unter den potentiellen AC-Kandidaten bekannt würden. Wir nahmen an, dass dieses Vorgehen keinen nennenswerten Einfluss auf die Zusammensetzung der Stichprobe haben würde.

Die Beobachter waren fast ausschließlich Studierende der Psychologie im Hauptstudium, die mit Hilfe eines eintägigen, intensiven Trainings auf ihre Aufgabe vorbereitet worden waren. Wie in der Praxis üblich führten dann jeweils zwei Beobachter das komplette Interview mit einem Teilnehmer durch.

4.2.1.1 Das Interview

Das in der Studie genutzte Multimodale Interview (vgl. Schuler, 1992) wurde von Borchert (2001) in Zusammenarbeit mit Praktikern aus dem Bereich der Personalauswahl entwickelt. Hierbei wurde die Zielposition „Management-Trainee“ gewählt, um Teilnehmern aus unterschiedlichen Studienfächern bzw. Bereichen eine möglichst realistische Bewerbungssituation zu ermöglichen. Entsprechend wurden in der Jobanalyse typische Alltagssituationen aus dem Arbeitsleben eines Management-Trainees bzw. Abteilungsleiters gesammelt. Diese konnten zu den vier Dimensionen Arbeitsorganisation/Planung (AP); Führungsverhalten (FV), Kooperation (KO) und Informationsverhalten (IV) zusammengefasst werden.

Im Training selbst wurden dann die drei nach Schuler und Moser (1995) zentralen Bestandteile des Multimodalen Interviews durchgeführt. Es handelt sich dabei um die Selbstpräsentation (SP, Dauer ca. 5 Minuten) sowie das Biographische und Situative Interview (BI bzw. SI, 3 Fragen je Dimension, Dauer pro Interview ca. 20 Minuten).

4.2.1.2 Experimentelle Manipulation

Die Bekanntgabe/Enthüllung der Dimensionen wurde analog zu Kleinmann et al. (1996) bzw. Kolk et al. (in press-b) durchgeführt. Die Teilnehmer der Transparenz-Bedingung erhielten zunächst ein 15-minütiges Training, in welchem die Dimensionen sowie die zugehörigen Verhaltensanker erläutert wurden. Als Merkhilfe bekamen die Teilnehmer anschließend ein Handout mit den Definitionen der Dimensionen und den Verhaltensbeispielen.

Die Teilnehmer der Intransparenz-Bedingung erhielten kein solches Training. Stattdessen erhielten sie Informationen über „irrelevante“ Themen, wie Stellensuche im Internet und Aufbau von Bewerbungsunterlagen.

4.2.2 Ergebnisse

4.2.2.1 Demographische Unterschiede

Zunächst wurde mit Hilfe von t-Tests und χ^2-Methoden (vgl. Bortz, 1999) geprüft, ob es in den beiden Experimentalgruppen Unterschiede hinsichtlich der demographischen Variablen gibt. Dabei zeigten sich signifikante Unterschiede in den Variablen „Alter“ (t = -2.61; p < .01) und „Anzahl der Semester“ (t = -3.86; p < .001), die beide unter der Transparenz-Bedingung höher waren. Keine Unterschiede ergaben sich jedoch hinsichtlich der Variablen „Geschlecht“ (χ^2 = 0.38; p = .54), „Studium bereits abgeschlossen“ (χ^2 = 0.08; p = .78) und „Vorerfahrung mit Bewerbungstrainings“ (t = .26; p = .61). Anders als erwartet scheint die sukzessive Erhebung der Daten somit zu einer unterschiedlichen Zusammensetzung der beiden Stichproben geführt zu haben. Die gefundenen demographischen Unterschiede wurden soweit möglich (vgl. Hypothese 1) bei der Auswertung kontrolliert, um eine fälschliche Interpretation unserer Daten zu vermeiden.

4.2.2.2 Realitätsnähe

Mit Hilfe verschiedener Fragen wurde die Realitätsnähe der Interviewsituation abgeschätzt. Auch hierbei ergaben sich keine Unterschiede zwischen den beiden Experimentalgruppen. Die Fragen lauteten im Einzelnen: 1. „Haben Sie die dargestellten Situationen als realistisch empfunden?“ (t = .65; p = .51); 2. „Haben Sie versucht, sich so zu verhalten, wie Sie es im Beruf auch tun würden?“ (t = -.81; p = .42); 3. „Haben Sie versucht, sich so zu verhalten, wie Sie es in einer echten Bewerbung auch tun würden?“ (t = .49; p = .63). Insbesondere für die letzten beiden

Fragen ergaben sich mit mittleren Werten von 1.69 bzw. 1.66 (5-stufige Skala mit 1 = ja, 3 = teilweise und 5 = nein) Hinweise auf die Generalisierbarkeit der gefundenen Ergebnisse.

4.2.2.3 Fragen zur experimentellen Manipulation

Um die Auswirkungen der experimentellen Manipulation abzuschätzen, wurden den Teilnehmern der Transparenz-Bedingung zwei weitere Fragen (vgl. Tabelle 18, 5-stufige Antwortskala von 1 = ja bis 5 = nein) vorgelegt. Es zeigte sich, dass die Anforderungskriterien zum größten Teil verstanden worden und die meisten Teilnehmer sich an ihnen orientiert hatten. Die Ergebnisse deuten darauf hin, dass die Transparenz-Bedingung erfolgreich realisiert werden konnte.

Tabelle 18: Ergebnisse des Manipulation Checks (Studie A)

Frage/Antwort	Anzahl				
	1 (ja)	2	3 (z.T.)	4	5 (nein)
1. Haben Sie den Eindruck, dass Ihnen durch die Vorbesprechung die Anforderungsdimensionen verständlich gemacht wurden?	14	23	20	2	-
2. Haben Sie während des Interviews versucht, sich den Anforderungen der Beobachter entsprechend darzustellen?	10	29	11	7	2

4.2.2.4 Hypothese 1

In Hypothese 1 wurde wie bei Kolk et al. (in press-b) angenommen, dass die Teilnehmer der Transparenz-Bedingung bessere Bewertungen erhalten als die Teilnehmer der Intransparenz-Bedingung. Der Bereich der Leistungsbeurteilungen reichte von 1 („übertrifft die Anforderungen weit“) bis 5 („starke Veränderungen wünschenswert“). Zur Prüfung der Hypothese wurden zunächst t-Tests für unabhängige Stichproben gerechnet. Je nachdem, ob der Levene-Test zur Überprüfung der Varianzgleichheit signifikant wurde oder nicht, wurde ein t-Test für homogene bzw. heterogene Varianzen angegeben (vgl. Tabelle 19).

Wie erwartet zeigten sich sowohl auf Ebene des Gesamtinterviews als auch auf Ebene der Selbstpräsentation und des SI signifikant bessere Beurteilungen für die Teilnehmer der Transparenz-Bedingung. Lediglich im BI schnitten die beiden Gruppen in etwa gleich ab. In der Transparenz-Gruppe ergab sich trotz des Leistungsanstiegs kein Deckeneffekt. Die Varianzen der Beurteilungen waren im Gegenteil eher größer, als in der Intransparenz-Bedingung.

Tabelle 19: T-Test für unabhängige Gruppen. Vergleich der Leistungen unter Intransparenz- und Transparenz-Bedingung (Studie A)

Übung	Intransparenz (N = 64)		Transparenz (N = 59)				
	M	*SD*	*M*	*SD*	*df*	*t*-Wert	*p* (1-seitig)
SP	2.84	.74	2.30	.83	121	t_{hom} = 3.81	.00***
BI	2.37	.36	2.28	.57	95.75	t_{het} = 1.02	.16
SI	2.63	.49	2.23	.58	121	t_{hom} = 4.11	.00***
MMI	2.61	.38	2.27	.56	102.11	t_{het} = 3.94	.00***

Anmerkung: SP = Selbstpräsentation. BI = Biographisches Interview. SI = Situatives Interview. MMI = Multimodales Interview. Leistungsbewertung von 1 = übertrifft Anforderungen bis 5 = Anforderungen nicht erfüllt. t_{hom} = t-Test für homogene Varianzen. t_{het} = t-Test für heterogene Varianzen.
***p < .001 (1-seitig)

Da sich in 4.2.2.1 Unterschiede zwischen den beiden Experimentalgruppen gezeigt hatten, sollten die Ergebnisse mit Hilfe einer multiplen linearen Regression (MLR) weiter abgesichert werden. Hierzu wurde die Gesamtbewertung im MMI sowohl mit der Variablen „Versuchsbedingung“ als auch mit Hilfe verschiedener Kontrollvariablen (Alter, Anzahl der Semester, Geschlecht, Studium bereits abgeschlossen, Vorerfahrung mit Bewerbungstrainings, dargestellte Situation realistisch?, wie im Beruf verhalten?, wie in echter Bewerbung verhalten?) vorhergesagt. Es zeigte sich, dass eine MLR (fallweiser Listenausschluss fehlender Werte) mit den simultan aufgenommenen Kontrollvariablen als Prädiktor keine signifikante Vorhersage der Werte im MMI erlaubte (R^2 = .12; p = .34), wohingegen die Hinzunahme der Gruppenvariable „Experimentalbedingung“ die Vorhersage signifikant verbesserte (R = .42; ΔR^2 = .06; p < .05). Dieses Ergebnis war zu erwarten, da keine der Kontrollvariablen signifikant mit der Leistung im MMI korrelierte.

4.2.2.5 Hypothese 2

In Hypothese 2 a und b wird die Annahme überprüft, dass ein transparent durchgeführtes MMI eine höhere interne Konstruktvalidität aufweist, als ein intransparent durchgeführtes MMI. Die entsprechenden MTMM-Matrizen sind in Tabelle 20 und Tabelle 21 aufgeführt.

Tabelle 20: Multitrait-Multimethod-Matrix (Studie A, Intransparenz)

				SP				BI				SI			
		M	*SD*	AP	FV	IV	KO	AP	FV	IV	KO	AP	FV	IV	KO
SP	AP	2.54	.94	-											
	FV	3.26	1.07	.49**	-										
	IV	2.89	.93	.51**	.44**	-									
	KO	2.66	.88	.45**	.49**	.45**	-								
BI	AP	2.27	.63	.40**	.11	.27*	.24	-							
	FV	2.71	.80	.22	.40**	.09	.29*	.08	-						
	IV	2.19	.50	.18	.10	.28**	.15	.20	.25*	-					
	KO	2.30	.65	-.01	-.03	.17	.14	.04	-.16	.12	-				
SI	AP	2.66	.65	.38**	-.02	.25*	.11	.37*	-.01	.23	.16	-			
	FV	2.74	.82	.25*	-.05	.22	.15	-.11	.10	.16	.23	.32**	-		
	IV	2.64	.51	.19	.07	.06	.09	-.01	.07	.11	.12	.29*	.24	-	
	KO	2.47	.77	-.06	-.21	-.02	-.05	.01	-.15	.07	.24	.35**	.39**	.40**	-

Anmerkungen: *N* = 64. SP = Selbstpräsentation. BI = Biographisches Interview. SI = Situatives Interview. AP = Arbeitsorganisation/Planung. FV = Führungsverhalten. IV = Informationsverhalten. KO = Kooperation.

*p < .05. **p < .01 (jeweils 2-seitig).

Tabelle 21: Multitrait-Multimethod-Matrix (Studie A, Transparenz)

				SP				BI				SI			
		M	*SD*	AP	FV	IV	KO	AP	FV	IV	KO	AP	FV	IV	KO
SP	AP	2.18	1.09	-											
	FV	2.50	1.26	.54**	-										
	IV	2.34	1.00	.56**	.50**	-									
	KO	2.17	.96	.50**	.14	.42**	-								
BI	AP	1.96	.60	.17	.27*	.19	.12	-							
	FV	2.67	.96	.35**	.69**	.42**	.07	.31*	-						
	IV	2.16	.72	.18	.29*	.38**	.07	.56**	.46**	-					
	KO	2.32	.81	.32*	.27*	.33*	.20	.42**	.33*	.38**	-				
SI	AP	2.20	.72	.29*	.39**	.22	.22	.50**	.43**	.44**	.48**	-			
	FV	2.17	.80	.13	.38**	.28*	-.05	.47**	.52**	.56**	.35**	.52**	-		
	IV	2.30	.66	.44**	.41**	.35**	.23	.41**	.43**	.51**	.51**	.50**	.50**	-	
	KO	2.24	.78	.14	.29*	.37**	.27*	.43**	.29*	.45**	.34**	.50**	.48**	.43**	-

Anmerkungen: *N* = 59. SP = Selbstpräsentation. BI = Biographisches Interview. SI = Situatives Interview. AP = Arbeitsorganisation/Planung. FV = Führungsverhalten. IV = Informationsverhalten. KO = Kooperation.

*p < .05. **p < .01 (jeweils 2-seitig).

Um die Höhe der konvergenten und diskriminanten Validität zu bestimmen, berechneten wir die mittleren monotrait-heteromethod-Korrelationen (konvergente Validität) und die mittleren heterotrait-monomethod-Korrelationen (diskriminante Validität) sowie die mittleren heterotrait-heteromethod-Korrelationen. Hierbei wurde beachtet, dass die Korrelationen aus Tabelle 20 und 21 vor der Mittelwertsbildung mit Hilfe der Fischer-Z-Transformation umgewandelt werden müssen (vgl. Bortz, 1999) und anschließend zurücktransformiert werden. Es zeigte sich ein Anstieg der konvergenten Validität von .20 in der Intransparenz-Bedingung auf .39 in der Transparenz-Bedingung. Die diskriminante Validität hingegen verschlechterte sich von .30 unter Intransparenz auf .46 unter Transparenz (vgl. Tabelle 22). Auch die mittlere heterotrait-heteromethod-Korrelation war in der Intransparenz-Bedingung besser als unter Transparenz (.10 vs. .32).

Tabelle 22: Mittlere konvergente und diskriminante Validität unter Intransparenz- und Transparenz-Bedingung (Studie A)

	Intransparenz (*N* = 64)	Transparenz (*N* = 59)
mthm-Korrelationen (Dimensionen)		
Arbeitsorganisation/Planung (AP)	.38	.33
Führungsverhalten (FV)	.16	.55
Informationsverhalten (IV)	.15	.41
Kooperation (KO)	.11	.27
Mittelwert (konvergente Validität)	.20	.39
htmm-Korrelationen (Übungen)		
Selbstpräsentation (SP)	.47	.45
Biographisches Interview (BI)	.09	.41
Situatives Interview (SI)	.33	.50
Mittelwert (diskriminante Validität)	.30	.46
Mittelwert hthm-Korrelationen	.10	.32

Anmerkungen: Mthm = monotrait-heteromethod. Htmm = heterotrait-monomethod.

Um die Auswirkungen der experimentellen Manipulation genauer zu untersuchen, wurden die Personen der Transparenz-Bedingung anhand ihrer Angaben im Manipulation Check in drei Gruppen aufgeteilt. Annahme war, dass die interne

Konstruktvalidität des MMI für die Teilnehmer besonders hoch ausfallen sollte, welche sich in besonderem Maße entsprechend der experimentellen Manipulation verhalten hatten. In Tabelle 23 sind die mittleren konvergenten und diskriminanten Korrelationen der drei Gruppen angegeben.

Tabelle 23: „Orientierung an den Dimensionen“ (Manipulation Check) und Auswirkungen auf die Konstruktvalidität im transparenten Interview

	Gruppe 1	Gruppe 2	Gruppe 3
konvergente Validität[a]	.59	.33	.25
diskriminante Validität[b]	.54	.48	.25
heterotrait-heteromethod Korrelationen	.44	.31	.16

Anmerkungen: Gruppe 1 = Dimensionen waren während der Übung bewusst und auch danach verhalten (*N* = 16). Gruppe 2 = Dimensionen teilweise bewusst und teilweise danach verhalten (*N* = 29). Gruppe 3 = Dimensionen nicht bewusst und auch nicht danach verhalten (*N* = 14).
[a]mittlere monotrait-heteromethod Korrelation. [b]mittlere heterotrait-monomethod Korrelation.

Wie erwartet fand sich in Gruppe 1 (Dimensionen bewusst und auch danach verhalten) die höchste mittlere konvergente Validität, welche dann für die Gruppen 2 und 3 geringer ausfiel. Nicht erwartet hatten wir, dass in Gruppe 1 auch die höchsten heterotrait-monomethod-Korrelationen (diskriminante Validität) und heterotrait-heteromethod-Korrelationen zu finden waren, die dann ebenfalls für die Gruppen 2 und 3 kleiner wurden.

Nach dieser informellen Auswertung der MTMM-Matrizen, wurde die interne Konstruktvalidität der beiden Datensätze als Ganzes überprüft. Dies geschieht in der AC-Forschung üblicherweise mit Hilfe von LISREL-Analysen wobei im traditionellen Vorgehen die Berechnung dreier hierarchisch aufgebauter Modelle erfolgt (vgl. Byrne, 1994). Dabei ist Modell 1 das sparsamste Modell, das für die geringste Konstruktvalidität steht. Hier wird angenommen, dass das gefundene Datenmuster allein durch Methodenfaktoren (Übungen) erklärt werden kann. Im Modell 2 wird dann ein zusätzlicher Faktor („Allgemeiner Trait Faktor“) und in Modell 3 insgesamt drei weitere (Dimensions-)Faktoren eingeführt. Durch Vergleich von Modell 3 und 1 lässt sich die konvergente Validität der Daten abschätzen, durch Vergleich von Modell 3 mit 2 die diskriminante.

Da der traditionelle Ansatz meist wegen Schätzproblemen nicht durchführbar ist, wurde von verschiedenen Autoren der sogenannte CU Ansatz (correlated

uniqueness approach) vorgeschlagen (vgl. Becker & Cote, 1994; Conway, 1996; Kenny & Kashy, 1992; Marsh, 1989). In einem an Byrne (1994) angelehnten Vorgehen werden die drei hierarchisch aufgebauten Modelle 1', 2' und 3' getestet, die anstatt dreier korrelierter Methodenfaktoren korrelierte Fehlervarianzen besitzen. Ein potentieller Nachteil des CU-Ansatzes Verfahren ist nach Lance et al. (2002), dass die angenommenen Nullkorrelationen zwischen den Methoden zu einer artifiziell erhöhten konvergenten Validität führen, während die diskriminante Validität künstlich erniedrigt wird. Daher berichten wir wie andere Autoren (z.B. Kolk et al., in press-b) sowohl die Ergebnisse der traditionellen Modelle als auch die Ergebnisse der CU-Modelle und führen bei den Modellen mit zulässigen Lösungen weitere Analysen durch. Die Ergebnisse unserer Modell-Analysen finden sich in Tabelle 24.

Tabelle 24: Goodness-of-Fit Statistiken aus LISREL-Analysen für verschiedene faktorenanalytische Modelle (Studie A)

Modellbezeichnung und Erläuterung		*df*	χ^2		*p*		*RMSEA*		*CFI*		Zulässig?	
			IT	T	IT	T	IT	T	IT	T	IT	T
traditionelle CFA												
3	4 korrelierende Dimensionen 3 korrelierende Methoden	33	23.48	28.71	.89	.68	.00	.00	1.00	1.00	a, b, d	a, b, c, d, e
2	1 Dimension, 3 korrelierende Methoden	39	35.08	40.31	.65	.41	.00	.02	1.00	.99	b	ja
1	3 korrelierende Methoden	51	71.88	76.94	.03	.01	.08	.09	.80	.88	ja	ja
CU-Ansatz												
3'	4 korrelierende Dimensionen, 4 korrelierende Fehler (CU)	30	30.01	37.09	.47	.18	.00	.06	.97	.96	a, b, d	ja
2'	1 Dimension, 4 korrelierende Fehler (CU)	36	45.85	52.13	.13	.04	.07	.09	.90	.92	ja	ja
1'	4 korrelierende Fehler (CU)	48	80.84	116.25	.00	.00	.10	.16	.70	.67	ja	ja

Anmerkungen: IT = Intransparent (*N* = 64). T = Transparent (*N* = 59). CFA = konfirmatorische Faktorenanalyse. CU = correlated uniqueness. *RMSEA* = Root Mean Square Error of Approximation. *CFI* = Comparative Fit Index. Zulässig? = sind alle geschätzten Modellparameter innerhalb zulässiger Bereiche.

[a]Konvergiert nicht. [b]Fehler-Varianz ist negativ. [c]Phi ist nicht positiv definit. [d]Theta-Delta ist nicht positiv definit. [e]Korrelation zwischen latenten Dimensions-Faktoren ist größer als 1.

Wie erwartet, ergeben sich bei Anwendung der traditionellen konfirmatorischen Faktorenanalyse Konvergenzprobleme für Modell 3 – und zwar sowohl für die Intransparenz-Bedingung als auch die Transparenz-Bedingung. Dieses

Konvergenzproblem blieb für die Daten der Intransparenz-Bedingung auch bei Nutzung des CU-Ansatzes bestehen, so dass ein Modellvergleich (χ^2-Test) zur Abschätzung von konvergenter und diskriminanter Validität lediglich für die Transparenz-Bedingung durchgeführt werden konnte (vgl. Tabelle 25).

Tabelle 25: Modellvergleiche (χ^2-Differenzen-Test) zur Abschätzung von konvergenter und diskriminanter Validität (Studie A, Transparenz)

Test auf	Modellvergleich	$\Delta\chi^2$	Δdf	$\Delta\chi^2_{krit.05}$	$\Delta\chi^2_{krit.01}$
diskriminante Validität	3' vs. 2'	15.04	6	12.59	16.81
konvergente Validität	3' vs. 1'	79.14	12	21.02	26.22

Anmerkungen: N = 59. $\Delta\chi^2$ = Differenz in χ^2. Δdf = Differenz der Freiheitsgrade. $\Delta\chi^2_{krit.05}$ = kritische χ^2-Differenz auf 5% Niveau bei entsprechenden Freiheitsgraden. 3' = Modell mit 3 korrelierenden Dimensionen und 3 korrelierenden Fehlern. 2' = Modell mit einer Dimension und 3 korrelierenden Fehlern. 1' = Modell mit 3 korrelierenden Fehlern.

Sowohl Modellvergleich 3' mit 2' als auch 3' mit 1' wurden in der Transparenz-Bedingung signifikant ($p < .05$ bzw. $p < .01$), was als Beleg für die diskriminante bzw. konvergente Validität des Interviews gewertet werden kann. Da weiterhin Modell 3' einen ausreichend guten Fit[5] besitzt, kann man davon sprechen, dass ein transparent durchgeführtes strukturiertes Interview über ein gewisses Maß an interner Konstruktvalidität verfügt. Da die Fallzahlen der LISREL-Analysen relativ gering sind, ist die Ergebnis jedoch mit der notwendigen Vorsicht zu interpretieren.

4.2.2.6 Hypothese 3

In Hypothese 3 wird die Annahme überprüft, dass unter Transparenz vermehrt bewertungsrelevantes Verhalten gezeigt wird, so dass die Beobachterurteile in einem höheren Ausmaß übereinstimmen, als unter Intransparenz. Zur Prüfung der Hypothese wurden zunächst die Beobachterübereinstimmungen (Pearson-Korrelationen) aller Einzelbewertungen berechnet. Hierbei zeigten sich entgegen unserer Erwartung keine systematisch höheren Beobachterübereinstimmungen in der Transparenz-Bedingung (vgl. Tabelle 26).

[5] Übliche Kriterien zur Beurteilung der Fit-Indices sind *CFI* größer gleich .95 und *RMSEA* kleiner gleich .05; manche Autoren sehen aber auch Werte kleiner gleich .10 als ausreichend an.

Tabelle 26: Vergleich der Beurteilerübereinstimmung unter Intransparenz- und Transparenz-Bedingung (Studie A)

	Intransparenz (*N* = 64)	Transparenz (*N* = 59)	z-Wert
Selbstpräsentation (SP)			
Systematisches Denken & Handeln	.77	.80	.31
Zusammenarbeit	.70	.66	-.35
Steuerung sozialer Prozesse	.78	.83	.81
Biographisches Interview (BI)			
1. Frage	.68	.58	-.91
2. Frage	.94	.95	.62
3. Frage	.77	.86	1.51 †
4. Frage	.85	.64	-2.68 **
5. Frage	.83	.83	-.04
6. Frage	.73	.78	.65
7. Frage	.83	.64	-2.22 *
8. Frage	.64	.59	-.45
9. Frage	.78	.84	.91
10. Frage	.84	.77	-1.11
11. Frage	.72	.78	.69
12. Frage	.87	.94	2.19 *
Situatives Interview (SI)			
1. Frage	.86	.79	-1.24
2. Frage	.55	.70	1.37 †
3. Frage	.61	.71	.90
4. Frage	.87	.85	-.49
5. Frage	.83	.72	-1.40 †
6. Frage	.77	.73	-.49
7. Frage	.84	.82	-.22
8. Frage	.80	.79	-.24
9. Frage	.75	.68	-.86
10. Frage	.87	.58	-3.60 **
11. Frage	.86	.79	-1.23
12. Frage	.85	.67	-2.49 **

Anmerkungen: †p < .10. *p < .05. ** p < .01 (jeweils 1-seitig)

Zusätzlich zu den verhaltensverankerten Beurteilungen nach jeder Frage sollten die Beobachter am Ende jeder Interviewkomponenten eine Gesamtbewertung der Teilnehmer vornehmen. Hierbei wurden sie ausdrücklich dazu aufgefordert, bei der Beurteilung auch erfolgsrelevante Verhaltensweisen zu beachten, welche nicht explizit durch die drei Dimensionen SDH, ZU und SSP abgedeckt wurden (z.B. rhetorisches Geschick). Entsprechend unserer Hypothese ergaben sich unter Transparenz-Bedingung deutlich höhere Übereinstimmungs-Werte, als unter Intransparenz (Tabelle 27).

Tabelle 27: Vergleich der Beurteilerübereinstimmung in der Gesamtbeurteilung unter Intransparenz- und Transparenz-Bedingung

	Intransparenz (N = 64)	Transparenz (N = 59)	z-Wert
Selbstpräsentation (SP)	.51	.78	2.54**
Biographisches Interview (BI)	.51	.70	1.63^{+}
Situatives Interview (SI)	.33	.78	3.78**

Anmerkungen: $^{+}p < .10$. $^{**}p < .01$ (jeweils 1-seitig).

Sowohl für die Selbstpräsentation als auch für das SI wurden diese Unterschiede signifikant (1-seitige Testung, $p < .01$). Für das BI ergab sich ein tendenzieller Unterschied (1-seitige Testung, $p < .10$).

Die Diskussion von Studie A erfolgt im Anschluss an die Darstellung der Studie B mit dieser zusammen.

4.3 Studie B

4.3.1 Methode

4.3.1.1 Übersicht

Studie B wurde ähnlich wie Studie A durchgeführt; es gab jedoch zwei wichtige Unterschiede: Neben der Verstärkung der experimentellen Manipulation (s.u.) sollten ferner die Auswirkungen der Transparenz im Prädiktor MMI auf die Kriteriumsvalidität überprüft werden (Hypothese 4), was die zusätzliche Erhebung eines Kriteriums notwendig machte. Dieses Kriterium hat verschiedene Kennzeichen zu erfüllen (u.a. Intransparenz der Anforderungsdimensionen, unterschiedliche Beurteiler in Prädiktor und Kriterium, Minimierung der „Ausfälle“ durch Teilnehmerschwund, Praktikabilität etc.), die natürlich nicht alle im Rahmen einer einzigen Studie zu erfüllen sind. Nach ausführlicher Diskussion der verschiedenen Möglichkeiten entschlossen wir uns analog zu Kleinmann (1997b), ein intransparentes AC als Kriterium im Vorfeld des Interviews durchzuführen. Um eine künstliche Konfundierung der Bewertungen in Prädiktor und Kriterium zu vermeiden (gemeinsame Beobachtervarianz, vgl. Kapitel 2), fand ein Wechsel der Beobachter zwischen AC-Übungen und MMI statt.

Ansonsten war Studie B ebenfalls in ein Bewerbungstraining eingebettet, das mit Hilfe des Internets bundesweit angeboten wurde. Insgesamt nahmen 178 Personen teil (99 Frauen und 79 Männer). Die meisten Teilnehmer waren Wirtschaftswissenschaftler (30.3%), gefolgt von Psychologen (15.7%) und Juristen (6.7%). Im Mittel hatten die Teilnehmer 9.33 Semester studiert (*SD* = 3.47) und waren 27.53 Jahre alt (*SD* = 4.67). 93 Personen (52.3%) verfügten über Berufserfahrung. Diese betrug im Mittel 5.09 Jahre (*SD* = 5.10) mit einem Range von 0.5 bis 22 Jahren. Entsprechend verfügten viele Teilnehmer (*N* = 139) über Erfahrung mit realen Bewerbungssituationen, wobei hier die Erfahrung mit Einstellungsinterviews überwog (*N* = 126) und Erfahrungen mit AC oder weiteren Bewerbungssituationen eher selten vorkamen (*N* = 14 bzw. 28; Mehrfachnennungen möglich). An Bewerbungstrainings hatten schon 74 Personen (41.6%) teilgenommen, wobei die Interviewtrainings überwogen (*N* = 58).

Die Motivation zur Teilnahme an dieser Studie bestand darin, einen umfassenden Auswahltag – bestehend aus AC, Einstellungsinterview und diversen Testverfahren – zu erleben. Eine ausführliche Beschreibung der verwendeten Übungen und Interviewkomponenten findet sich in Kapitel 2. Zum Abschluss des Trainings erhielten die Teilnehmer ein individuelles Einzelfeedback, um so besser auf zukünftige Bewerbungssituationen vorbereitet zu sein. Analog zur Studie A wurde zur Erhöhung des Commitments eine geringe Teilnahmegebühr erhoben.

Studie B fand ebenfalls als Doppelblindstudie statt. Wiederum wurden die Teilnehmer zufällig der Intransparenz-Gruppe (*N* = 110, komplette Daten von 108 Personen, vgl. Kapitel 2) bzw. der Transparenz-Bedingung (*N* = 68) zugewiesen. Die eingesetzten Beobachter (meist Studierende der Psychologie) kamen jeweils nur in einer der beiden Experimental-Bedingungen zum Einsatz und waren mit Hilfe eines eintägigen, intensiven Frame-of-Reference-Trainings (vgl. Arthur et al., 2000) auf ihre Aufgabe vorbereitet worden.

4.3.1.2 Experimentelle Manipulation

Die Bekanntgabe der Dimensionen wurde zunächst analog Kleinmann et al. (1996) und Kolk et al. (in press-b) durchgeführt (siehe Studie A). So erhielten auch in Studie B die Teilnehmer vor Beginn des MMI ein Merkblatt sowie ein 15-minütiges Training, in welchem die Dimensionen und die zugehörigen Verhaltensanker erläutert wurden. In Studie A bekamen die Teilnehmer nach diesem Kurztraining und dem Merkblatt keine weiteren Informationen über die Dimensionen. Dies hatte zur Folge, dass die

Probanden nicht genau wussten, welche Anforderungsdimension mit der jeweils gerade gestellten Interviewfrage erfasst werden sollte. Die Interviews blieben sozusagen „halbtransparent“. In Studie B wollten wir die experimentelle Manipulation verstärken. Daher bekamen die Teilnehmer nun zusätzlich vor jeder einzelnen Frage einen kurzen Hinweis auf die zu erfassende Dimension.

4.3.2 Ergebnisse

4.3.2.1 Demographische Unterschiede

Wiederum wurden die beiden Experimentalgruppen auf Unterschiede hinsichtlich der demographischen Variablen untersucht. Dabei ergaben sich die Intransparenz-Bedingung tendenziell höhere Werte in den Variablen „Alter“ (t_{het} = -1.84; p = .07), „Anzahl der absolvierten Semester“ (t_{hom} = 1.75; p = .08) und bezüglich des Anteils der Personen mit Berufserfahrung (χ^2 = 2.86; p = .09). In der Transparenz-Bedingung hingegen gab es mehr Personen mit realer Bewerbungserfahrung (χ^2 = 8.68; p < .01) und auch der Anteil an Frauen war hier höher (χ^2 = 6.45; p < .05). Dafür hatten in der Transparenz-Bedingung signifikant weniger Personen bereits Erfahrungen im Rahmen von Bewerbungstrainings gesammelt (χ^2 = 25.93; p = .001). Falls möglich wurden die Unterschiede zwischen den beiden Experimentalbedingungen bei der Auswertung der Hypothesen beachtet.

4.3.2.2 Realitätsnähe

Mit Hilfe verschiedener Fragen (vierstufiges Antwortformat mit 1 = ja, 2 = eher ja, 3 = eher nein, 4 = nein) wurde auch in Studie B die Realitätsnähe des simulierten Auswahltages abgeschätzt. Der Frage „Haben Sie das AC-Bewerbungstraining als realistisch wahrgenommen?“ stimmten 92.5% der Teilnehmer zu. Auch der zweiten Frage („Haben Sie sich im heutigen AC-Bewerbungstraining so verhalten, wie Sie es bei einer echten Bewerbung tun würden?“) stimmte ein Anteil von 85.9% der Teilnehmer zu. Auch der letzten Frage („Haben Sie sich gut in die Rolle eines Bewerbers hineinversetzten können?“) stimmten immerhin 81.9% der Teilnehmer zu. Lediglich bei dieser Frage gab es signifikante Unterschiede zwischen den beiden Experimentalbedingungen: Die Teilnehmer der intransparenten Gruppe hatten sich nach eigenen Angaben besser in die Bewerberrolle hineinversetzt (t_{hom} = -2.03; p < .05). Insgesamt ergeben sich starke Hinweise auf die Realitätsnähe unserer

Bewerbungstrainings. Die nachfolgend berichteten Ergebnisse sollten daher auf Bewerbungssituationen mit echten Bewerbern übertragbar sein.

4.3.2.3 Fragen zur experimentellen Manipulation

Um die Auswirkungen der experimentellen Manipulation abzuschätzen, wurden den Teilnehmern in der Transparenz-Bedingung zwei Fragen vorgelegt, die auf einer 4-stufigen Skala von 1 (ja) bis 4 (nein) beantwortet werden konnten. Hierbei zeigte sich, dass die Anforderungsdimensionen in der Vorbesprechung zum größten Teil verständlich gemacht werden konnten und sich auch die meisten Teilnehmer zumindest teilweise daran orientiert hatten (vgl. Frage 1 und 2 in Tabelle 28). In der Intransparenz-Bedingung bekamen die Teilnehmer am Ende des Trainings eine Liste mit acht potentiellen Dimensionen vorgelegt. Sie sollten dann die im Rahmen des Trainings erhobenen Anforderungsdimensionen markieren. Nur ein einziger Teilnehmer markierte hierbei die richtigen Dimensionen, so dass die Intransparenz der Bewertungskriterien als gegeben angenommen werden kann.

Tabelle 28: Fragen zur experimentellen Manipulation (Studie B)

Frage/Antwort	Anzahl			
	1 (ja)	2 (eher ja)	3 (eher nein)	4 (nein)
1. Haben Sie den Eindruck, dass Ihnen durch die Vorbesprechung die Anforderungsdimensionen verständlich gemacht wurden?	45	21	2	-
2. Haben Sie versucht, sich nach den bekannt gegebenen Anforderungsdimensionen zu verhalten?	13	46	7	2

4.3.2.4 Hypothese 1

Zur Prüfung von Hypothese 1 (Leistungsverbesserung unter Transparenz) wurden zunächst t-Tests für unabhängige Stichproben gerechnet (vgl. Tabelle 29).

Tabelle 29: T-Test für unabhängige Gruppen. Vergleich der Leistungen unter Intransparenz- und Transparenz-Bedingung (Studie B)

Übung	Intransparenz (N = 108)		Transparenz (N = 68)			
	M	*SD*	*M*	*SD*	*df*	*t*-Wert
Selbstpräsentation	2.76	.77	2.31	.77	174	t_{hom} = 3.77**
Biographisches Interview	2.73	.54	2.46	.48	174	t_{hom} = 3.34**
Situatives Interview	2.63	.48	2.40	.42	174	t_{hom} = 3.28**
Multimodales Interview	2.71	.46	2.39	.43	174	t_{hom} = 4.58**

Anmerkung: Beurteilung der Leistung von 1 = erfüllt die Anforderungen vollständig bis 5 = starke Veränderungen wünschenswert. t_{hom} = t-Test für homogene Varianzen.

**p < .01 (1-seitig)

Wie erwartet ergaben sich für alle drei Interviewkomponenten und für das Gesamtinterview bessere Beurteilungen unter Transparenz. Zur weiteren Absicherung der Ergebnisse wurde, wie auch schon in Studie A, eine multiple lineare Regression (MLR) gerechnet. Wiederum zeigte sich, dass eine MLR, mit den insgesamt 13 verschiedenen simultan aufgenommenen Kontrollvariablen[6] als Prädiktoren, keine signifikante Vorhersage der Werte im MMI erlaubte (R^2 = .15, p = .93), während die Hinzunahme der Gruppenvariable „Experimentalbedingung“ die Vorhersage signifikant verbesserte (R = .53, ΔR^2 = .12; p < .05).

Die Daten in Studie B erlaubten auch die Testung einer weiteren Annahme, die sich auf die unterschiedliche Leistungsfähigkeit von Teilnehmern mit bzw. ohne Arbeitserfahrung bezieht. Kolk et al. (in press-b) erwarten, dass Bewerber mit Arbeitserfahrung im Vergleich zu berufsunerfahrenen Bewerbern eher in der Lage sein sollten, ihr Verhalten aufgrund von Hinweisreizen zu verändern. Dies müsste sich in einem überproportionalen Leistungsanstieg joberfahrener Bewerber unter Transparenz niederschlagen. Zur Testung dieser Annahme wurde eine zweifaktorielle Varianzanalyse mit den Faktoren Berufserfahrung (ja vs. nein) und Experimentalbedingung (Transparenz vs. Intransparenz) durchgeführt.

[6] Alter, Geschlecht, Studienfach, Anzahl Semester, erreichter Abschluss, Studium beendet?, Note im Abschluss, Berufserfahrung? Länge Berufserfahrung?, in Bewerberrolle hineinversetzt?, als Bewerber gefühlt?, wie in echter Bewerbung verhalten?, erfolgreich verhalten?

Tabelle 30: Mittelwert im Multimodalen Interview über alle Dimensionen - Deskriptive Statistiken einer zweifaktoriellen Varianzanalyse mit den Faktoren Berufserfahrung und Experimentalbedingung

Versuchsbedingung	Berufserfahrung?	*M*	*SD*	*N*
Intransparenz	ja	2.59	.44	51
	nein	2.81	.45	57
	Gesamt	2.71	.46	108
Transparenz	ja	2.35	.45	41
	nein	2.46	.41	27
	Gesamt	2.39	.43	68
Gesamt	ja	2.48	.46	92
	nein	2.70	.47	84
	Gesamt	2.59	.47	176

Anmerkung: Leistungsbeurteilung von 1 = erfüllt die Anforderungen vollständig bis 5 = starke Veränderungen wünschenswert

Wie erwartet wurden sowohl der Haupteffekt „Berufserfahrung“ ($F = 5.84$; $p < .05$) als auch der Haupteffekt „Experimentalbedingung“ ($F = 18.72$; $p < .001$) signifikant. Es ergab sich jedoch keine signifikante Wechselwirkung ($F = 0.55$; $p = .46$). Die Annahme, dass berufserfahrene Bewerber die Hinweise in der Transparenz-Bedingung in einem stärkeren Ausmaß zur Verbesserung ihres Verhaltens nutzen können, kann damit aufgrund unserer Daten nicht gestützt werden.

4.3.2.5 Hypothese 2

In Hypothese 2 wurde mit Hilfe von LISREL-Analysen die Annahme überprüft, dass die interne Konstruktvalidität eines transparent durchgeführten MMI gegenüber dem traditionell durchgeführten intransparenten MMI ansteigt. Die zugehörigen MTMM-Matrizen sind in Tabelle 31 und Tabelle 32 aufgeführt.

Tabelle 31: Multitrait-Multimethod-Matrix (Studie B, Intransparenz)

				SP			BI			SI		
		M	*SD*	SDH	ZU	SSP	SDH	ZU	SSP	SDH	ZU	SSP
SP	SDH	2.53	.81	-								
	ZU	2.51	.98	.46**	-							
	SSP	3.26	1.06	.52**	.50**	-						
BI	SDH	2.64	.68	.02	.24*	.11	-					
	ZU	2.73	.59	.11	.32**	.23*	.45**	-				
	SSP	2.82	.85	.29**	.44**	.41**	.28**	.37**	-			
SI	SDH	2.66	.69	.22*	.24*	.17	.31**	.33**	.38**	-		
	ZU	2.55	.65	.04	.15	-.15	.22*	.32**	.09	.27**	-	
	SSP	2.69	.57	.28**	.28**	.12	.10	.37**	.31**	.43**	.42**	-

Anmerkungen: *N* = 108. SP = Selbstpräsentation. BI = Biographisches Interview. SI = Situatives Interview. SDH = Systematisches Denken & Handeln. ZU = Zusammenarbeit. SSP = Steuerung sozialer Prozesse.

*p < .05. ** p < .01 (jeweils 2-seitig).

Tabelle 32: Multitrait-Multimethod-Matrix (Studie B, Transparenz)

				SP			BI			SI		
		M	*SD*	SDH	ZU	SSP	SDH	ZU	SSP	SDH	ZU	SSP
SP	SDH	2.18	.87	-								
	ZU	2.01	.92	.32**	-							
	SSP	2.76	1.12	.53**	.46**	-						
BI	SDH	2.35	.61	.16	.23	.23	-					
	ZU	2.45	.61	.07	.30*	.25*	.31*	-				
	SSP	2.58	.72	.05	.37**	.42**	.25*	.41**	-			
SI	SDH	2.43	.63	.09	.31*	.21	.40**	.27*	.42**	-		
	ZU	2.25	.55	.03	.27*	-.08	.28*	.45**	.14	.28*	-	
	SSP	2.53	.56	.06	.11	.21	.29*	.19	.21	.35**	.28*	-

Anmerkungen: *N* = 68. SP = Selbstpräsentation. BI = Biographisches Interview. SI = Situatives Interview. SDH = Systematisches Denken & Handeln. ZU = Zusammenarbeit. SSP = Steuerung sozialer Prozesse.

*p < .05. ** p < .01 (jeweils 2-seitig)

In einem ersten Schritt wurden wiederum die mittleren monotrait-heteromethod-Korrelationen (konvergente Validität) und die mittleren heterotrait-monomethod-Korrelationen (diskriminante Validität) sowie die mittleren heterotrait-heteromethod-Korrelationen berechnet. Für die Transparenz-Bedingung ergaben sich sowohl für

die mittlere konvergente Validität als auch für die mittlere diskriminante leichte Verbesserungen gegenüber der Intransparenz-Bedingung. Die mittlere heterotrait-heteromethod-Korrelation blieb hingegen nahezu unverändert (vgl. Tabelle 33).
Um die Auswirkungen der experimentellen Manipulation genauer zu untersuchen, wurden die mittleren konvergenten und diskriminanten Korrelationen für die 58 Personen berechnet, welche den beiden Fragen des Manipulation Checks zugestimmt hatten. Anders als in Studie A ergaben sich hierdurch nur geringfügige Verbesserungen bezüglich konvergenter und diskriminanter Validität (siehe ebenfalls Tabelle 33).

Tabelle 33: Mittlere konvergente und diskriminante Validität unter Intransparenz- und Transparenz-Bedingung (Studie B)

	Intransparenz (*N* = 108)	Transparenz (*N* = 68)	Transparenz[a] (*N* = 58)
mthm-Korrelationen (Dimensionen)			
Systematisches Denken & Handeln (SDH)	.18	.22	.22
Zusammenarbeit (ZU)	.26	.34	.37
Steuerung sozialer Prozesse (SSP)	.28	.28	.30
Mittelwert (konvergente Validität)	.24	.28	.30
htmm-Korrelationen (Übungen)			
Selbstpräsentation (SP)	.49	.44	.45
Biographisches Interview (BI)	.37	.32	.31
Situatives Interview (SI)	.38	.30	.20
Mittelwert (diskriminante Validität)	.41	.36	.32
Mittelwert hthm-Korrelationen	.22	.21	.20

Anmerkungen: Mthm = monotrait-heteromethod. Htmm = heterotrait-monomethod.
[a]es wurden zehn Personen eliminiert, die im Manipulation Check angegeben hatten, die Erläuterung der Dimensionen nicht verstanden zu haben bzw. sich nicht an den Dimensionen orientiert zu haben.

Nach dieser informellen Auswertung der MTMM-Matrizen wurde in einem zweiten Schritt die Konstruktvalidität des MMI in den beiden Bedingungen mit Hilfe einer Serie von konfirmatorischen Faktorenanalysen als Ganzes geprüft. Wie auch schon in Studie A wurden sowohl im Rahmen des klassischen Ansatzes als auch im Rahmen des CU-Ansatzes jeweils drei hierarchische Modelle berechnet (vgl. Kapitel 2). Anschließend konnte für die konvergierenden Modelle mit Hilfe von Modellvergleichen (χ^2-Differenzen-Test) das Ausmaß an konvergenter und

diskriminanter Validität abgeschätzt werden. Die Ergebnisse dieser Analysen finden sich in Tabelle 34 und Tabelle 35.

Tabelle 34: Goodness-of-Fit Statistiken aus LISREL-Analysen für verschiedene faktorenanalytische Modelle (Studie B)

Modellbezeichnung und Erläuterung		*df*	χ^2		*p*		*RMSEA*		*CFI*		Zulässig?	
			IT	T	IT	T	IT	T	IT	T	IT	T
traditionelle CFA												
3	4 korrelierende Dimensionen, 3 korrelierende Methoden	12	20.39	10.99	.06	.53	.08	.00	.96	1.00	[b, e]	[a, b, c, d]
2	1 Dimension, 3 korrelierende Methoden	15	21.62	20.99	.12	.14	.06	.08	.97	.94	[a, b, d]	ja
1	3 korrelierende Methoden	24	57.67	37.41	.00	.04	.12	.09	.86	.87	ja	ja
CU-Ansatz												
3’	4 korrelierende Dimensionen, 4 korrelierende Fehler (CU)	15	32.94	18.25	.00	.25	.11	.06	.91	.96	ja	ja
2’	1 Dimension, 4 korrelierende Fehler (CU)	18	35.45	28.64	.01	.05	.09	.09	.92	.89	ja	ja
1’	4 korrelierende Fehler (CU)	27	59.58	70.64	.00	.00	.13	.16	.72	.58	ja	ja

Anmerkungen: IT = Intransparent (*N* = 108). T = Transparent (*N* = 68). CFA = konfirmatorische Faktorenanalyse. CU = correlated uniqueness. *RMSEA* = Root Mean Square Error of Approximation. *CFI* = Comparative Fit Index. Zulässig? = sind alle geschätzten Modellparameter innerhalb zulässiger Bereiche.

[a]Konvergiert nicht. [b]Fehler-Varianz ist negativ. [c]Phi ist nicht positiv definit. [d]Theta-Delta ist nicht positiv definit. [e]Korrelation zwischen latenten Dimensions-Faktoren ist größer als 1.

Erwartungsgemäß ergeben sich beim traditionellen Ansatz wiederum Konvergenzprobleme. In der weiteren Auswertung werden daher nur die Modelle des CU-Ansatzes beachtet, die diesmal sämtliche konvergieren und zwar für beide Experimentalbedingungen (Tabelle 34). Hypothesenkonform ergibt sich in der Intransparenz-Bedingung der beste Modell-Fit für Modell 2’ (1 Dimension, 3 korrelierende Übungen), während in der Transparenz-Bedingung der beste Modell-Fit für Modell 1’ (3 Dimensionen, 3 korrelierende Übungen) erzielt wird.

Hypothesenkonforme Ergebnisse liefern auch die Modellvergleiche zur Abschätzung von konvergenter und diskriminanter Validität (Tabelle 35). Während unter Intransparenz lediglich der Modellvergleich 3’ mit 1’ (konvergente Validität) signifikant wird ($p < .01$), werden in der Transparenz-Bedingung sowohl der Modellvergleich 3’ mit 1’ (konvergente Validität) als auch Modellvergleich 3’ mit 2’ (diskriminante

Validität) signifikant (p < .01 bzw. *p* < .05). Weiterhin zeigt Modell 3' (drei Dimensionen) unter Transparenz einen ausreichend guten Fit (*CFI* > .95, *RMSEA* < .10, vgl. Studie A), während unter Intransparenz Modell 2' (eine Dimension) den besten Fit zeigt, der jedoch unbefriedigend ausfällt (*CFI* < .95). Insgesamt weisen die Ergebnisse darauf hin, dass die Bekanntgabe der Anforderungsdimensionen im MMI zu einer Verbesserung der internen Konstruktvalidität führt.

Tabelle 35: Modellvergleiche (χ^2-Differenzen-Test) zur Abschätzung von konvergenter und diskriminanter Validität (Studie B)

	Modellvergleich	$\Delta\chi^2$	Δdf	$\Delta\chi^2_{krit.05}$	$\Delta\chi^2_{krit.01}$
Intransparenz (*N* = 108)					
diskriminante Validität	3' vs. 2'	2.51	3	7.81	11.34
konvergente Validität	3' vs. 1'	26.64	12	21.02	26.22
Transparenz (*N* = 68)					
diskriminante Validität	3' vs. 2'	10.39	3	7.81	11.34
konvergente Validität	3' vs. 1'	52.39	12	21.02	26.22

Anmerkungen: $\Delta\chi^2$ = Differenz in χ^2. Δdf = Differenz der Freiheitsgrade. $\Delta\chi^2_{krit.05}$ = kritische χ^2-Differenz auf 5% Niveau bei entsprechenden Freiheitsgraden. 3' = Modell mit 3 korrelierenden Dimensionen und 3 korrelierenden Fehlern. 2' = Modell mit einer Dimension und 3 korrelierenden Fehlern. 1' = Modell mit 3 korrelierenden Fehlern.

4.3.2.6 Hypothese 3

Im Folgenden wurde die Annahme überprüft, dass die Beobachterurteile in der Transparenz-Bedingung in einem höheren Ausmaß übereinstimmen als in der Intransparenz-Bedingung. Hierzu wurden wiederum die Beobachterübereinstimmung aller vorgenommenen primären Bewertungen berechnet (vgl. Tabelle 36). Insgesamt wurden 30 Beurteilerübereinstimmungen miteinander verglichen. Hypothesenkonform nahmen in der Transparenz-Bedingung bei deskriptiver Betrachtung 22 Koeffizienten einen höheren Wert an. Davon waren 5 Unterschiede auf dem .05 bzw. .01 Niveau signifikant und 6 Unterschiede zeigten immerhin einen tendenziellen Effekt (*p* < .10).

Tabelle 36: Vergleich der Beurteilerübereinstimmung unter Intransparenz- und Transparenz-Bedingung (Studie B)

	Intransparenz (*N* = 108)	Transparenz (*N* = 68)	z-Wert
Selbstpräsentation (SP)			
Systematisches Denken & Handeln	.66	.84	-2.72 **
Zusammenarbeit	.72	.85	-2.19
Steuerung sozialer Prozesse	.79	.85	-1.21
Gesamtbeurteilung	.72	.82	-1.46 †
Biographisches Interview (BI)			
1. Frage	.73	.76	-0.42
2. Frage	.79	.87	-1.54 †
3. Frage	.85	.80	1.13
4. Frage	.92	.90	0.44
5. Frage	.64	.74	-1.31 †
6. Frage	.87	.81	1.36 †
7. Frage	.78	.82	-0.75
8. Frage	.75	.81	-1.03
9. Frage	.83	.80	0.69
10. Frage	.74	.87	-2.41 **
11. Frage	.73	.78	-0.77
12. Frage	.70	.78	-1.15
Gesamtbeurteilung	.66	.77	-1.43 †
Situatives Interview (SI)			
1. Frage	.82	.86	-0.95
2. Frage	.76	.80	-0.72
3. Frage	.64	.58	0.54
4. Frage	.45	.73	-2.80 **
5. Frage	.75	.77	-0.35
6. Frage	.75	.72	0.51
7. Frage	.63	.46	1.59 †
8. Frage	.76	.85	-1.66 *
9. Frage	.62	.82	-2.65 **
10. Frage	.89	.90	-0.33
11. Frage	.79	.77	0.23
12. Frage	.76	.76	-0.02
Gesamtbeurteilung	.67	.77	-1.30 †

Anmerkungen: †p < .10. *p < .05. ** p < .01 (jeweils 1-seitig)

Zur weiteren statistischen Absicherung der Hypothese wurde ein t-Test für abhängige Gruppen durchgeführt, wobei die Korrelationen nach Fischer-Z-Transformation als Datenpunkte betrachtet wurden (vgl. Silverman et al., 1986). Dieser Test wurde signifikant (t = -3.44; p < .001, 1-seitige Testung). Die Beurteilerübereinstimmung in der Intransparenz-Bedingung ist also niedriger als in der Transparenz-Bedingung.

4.3.2.7 Hypothese 4

Im Folgenden wird die Annahme geprüft, dass die korrelative Übereinstimmung zwischen strukturiertem Interview (Prädiktor) und den Einschätzungen auf einem

Kriterium (intransparentes AC) in der Transparenz-Bedingung geringer ausfällt als unter der Intransparenz-Bedingung. Im Kriterium (AC) wurden dabei zwei Gesamtwerte bestimmt: Der erste Wert basiert auf den Bewertungen der einzelnen Dimensionen, der zweite Wert basiert auf den Gesamtbeurteilungen der Übungen. Korreliert wurden diese Werte sowohl mit dem Gesamtergebnis im Multimodalen Interview (MMI) als auch mit dem Gesamtergebnis im BI und SI (vgl. Tabelle 37).

Tabelle 37: Produkt-Moment-Korrelationen zwischen Assessment Center (AC) und Multimodalem Interview (MM) für Intransparenz und Transparenz

Korrelation von	Interview Intransparent *N* = 107	Interview Transparent *N* = 66	z-Wert[a]
AC (Durchschnitt Dimensionen) mit MMI (SP, BI, SI)	.44	.37	0.56
AC (Durchschnitt Urteile Übungen) mit MMI (SP, BI, SI)	.43	.31	0.89
AC (Durchschnitt Dimensionen) mit Interview (BI & SI)	.43	.29	1.01
AC (Durchschnitt Urteile Übungen) mit Interview (BI & SI)	.41	.22	1.39^{+}

Anmerkungen: SP = Selbstpräsentation. BI = Biographisches Interview. SI = Situatives Interview.

[a]Berechnung siehe Bortz (1999); $z_{krit}(10\%) = 1.28$; $z_{krit}(5\%) = 1.65$.

$^{+}p < .10$ (1-seitig).

Hypothesenkonform fielen die Zusammenhänge im transparenten Interview deskriptiv niedriger aus, als im intransparenten Interview. Die Unterschiede wurden jedoch lediglich für die Korrelation AC (Durchschnittswert Dimensionen) mit Interview (BI plus SI) tendenziell signifikant.

Wichtig zu erwähnen ist, dass durch die experimentelle Manipulation (Transparenz) keine Einschränkung der Varianz in den Bewertungen stattfand (Prüfung mit dem Levene-Test auf Varianzgleichheit, vgl. Hypothese 1). D.h. die numerische Verringerung der Korrelationen unter Transparenz ist nicht auf ein methodisches Artefakt zurückführbar.

4.4 Diskussion

Dieses Kapitel beschäftigt sich mit den Auswirkungen der Transparenz auf das strukturierte Interview. Dabei zeigte sich, dass die Bekanntgabe der Dimensionen zu einem Leistungsanstieg der Bewerber führt. Weiterhin kam es zu einer höheren Beurteilerübereinstimmung und zu einem Anstieg der internen Konstruktvalidität des MMI unter Transparenz. Der Zusammenhang zwischen MMI und einem Kriterium (AC) sank unter Transparenz. Der Unterschied wurde jedoch nicht signifikant.

In unserer ersten Hypothese postulierten wir einen Leistungsanstieg unter Transparenz. Dieser Effekt konnte in beiden Studien statistisch abgesichert bestätigt werden. Lediglich bei einer Interviewkomponenten – dem BI in Studie A – ergab sich kein signifikanter Anstieg, was folgendermaßen erklärt werden kann. Wie in Kapitel 3.1 ausgeführt, versuchen Teilnehmer in einer Bewerbungssituation sich möglichst optimal zu präsentieren. Doch selbst wenn die Beurteilungskriterien den Teilnehmern bekannt gegeben werden (Transparenz), ist eine optimale Beantwortung der Fragen nicht ohne weiteres möglich. Dies gilt insbesondere dann, wenn – wie im BI – das zu berichtende Verhalten in der Vergangenheit stattgefunden hat bzw. es sich um prinzipiell überprüfbare Fragen mit eindeutiger „wahrer“ Antwort (z.B. „Welche Ämter in Vereinen oder sonstigen Organisationen haben Sie bereits inne gehabt?“) handelt. In diesen Fällen können „geschönte“ Antworten der Bewerber unter Umständen als Lüge entlarvt werden – mit den entsprechend negativen Konsequenzen in Bezug auf eine Einstellung. Obwohl im BI der Studie B ebenfalls ein signifikanter Leistungsanstieg gefunden wurde, scheint es uns lohnenswert die Eigenschaften biographischer Fragen, welche zu einem Leistungsanstieg in einem transparent durchgeführten strukturierten Interview führen, in zukünftigen Studien näher zu untersuchen (z.B. Fragen sind überprüfbar vs. nicht überprüfbar oder Fragen haben nur eine gültige Antwort vs. lassen die Beschreibung verschiedener Situationen zu).

Bei der Untersuchung der Leistungsveränderungen unter Transparenz in einem AC, bestehend aus Rollenspielen und einer Präsentation, konnten Kolk et al. (in press-b) im Gegensatz zur vorliegenden Untersuchung keinen Anstieg der Leistung beobachten. Wir vermuten, dass zwei Erklärungen bei der Analyse dieser unterschiedlichen Ergebnisse besonders wichtig sind: Zum einen ist zu vermuten, dass die Beurteilung im AC relativ zu den anderen Teilnehmern erfolgt, insbesondere wenn es sich um Gruppenübungen handelt. Daher sollte die mittlere Beurteilung in solchen Gruppen immer ähnlich gut ausfallen, unabhängig davon ob es sich bei

objektiver Betrachtung um eine eher gute oder eher mittelmäßige Gruppe gehandelt hat. Hingegen erfolgt die Beurteilung der Teilnehmer im MMI anhand von verhaltensverankerten Skalen und damit anhand eines absoluten und sensitiven Maßstabes, so dass sich die Leistungssteigerung der Teilnehmer unter Transparenz auch entsprechend in den Beurteilungen niederschlagen kann. Zum anderen müssen die Teilnehmer im Interview lediglich ihre verbale Aussage an die Anforderungskriterien anpassen, während in den AC-Übungen auch das entsprechende verbale und nonverbale Verhalten gezeigt werden muss. Letzteres ist mit Sicherheit deutlich schwieriger und bedarf sehr viel mehr Übung bzw. einer stärkeren Intervention. Weitere Studien, in denen der Einfluss verschiedener Interventionen (Coaching, Training) auf die Leistung untersucht wurde, stützen diese Erklärungen. So führte beispielsweise ein echtes Coaching zu einem Leistungsanstieg in einer führerlosen Gruppendiskussion, nicht jedoch eine Placebo Intervention (Kurecka et al., 1982) oder die vorherige Teilnahme an einer Diskussion (Petty, 1974). Da in diesen Studien nun die Teilnehmer aus den unterschiedlichen Interventionsbedingungen und damit mit unterschiedlichem Leistungsniveau im Rahmen einer gemeinsamen Gruppendiskussion beurteilt wurden, spielt der Effekt der relativen Beurteilung – anders als bei Kolk et al. (in press-b) – eine unbedeutende Rolle. Auch die Teilnahme an einem intensiven Verhaltenstraining führte zu besseren Leistungen bei einer Problemlöse-Diskussion (Moses & Ritchie, 1976). Weiterhin kam es schon nach einem kurzen Training zu höheren Leistungen in einer Postkorbaufgabe (Brannick, Michaels & Baker, 1989; Gill, 1982). Die Bewertung der Teilnehmer erfolgte hierbei anhand von Checklisten und somit anhand von absoluten Vergleichsmaßen. Dies erklärt, warum die relativ schwachen Interventionen entsprechende Unterschiede in der Leistungsbeurteilung zur Folge haben konnten.

Im Rahmen der Studie B wurde auch die Annahme überprüft, dass berufserfahrene Personen besonders von der Bekanntgabe der Beurteilungskriterien profitieren. So sollen nach Kolk et al. (in press-b) Berufstätige über ein entsprechend differenziertes Verhaltensrepertoire verfügen, dass bei entsprechenden Hinweisen aktiviert werden kann. Diese Annahme konnte für das MMI nicht bestätigt werden. Hierfür gibt es unterschiedliche Erklärungen: Zum einen war die Berufserfahrung unserer Teilnehmer zum Teil sehr gering und betrug in vielen Fällen nur 6 Monate. Eine Subgruppenanalyse war jedoch aufgrund der geringen Fallzahlen nicht möglich.

Weiterhin ist zu vermuten, dass der von Kolk et al. (in press-b) postulierte Effekt im Interview generell geringer ausfällt, da das zu bewertende Verhalten lediglich verbal beschrieben und nicht im Rahmen einer sozialen Interaktion gezeigt werden muss. Zudem ist es plausibel, dass der Nutzen transparenter Kriterien für unterschiedliche Dimensionen verschieden ausfällt (z.B. Verhandlungsgeschick oder Präsentationsfähigkeit vs. Systematisches Denken). Zur Klärung der Fragestellung sind deswegen weitere Untersuchungen nötig.
Ein weiteres wichtiges Ergebnis ist, dass die Varianz des transparent durchgeführten Interviews nicht sinkt, sondern eher sogar steigt. Es kommt also nicht zu einem Deckeneffekt. Damit bleibt die differenzierte Beurteilung erhalten, und Korrelationen mit anderen Maßen und Konstrukten sind weiterhin möglich.

In der zweiten Hypothese wurden die Folgen transparenter Dimensionen auf die interne Konstruktvalidität des MMI untersucht. Kolk (2001) fasst den Stand der Forschung zum AC folgendermaßen zusammen: „Transparency seems to have little effect on inexperienced participants taking part in individual exercises, yet a greater effect on inexperienced participants when they are interacting with others. Also, transparency has an effect on the performance of actual job candidates with work experience taking part in individual exercises" (S. 201). In den beiden vorliegenden Studien wurde dieser Forschungsansatz auf das MMI angewandt. Die Auswertung erfolgte mit Hilfe von MTMM-Matrizen und der konfirmatorischen Faktorenanalyse.
Aufgrund von Konvergenzproblemen konnten in Studie A nur die CU-Modelle der Transparenz-Bedingung mit Hilfe von χ^2-Tests statistisch auf konvergente und diskriminante Validität geprüft werden. Hierbei zeigte sich eine zufriedenstellende Konstruktvalidität des transparent durchgeführten Interviews. Auch bei Betrachtung der mittleren monotrait-heteromethod-Korrelationen und Aufteilung der Teilnehmer aufgrund der Werte im Manipulation Check ergab sich ein deutlicher Anstieg der konvergenten Validität unter Transparenz. Nicht im Einklang mit unserer Hypothese ist jedoch der Anstieg der heterotrait-monomethod- und heterotrait-heteromethod-Korrelationen (diskriminante Validität), für den wir bisher keine Erklärung haben. Nicht auszuschließen ist, dass es sich um Stichprobeneffekte handelt.
In Studie B konvergierten alle postulierten CU-Modelle, so dass die interne Konstruktvalidität für beide Experimentalbedingungen mit Hilfe von hierarchischen χ^2-Tests überprüft werden konnte. Hypothesenkonform erklärte ein Modell mit einer

Dimension in der Intransparenz-Bedingung die Daten am besten, während in der Transparenz-Bedingung ein Modell mit drei Dimensionen den besten Fit erzielte. Anders als in Studie A zeigten sich bei Betrachtung der mittleren monotrait-heteromethod- und heterotrait-heteromethod-Korrelationen, dass vor allem die Verbesserungen der diskriminanten Validität und weniger der konvergenten Validität zu der Erhöhung der internen Konstruktvalidität geführt hatten. Für diese unterschiedlichen Effekte in Studie A und B haben wir bisher keine befriedigende Erklärung. So hatten wir erwartet, dass die Verstärkung der experimentellen Manipulation in Studie B eher zu einem weiteren Anstieg der konvergenten Validität führen würde, was jedoch nicht beobachtet werden konnte.

Insgesamt konnte Hypothese 2 durch die beiden Studien empirisch gestützt werden. Einschränkend möchten wir jedoch darauf hinweisen, dass die Stabilität der durchgeführten LISREL-Analysen, aufgrund der großen Anzahl an Variablen und der teilweise geringen Versuchspersonenanzahl, unter einem gewissen Vorbehalt zu betrachten sind. Unsere Ergebnisse sind insofern überraschend, als dass es sich beim Interview um eine Einzelübung handelt und die Teilnehmer zum Teil Studierende waren. Diese sollten nach Kolk et al. (in press-b) über ein geringeres Verhaltensrepertoire verfügen. Hinzu kommt, dass die Datenerhebung im Rahmen eines Bewerbungstrainings stattfand, wodurch die Motivation, den Hinweisreizen zu folgen, unter Umständen eingeschränkt ist. Wir kommen trotzdem zu dem Schluss, dass die Steigerung der internen Konstruktvalidität durch Bekanntgabe der Anforderungsdimensionen ein relativ stabiler Effekt ist, der bei unterschiedlichen Verfahren (Gruppendiskussion, Rollenspiele, strukturierte Interviews) eine Rolle spielt. Die genauen Wirkmechanismen, welche zu einer Erhöhung der internen Konstruktvalidität bei einem transparent durchgeführten Auswahlverfahren führen, sind jedoch noch nicht ausreichend geklärt (vgl. auch Hypothese 3).

In der dritten Hypothese wurde die Annahme überprüft, dass die Beurteilerübereinstimmung unter Transparenz steigt. So vermuteten wir, dass nach Bekanntgabe der Anforderungskriterien mehr dimensionsrelevantes Verhalten gezeigt wird, wodurch es den Beobachtern leichter fallen sollte, die gezeigte Leistung anhand der Verhaltensanker zu beurteilen. Folglich sollten die Beobachter schnell zu einem Konsens gelangen und die Beobachterübereinstimmung ansteigen. Aus der Forschung wissen wir jedoch, dass verhaltensverankerte Beurteilungsskalen eine

relativ hohe Beobachterübereinstimmung besitzen (vgl. Conway et al., 1995). Unter diesen Umständen ist eine weitere Steigerung der hohen Beurteilerübereinstimmung – hier durch Bekanntgabe der Anforderungsdimensionen – ein sehr hochgestecktes Ziel. Es ist so gesehen nicht verwunderlich, dass in Studie A kein systematischer und signifikanter Anstieg der Beurteilerübereinstimmung für die einzelnen Interviewfragen aufgezeigt werden konnte. Es ergaben sich jedoch für die Gesamturteile in den Übungen, welche nicht mit Verhaltensankern versehen waren, signifikant höhere Übereinstimmungen in den Beurteilungen.

Ein ähnlicher Effekt zeigte sich auch in Studie B. Dort waren die Beurteilerübereinstimmungen in den Gesamturteilen der verschiedenen Interviewkomponenten jedoch nur tendenziell höher. Dafür zeigte sich hier, dass nun in den verhaltensverankerten Interviewfragen unter Transparenz eine signifikant höhere Beobachterübereinstimmung erzielt wurde als unter Intransparenz.

Die Untersuchung dieser Fragestellung an einem transparent durchgeführten AC erscheint sehr wünschenswert. Außerdem stellt sich die Frage, inwieweit ein Anstieg der Beobachterübereinstimmung auch zur Steigerung der internen Konstruktvalidität beiträgt. Bisher wurde meist ein Anstieg der konvergenten Validität bei Bekanntgabe der Anforderungsdimensionen diskutiert (vgl. Kleinmann, 1993, 1997b), wodurch es natürlich auch Effekte für die Konstruktvalidität als Ganzes geben sollte. Es ist jedoch auch denkbar, dass es unter Transparenz zu einem Anstieg der Messgenauigkeit kommt, wodurch sich konvergente und diskriminante Korrelationen weiter den „wahren“ Werten annähern. Eine genaue Analyse der Wirkmechanismen, welche zu einem Anstieg der Konstruktvalidität unter Transparenz führen, erscheint sinnvoll.

In der vierten und letzten Hypothese wurde der Zusammenhang zwischen MMI (intransparent/transparent) und einem intransparentem Kriterium (AC) untersucht. Auch diese Fragestellung sollte wichtige Hinweise auf die Frage liefern, was das strukturierte Interview misst. Ein Sinken der Korrelation unter Transparenz würde darauf hinweisen, dass durch die experimentelle Manipulation gemeinsame Varianz in Prädiktor und Kriterium verloren geht und damit Hinweise auf das im strukturierten Interview gemessene Konstrukt liefern, nämlich die Fähigkeit, Anforderungs-dimensionen zu erkennen (vgl. Kapitel 3).

Bei deskriptiver Betrachtung zeigen sich hypothesenkonform höhere Kriteriumsvaliditäten für das intransparente MMI. Die Unterschiede werden jedoch

nicht signifikant, obwohl es zum Teil fast zu einer Halbierung der Korrelationen unter Transparenz kommt. Eine wichtige Anmerkung an dieser Stelle ist, dass die geringeren Korrelationen unter Transparenz nicht auf eine Varianzeinschränkung im MMI zurückzuführen sind (vgl. Ergebnisse Hypothese 1). In Studie A kam es im Gegenteil eher zu einem Anstieg der Varianz der Beobachterurteile unter Transparenz.

Noch zu klären ist, warum die Korrelationsunterschiede in den Kriteriumsvaliditäten größer sind, wenn man nicht das gesamt MMI betrachtet, sondern lediglich die beiden strukturierten Interviews (BI und SI). Die Selbstpräsentation ist durch eine relativ geringe Beurteilerübereinstimmung (vgl. Schuler, 1989a, 1992) bei gleichzeitig hoher innerer Konsistenz (siehe Kapitel 2) gekennzeichnet, was zusammengenommen auf einen starken Halo-Effekt hinweist. Daher vermuten wir, dass unsere im Vorfeld definierten Anforderungskriterien in dieser Übung eine verhältnismäßig geringe Rolle spielen könnten. Vielmehr ziehen die unterschiedlichen Beobachter bei der Beurteilung der Teilnehmer verschiedene Kriterien heran, die sich z.T. widersprechen (z.B. Beobachter A, der einen ruhigen Vortrag schätzt, und Beobachter B, der ausgefallene Selbstvorstellungen bevorzugt). Die Leistung der Teilnehmer würde nach den Ergebnissen von Kleinmann (1993) davon abhängen, inwieweit sie diese impliziten Kriterien der Beobachter erkennen. Daher vermuten wir, das die Fähigkeit, Anforderungsdimensionen zu erkennen, auch bei einem transparent durchgeführten MMI eine gewisse Rolle spielt, und zwar vor allem in der Selbstpräsentation, wodurch sich bei Ausschluss der Selbstpräsentation größere Unterschiede in den Kriteriumskorrelationen ergeben sollten.

Da die Bekanntgabe der Dimensionen den Teilnehmern eine gezieltere Darstellung ihres Maximalverhaltens ermöglicht – während in intransparenten Übungen unter Umständen eher typisches Verhalten gezeigt wird – sei in diesem Zusammenhang auch auf die Ergebnisse von Sackett, Zedeck und Fogli (1988) verwiesen. Auch diese Autoren vermuten, dass die typische Leistung besser zur Vorhersage von Berufserfolg geeignet ist als maximale Leistung. Anders als Kleinmann (1993, 1997b) sehen sie jedoch nicht die Fähigkeit, Anforderungsdimensionen zu erkennen, als Bindeglied zwischen Prädiktor und Kriterium, sondern einen Motivationsfaktor. So ist nach Sackett et al. (1988) die normale Arbeitsleistung eine typische Leistung, die sich daher aus Können und „typischer“ Motivation zusammensetzt, während Maximalleistung aus Können und maximaler Motivation resultiert.

Eine generelle Einschränkung der Studien ist, dass die Datenerhebung im Rahmen von Bewerbungstrainings stattfand. Daher ist eine Replikation der Ergebnisse mit echten Bewerbern auf einen real ausgeschriebenen Job und einem echten Kriterium (prognostische statt konkurrente Validität) wünschenswert. Wir gehen davon aus, dass die stärkere Motivation echter Bewerber eine effektivere Manipulation ermöglichen würde, wodurch die gezeigten Effekte stärker werden könnten.
Tomás et al. (2000) weisen darauf hin, dass MTMM-Matrizen mit 3 Übungen und 3 Dimensionen besonders anfällig für Konvergenzprobleme sind. Um die Stabilität der Ergebnisse zu erhöhen sollten Folgestudien daher mindestens 4 Dimensionen untersuchen sowie mit einer größeren Stichprobe arbeiten. Da sich unsere beiden Experimentalgruppen in einigen demographischen Variablen signifikant unterschieden haben, ist es auch denkbar, in einer zukünftigen Studie zur Konstruktvalidität die Transparenz *innerhalb* der Versuchpersonen zu manipulieren. So könnten beispielsweise zwei Dimensionen bekannt gegeben werden und zwei Dimensionen intransparent bleiben.
Weiterhin ist zu wünschen, dass, wie von Kolk (2001) gefordert, in der zukünftigen Forschung zur Transparenz der Anforderungsdimension die Auswirkungen auf die Wahrnehmung der Bewerber verstärkt untersucht werden. Hierzu gehören die Wahrnehmung von Kontrolle, die Annahme von Feedback, Änderungen im emotionalen Erleben (z.B. Aufgeregtheit, Kleinmann, 1997b) sowie Aspekte der sozialen Validität (Schuler, 1993).

In den beiden durchgeführten Studien zum MMI wurden die Auswirkungen der Transparenz untersucht. Beide Studien kamen zu ähnlichen Ergebnissen, die darauf hinweisen, dass sowohl Konstruktvalidität als auch Beurteilerübereinstimmung in einem transparent durchgeführten Interview ansteigen könnten. Zusammen mit Kleinmann et al. (1996) sowie Kolk et al. (in press-b) kann man daher vermuten, dass unter Transparenz sowohl im AC als auch im strukturierten Interview die beabsichtigten Anforderungsdimensionen besser erfasst werden. Der Einsatz transparent durchgeführter Verfahren im Rahmen der Personalentwicklung erscheint somit sinnvoll. Die vorliegenden Ergebnisse weisen jedoch gleichfalls darauf hin, dass konkurrente und damit wahrscheinlich auch prognostische Validität negativ beeinflusst werden. Weitere Forschung zu diesem Fragenkomplex ist jedoch nötig, da sich wichtige Konsequenzen für die Durchführung von Auswahlverfahren ergeben könnten.

5 Gesamtdiskussion

Nachdem eine Reihe von Metaanalysen (Huffcutt & Arthur, 1994; Schmidt & Hunter, 1998; Schmidt & Rader, 1999; Wiesner & Cronshaw, 1988) die prognostische Validität strukturierter Einstellungsinterviews belegt haben, werden in der vorliegenden Arbeit erstmals umfassende Untersuchungen ihrer Konstruktvalidität mit Hilfe von MTMM-Matrizen und konfirmatorischen Faktorenanalysen durchgeführt.

Zunächst konnten wir zeigen (vgl. Kapitel 2), dass die interne Konstruktvalidität strukturierter Interviews als eher gering einzuschätzen ist, obwohl es plausible Korrelationen mit externen Verfahren und Konstrukten gibt. In weiteren Untersuchungen erwiesen sich die Fähigkeit, Anforderungsdimensionen zu erkennen (CDD, vgl. Kapitel 3) und die Bekanntgabe der Anforderungsdimensionen (Transparenz, Studie A und B in Kapitel 4) als effektive Moderatoren der internen Konstruktvalidität. Der Einflussfaktor gemeinsame Beobachtervarianz (vgl. Kapitel 2) erbrachte hingegen keine nennenswerten Effekte. Dies weist darauf hin, dass sich die Ergebnisse der AC-Forschung nicht ohne weiteres auf strukturierte Interviews übertragen lassen. Insgesamt konnten damit sowohl ein Personenmerkmal als auch eine „technische“ Manipulation, als Einflussfaktoren auf die interne Konstruktvalidität strukturierter Interviews nachgewiesen werden.

Als Quintessenz der Arbeit lässt sich somit formulieren: Die Suche nach Einflussfaktoren auf die interne Konstruktvalidität strukturierter Interviews ist – ähnlich wie beim AC – sinnvoll und erfolgsversprechend. Die gefundenen Ergebnisse stellen einen bedeutsamen Erkenntnisgewinn dar und könnten den Auftakt zu einer ganzen Reihe von Studien darstellen. Im Folgenden werden daher die generellen Beschränkungen der Studien sowie Konsequenzen für die weitere Forschung und Praxis dargestellt.

5.1 Generelle Grenzen der durchgeführten Studien

Ausführliche Diskussionen der einzelnen Studien und zugehörigen Ergebnisse finden sich in den entsprechenden Kapiteln 2 bis 5. Im Folgenden werden daher nur kurz die generellen Beschränkungen unserer Untersuchungen diskutiert, also Fragen der Generalisierbarkeit und der ökologischen Validität.

Wie oben erwähnt, erfreuen sich Einstellungsinterviews großer Akzeptanz und weiter Verbreitung (Schuler et al., 1993; Schulz et al., 1985) Doch anders als in der Wissenschaft, spielen strukturierte Interviews in der Wirtschaft immer noch eine relativ untergeordnete Rolle. Hierfür sind unterschiedliche Gründe denkbar, wie Unwissenheit, beschränkte Ressourcen usw. Wichtigster Grund war in einer Untersuchung von Terpstra und Rozell (1997), dass HR-Manager trotz zahlreicher wissenschaftlicher Belege immer noch nicht von der Nützlichkeit strukturierter Interviews überzeugt sind. Dies macht deutlich, wie wichtig es ist, sich um die Verbreitung wissenschaftlicher Erkenntnisse zu kümmern. Daher ist es sehr erfreulich, dass in einer aktuellen Studie von van der Zee et al. (2002) die Bedingungen näher untersucht werden, die eine Vorhersage darüber erlauben, ob ein Praktiker bevorzugt strukturierte oder unstrukturierte Interviews einsetzt. Ziel unserer Forschung ist hingegen die Verbesserung der psychometrischen Eigenschaften von Interviews. Daher wäre es sicherlich wenig sinnvoll, konventionelle Einstellungsgespräche zu untersuchen und somit die Standardisierung von Interviews – welche den wichtigsten Faktor zur Erhöhung der prognostischen Validität darstellt – zu vernachlässigen.

Ein weiteres Problem bei der Frage der Generalisierbarkeit unserer Ergebnisse betrifft die Auswahl der Dimensionen. So ist es wahrscheinlich, dass in anderen Interviews auch andere Dimensionen verwendet werden. Wie in der Einleitung ausgeführt, sehen wir strukturierte Interviews, ähnlich wie das AC oder Fragebogenverfahren, als eine Hülle an, „die inhaltlich mit verschiedenen Konstrukten gefüllt [...] werden kann.“ (Schuler, 1992, S. 284). Um welche Konstrukte bzw. Dimensionen es sich dabei im Einzelnen handelt, ist vor allem interessant, wenn man sich die Frage nach dem nomologischen Netzwerk des Interviews stellt. So ist plausibel, dass ein Interview zur Erfassung der sozialen Kompetenz mit anderen Verfahren korreliert, als ein Interview zur Erfassung der allgemeinen Intelligenz. In unseren Studien hingegen liegt der Fokus auf der internen Konstruktvalidität (vgl. Kolk, 2001), also der Frage nach konvergenter und diskriminanter Validität der Dimensionen innerhalb des Interviews. Wir gehen daher davon aus, das sich unsere Ergebnisse auch auf andere strukturierte Interviews übertragen lassen, sofern deren Dimensionen ähnlich sorgfältig konstruiert werden (vgl. Kapitel 2.2.2), d.h. unter Beachtung der einschlägigen Ergebnisse aus der AC-Forschung (Lievens, 1998). Genau genommen müsste diese Annahme natürlich

empirisch überprüft werden, was bisher jedoch selbst in der Forschung zur internen Konstruktvalidität des AC nicht geschehen ist.

Schwerwiegender ist unseres Erachtens die Frage, inwieweit Ergebnisse aus einem Bewerbungstraining auf reale Bewerbungssituationen mit einem realen Job übertragen werden können. Allgemein kann man sagen, dass die Leistungen der Teilnehmer durch ihre Fähigkeiten (Können), ihr Talent relevante Kriterien zu erkennen (Erkennen) und durch ihre Motivation (Wollen) beeinflusst werden. Es ist unseres Erachtens plausibel, dass sich Bewerber und Trainingsteilnehmer hinsichtlich ihrer Motivation und in der Heterogenität ihrer Voraussetzungen bzw. Fähigkeiten unterscheiden. So dürften Bewerber ein stärkeres Interesse daran haben, besonders gut in den verschiedenen Auswahlverfahren abzuschneiden. Außerdem findet durch Selbstselektion und Vorauswahl anhand von Bewerbungsunterlagen eine Homogenisierung der Bewerber statt, wodurch es zur Varianzeinschränkung der Leistungswerte und damit zu geringeren Korrelationen kommen kann. Die Ziele der Trainingsteilnehmer sind hingegen weniger eindeutig. So gibt es beispielsweise Teilnehmer, die sich „einfach ganz natürlich verhalten wollen“, oder aber andere die sich wünschen „mal was ganz Neues auszuprobieren“. Ferner findet im Bewerbungstraining keine Vorselektion der Teilnehmer durch die Organisatoren statt. Die oben genannten Faktoren der Teilnehmerleistung (Können, Erkennen, Wollen) könnten theoretisch sowohl additiv als auch multiplikativ verknüpft sein. So ist es sowohl denkbar, dass Personen ohne Erkennensfähigkeit, unabhängig vom Ausmaß der Motivation, gleich schlecht abschneiden. Andererseits ist es denkbar, dass eine höhere Motivation dazu führt, dass Hinweisreize genauer verarbeitet werden, so dass sich eine höhere Erkennensfähigkeit zeigt. Diese Gedanken machen deutlich, wie schwierig es ist vorherzusagen, in welcher Art und Weise sich die Teilnehmerunterschiede auswirken werden. Zum Glück sind wir bei dieser Frage nicht nur auf Spekulationen angewiesen, sondern können auf den Manipulation Check in Studie 3 (vgl. Kapitel 4.2.1.2 und 4.3.1.2) zurückgreifen. Dort zeigte sich, dass die Effekte in zwei Hypothesen (Leistungsanstieg und Anstieg der internen Konstruktvalidität) stärker werden, wenn die motivierteren Personen ausgewählt wurden. Das Problem der Varianzeinschränkung (Homogenisierung der Bewerber durch Vorauswahl) lässt sich hingegen relativ einfach durch Anpassung der Verhaltensanker in den verhaltensverankerten Beurteilungsskalen beseitigen.

Insgesamt gehen wir folglich davon aus, dass die gefundenen Ergebnisse auf eine Vielzahl von Bewerbungssituationen – auch reale – übertragen werden können.

5.2 Konsequenzen für die weitere Forschung

Im Verlauf der Arbeit wurden an zahlreichen Stellen alternative Forschungsansätze und weiterführende Fragestellungen angesprochen bzw. diskutiert. Fasst man diese zusammen, so ergeben sich die folgenden zentralen Forschungsgebiete und -aufgaben, die Gegenstand weiterer Forschung sein sollten:

1. *Verstärkte Forschung zur internen Konstruktvalidität von strukturierten Interviews.* Studien und Metaanalysen zur Konstruktvalidität des strukturierten Interviews basieren auf dem Konzept des nomologischen Netzwerkes. D.h. es werden (mittlere) Korrelationen zwischen Interview und verschiedenen Außenkriterien berechnet. Wie oben dargestellt können jedoch mit Interviews eine Vielzahl von Dimensionen erfasst werden. Die Frage nach der Konstruktvalidität strukturierter Interviews ist daher nicht gleichzusetzten mit der Frage nach den mit strukturierten Interviews erfassten Konstrukten (externe Konstruktvalidität; vgl. Kolk, 2001). Vielmehr gilt es herauszufinden, wie man strukturierte Interviews so verbessern kann, dass sie die jeweils beabsichtigten, *im Vorfeld definierten Dimensionen* mit einer höheren Genauigkeit erfassen (interne Konstruktvalidität, vgl. Kolk, 2001). Dieser Ansatz hat sich in der AC-Forschung als äußerst anregend erwiesen (vgl. die Überblicksarbeiten von Kolk et al., 2001; Lievens, 1998). In der vorliegenden Arbeit konnten die Fähigkeit, Anforderungsdimensionen zu erkennen, und die Bekanntgabe der Dimensionen (Transparenz) als Moderatoren der internen Konstruktvalidität des strukturierten Interviews bestätigt werden, während die gemeinsame Beobachtervarianz keinen Einfluss zu haben scheint. Die Suche nach weiteren solchen Moderatoren scheint daher sinnvoll und erfolgsversprechend.

2. *Übertragung unserer Ergebnisse auf reale Auswahlsituationen.* Wie oben ausgeführt, gibt es mindestens zwei Unterschiede zwischen Teilnehmern eines Bewerbungstrainings und Bewerbern auf eine reale Stelle, nämlich Motivation und Homogenität der Teilnehmer. Auch wenn die empirischen

Ergebnisse aus Kapitel 4 darauf hinweisen, dass die gefundenen Effekte – zumindest bezüglich Leistungsanstieg und Anstieg der Konstruktvalidität unter Transparenz – auch für Bewerbergruppen gelten, so ist es doch notwendig entsprechende Studien in Unternehmen durchzuführen. Solche Studien und die damit verbundenen Erfahrungen sind auch sinnvoll, wenn es darum geht die Akzeptanz bei Entscheidern in der Wirtschaft zu erhöhen und somit die Anwendung wissenschaftlicher Ergebnisse in der Praxis zu ermöglichen. Die Beachtung und Anwendung wissenschaftlicher Ergebnisse in der Praxis ist ein wichtiges Kriterium für eine anwendungsorientierte Wissenschaft.

3. *Untersuchung der Auswirkungen auf die (prognostische) Kriteriumsvalidität.* Obwohl wir die vermehrte Beachtung der Konstruktvalidität und die Suche nach weiteren Moderatoren der internen Konstruktvalidität strukturierter Interviews fordern, so ist damit nicht die Abkehr von Studien zur prognostischen Validität gemeint. Die (prognostische) Kriteriumsvalidität ist und bleibt zentrales Gütekriterium für den Anwender in der Wirtschaft. Dies gilt insbesondere dann, wenn das Interview zu Zwecken der Personalselektion genutzt wird. Die Untersuchung der Konstruktvalidität eines Auswahlverfahren hat immer auch den Sinn, die spezifischen Wirkmechanismen eines Verfahrens zu verstehen und damit die Qualität der Selektion unter wandelnden Rahmenbedingungen (z.B. Übertragung auf andere Jobgruppen) zu erhalten bzw. sogar zu verbessern. Die Moderatoren der internen Konstruktvalidität sind daher auch bezüglich ihres Einflusses auf die (prognostische) Kriteriumsvalidität zu untersuchen. In diesem Punkt besteht auch in der AC-Forschung ein gewisser Nachholbedarf, der möglichst bald abgebaut werden sollte.

   Als Beispiel sei auf die Forschung zu CDD („capability to discern dimensions“, Fähigkeit, Anforderungsdimensionen zu erkennen) verwiesen. Wie gezeigt werden konnte, hat diese moderierenden Einfluss auf die konvergente Validität und somit Konstruktvalidität von Interview und AC. Darüber hinaus scheint CDD einen Teil der gemeinsamen Varianz zwischen Prädiktor und Kriterium zu erklären (siehe Kapitel 3 und 4). Beim Versuch der Konstruktaufklärung (vgl. Hartstein & Kleinmann, 2002) hat sich ein neu entwickeltes videogestütztes Verfahren zur Erfassung der sozialen Wahrnehmung (Becker & Staufenbiel, 2002) als besonders erfolgreich

erwiesen. In Folgestudien sollte nun untersucht werden, ob dieses Instrument eine wie erwartet hohe (prognostische) Kriteriumsvalidität besitzt. Falls dies der Fall ist, so könnte das Verfahren zukünftig zusätzlich zu einem transparenten AC bzw. Interview durchgeführt werden. Als Ergebnis würde man möglicherweise ein konstruktvalides Auswahlverfahren erhalten, dass gleichzeitig eine hohe prognostische Validität besitzt.

5.3 Konsequenzen für die Praxis

Praktiker wünschen sich Auswahlverfahren, die bei der Durchführung wenig Aufwand verlangen und trotzdem gute Vorhersagen erlauben. Ein besonders beliebtes Selektionsverfahren ist das Einstellungsinterview. Die wissenschaftliche Forschung hat einige Faktoren herausgefunden, mit deren Hilfe sich die Qualität von Interviews verbessern lässt. Zur Verbesserung der Reliabilität und prognostischen Kriteriumsvalidität haben sich insbesondere die folgenden Einflussfaktoren als bedeutsam erwiesen:

1. *Strukturierung/Standardisierung.* Dies betrifft sowohl der Interviewfragen als auch die Bewertung der Antworten (vgl. Campion et. al., 1997).

2. *Anforderungsanalyse im Vorfeld.* Eine solche wird inzwischen bei den meisten strukturierten Interviews standardmäßig durchgeführt (vgl. Janz, 1982; Latham et al., 1980).

Der positive Einfluss dieser Faktoren wurde in vielen Studien demonstriert. In der vorliegenden Arbeit konnten nun in ersten Untersuchungen Einflussfaktoren auf die interne Konstruktvalidität des strukturierten Interviews identifiziert werden. Als besonders wichtig erwies sich dabei die Bekanntgabe der Anforderungsdimensionen (Transparenz, vgl. Kapitel 4). Aufgabe der Forschung ist es nun, diese Ergebnisse weiter zu sichern und so aufzubereiten, dass sie für den interessierten Praktiker verständlich werden. Nur so kann die Umsetzung wissenschaftlicher Erkenntnisse in den Praxisalltag gelingen.

6 Zusammenfassung

6.1 Theoretischer Hintergrund

6.1.1 Kriteriumsvalidität und Akzeptanz von Einstellungsinterviews

Einstellungsinterviews erfreuen sich großer Akzeptanz und weiter Verbreitung (vgl. Schuler et al., 1993; Schulz et al., 1985), obwohl ihre prognostische Validität lange Zeit als sehr schlecht galt (z.B. Reilly & Chao, 1982; Schmitt, 1976). Inzwischen liegen jedoch eine Reihe von Metaanalysen vor, welche die Validität und Reliabilität des Interviews belegen (z.B. Conway et al., 1995; Huffcutt & Arthur, 1994; Schmidt & Hunter, 1998; Schmidt & Rader, 1999; Wiesner & Cronshaw, 1988; Wright et al., 1989). Als Moderatoren der Kriteriumsvalidität wurden nach Schuler (1992) insbesondere Anforderungsbezug/Arbeitsanalyse, die Anzahl der Interviewer und Standardisierung/Strukturierung des Interviews diskutiert, wobei sich letzterer als wichtigster Faktor erwiesen hat.

Unter den strukturierten Interviews haben sich in der wissenschaftlichen Literatur insbesondere das Biographische Interview (BI; vgl. Janz, 1982, 1989) und das Situative Interview (SI; vgl. Latham & Saari, 1984; Latham et al., 1980) durchgesetzt. Hauptunterschied zwischen den beiden Interviews ist, dass biographische Fragen vergangenheitsbezogen und situative Fragen zukunftsbezogen sind.

Da sich eine zu starke Standardisierung des Interviews jedoch negativ auf die Akzeptanz durch die Bewerber und Anwender auswirkt (siehe Campion et al., 1997; Latham, 1989) wurde von Schuler (1989, 1992) das sogenannte Multimodale Interview (MMI) entwickelt. Dieses besteht aus maximal acht strukturierten und unstrukturierten Komponenten und ermöglicht eine hohe prognostische Validität bei gleichzeitiger Tauglichkeit zur Selbstselektion und unter Wahrung der sozialen Akzeptanz und Praktikabilität (vgl. Schuler, 1992; Schuler & Moser, 1995).

6.1.2 Die Konstruktvalidität strukturierter Interviews

Wie oben dargestellt ist die Kriteriumsvalidität strukturierter Interviews gut belegt. Es gibt jedoch erst wenige Studien zur Konstruktvalidität strukturierter Interviews. Dieser Mangel an einschlägigen Studien ist sehr bedauerlich, denn der Prozess der

Konstruktvalidierung hilft dabei zu verstehen, unter welchen Bedingungen Selektionsinterviews prädiktiv valide sind (vgl. Cronshaw & Wiesner, 1989).

Grundsätzlich lassen sich nach Kolk (2001) zwei Ansätze der Konstruktvalidierung unterscheiden. Bei Untersuchung der *externen* Konstruktvalidität (vgl. nomologisches Netzwerk von Cronbach & Meehl, 1955) werden Zusammenhänge zwischen dem interessierenden Verfahren und externen Maßen und Kriterien überprüft. Diese Maße (z.B. soziale Erwünschtheit, verbale Intelligenz) werden mit Verfahren außerhalb des eigentlich interessierenden Messinstrumentes (z.B. AC oder Interview) erfasst, sollen aber in charakteristischer Weise mit diesem zusammenhängen. Bei der *internen* Konstruktvalidität interessiert hingegen die Struktur innerhalb des Messinstrumentes, also konvergente und diskriminante Validität der innerhalb des Verfahrens benutzten Dimensionen. Diese Struktur wird meist mit Hilfe von Multitrait-Multimethod-Matrizen (MTMM, Campbell & Fiske, 1959) untersucht, wobei häufig die konfirmatorische Faktorenanalyse (CFA, Byrne, 1994; Widaman, 1985) zur Auswertung herangezogen wird.

Die wenigen bisher vorliegenden Metaanalysen zur Konstruktvalidität strukturierter Interviews (Huffcutt et al., 2001; Huffcutt et al., 1996; Mumford et al., 1996; Salgado & Moscoso, 2002) arbeiten mit dem Konzept der externen Konstruktvalidierung. Kritisch an diesem Vorgehen ist, dass mit der Methode Interview – ähnlich wie mit anderen Methoden (z.B. AC, Fragebögen) auch – je nach Zielsetzung und Art der Konstruktion prinzipiell eine Vielzahl unterschiedlicher Konstrukte erfasst werden kann (vgl. Huffcutt et al., 2001; Landy, 2002; Posthuma et al., 2002; Schuler, 1992). Folglich sind Aussagen wie „behavior interviews assess GMA to a lower degree than conventionell interviews“ (Salgado & Moscoso, 2002, S. 312) nur bedingt sinnvoll. Interessant ist vielmehr, ob strukturierte Interviews prinzipiell in der Lage sind, die beabsichtigten Dimensionen zu erfassen, und zwar unabhängig davon, um welche konkreten Dimensionen es sich im spezifischen Fall handelt. Weiterhin sollte herausgefunden werden, unter welchen Bedingungen (z.B. Unabhängigkeit der Dimensionen, vgl. Kleinmann et al., 1995) eine valide Erfassung dieser Dimensionen begünstigt wird. Wir stimmen daher mit Maurer et al. (1999) überein, dass wie in der AC-Forschung auch, die interne Konstruktvalidität strukturierter Interviews stärker untersucht werden sollte. Bisher ist uns nur eine Studie bekannt (Schuler, 1989 bzw. Schuler & Funke, 1989), in welcher die interne Konstruktvalidität strukturierter Interviews untersucht wurde. Dabei kamen die Autoren zu dem Ergebnis, dass die

Konstruktvalidität des MMI gering ist. Leider wurde die zugehörige MTMM-Matrix nicht berichtet, so dass eine Reanalyse der Daten mit Hilfe der konfirmatorischen Faktorenanalyse nicht möglich ist.

6.1.3 Zielrichtung der Arbeit

Wie oben dargestellt, ist die interne Konstruktvalidität strukturierter Interviews ein bisher vernachlässigtes Thema. Dies ist sehr erstaunlich, da die valide Erfassung der durch die Anforderungsanalyse erhobenen Dimensionen elementare Voraussetzung für die Vorhersage zukünftiger Leistungen sein sollte. Die zentrale Fragestellung der ersten Studie (Kapitel 2) ist daher die Untersuchung der internen Konstruktvalidität strukturierter Interviews. In einem zweiten Schritt werden dann mögliche Einflussfaktoren auf die interne Konstruktvalidität untersucht. Grundlegend hierfür sind die Arbeiten von Kleinmann und Kollegen (Kleinmann, 1993, 1997a, 1997b; Kleinmann et al., 1996), in denen solche Einflussfaktoren zum Assessment Center untersucht wurden. Entsprechend lauten die zentralen Fragestellungen der Studien 2 und 3, ob das Ausmaß, in dem Bewerber die Anforderungsdimensionen eines strukturierten Interviews erkennen, Einfluss auf die konvergente Validität und die gezeigte Leistung hat (Kapitel 3) und ob die Bekanntgabe/Enthüllung der Anforderungsdimensionen (Transparenz) Einfluss auf die Konstrukt- und Kriteriumsvalidität des strukturierten Interviews hat (Kapitel 4).

6.2 Konvergente und diskriminante Validität des strukturierten Interviews (Studie 1)

Hauptziel der ersten Studie war die Überprüfung der internen Konstruktvalidität des strukturierten Einstellungsinterviews mit Hilfe der konfirmatorischen Faktorenanalyse. Weiterhin wurde eine explorative Analyse zu Korrelaten des strukturierten Interviews durchgeführt und der Einfluss der gemeinsamen Beobachtervarianz auf konvergente und diskriminante Validität untersucht.

Zur Erhebung der Daten wurde ein zweitägiges Bewerbungstraining für Hochschulabsolventen und Studierende am Ende des Studiums mit insgesamt 110 Teilnehmern durchgeführt. Dabei kamen ein Multimodales Interview (MMI, vgl. Schuler, 1992) bestehend aus Selbstpräsentation (SP), Biographischem Interview (BI) und Situativem Interview (SI) sowie ein Assessment Center (AC), bestehend aus

einer Postkorbaufgabe (PK) und zwei Gruppendiskussionen (GD1 und GD2), zum Einsatz. Bei der Entwicklung der Anforderungsdimensionen (Systematisches Denken & Handeln, Zusammenarbeit, Steuerung sozialer Prozesse) wurden potentielle Einflussfaktoren auf die interne Konstruktvalidität beachtet, welche sich in der AC-Forschung als relevant herausgestellt hatten (vgl. Lievens, 1998).

Weiterhin wurden durchgeführt: Ein Intelligenztest (IST 2000, Amthauer et al., 1999), eine PC-gestützte Simulation zur Erfassung der Problemlösefähigkeit (Textilfabrik, Hasselmann & Strauß 1995), ein Integritätstest (FES, Marcus & Schuler 1998), ein Fragebogen zum Self-Monitoring (Mielke & Kilian 1990) sowie ein Fragebogen zur sozialen Erwünschtheit (Lück & Timaeus, 1969).

Die Überprüfung der internen Konstruktvalidität des MMI erfolgte mit Hilfe verschiedener Auswertungsmethoden (mittlere konvergente und diskriminante Korrelationen aus der MTMM-Matrix, traditionelle konfirmatorische Faktorenanalyse und CU-Ansatz, χ^2-Differenzen-Test, erklärte Varianzanteile der Dimensionsfaktoren), da nach Conway (1996) unterschiedliche Methoden teilweise zu unterschiedlichen Ergebnissen führen können. In der vorliegenden Studie ergaben die Analysen der internen Konstruktvalidität jedoch ein einheitliches Bild. So gibt es keine Anzeichen auf diskriminante Validität und nur wenige Hinweise auf konvergente Validität, wobei insgesamt die interne Konstruktvalidität als niedrig einzuschätzen ist. Sie fällt damit ähnlich unbefriedigend aus, wie noch vor einigen Jahren in der AC-Forschung (vgl. Sackett & Dreher, 1982).

Im Rahmen der zweiten Fragestellung wurden die Zusammenhänge zwischen den einzelnen Interviewdimensionen und verschiedenen externen Maßen untersucht. Während die Dimension Zusammenarbeit vor allem mit dem Integritätstest korrelierte und tendenziell mit Self-Monitoring und sozialer Erwünschtheit zusammenhing, ergaben sich für die Dimension Steuerung sozialer Prozesse signifikante Korrelationen mit verbaler und numerischer Intelligenz. Die Dimension Systematisches Denken & Handeln wiederum korrelierte mit dem Integritätstest und mit sozialer Erwünschtheit, vor allem aber auch mit verbaler und numerischer Intelligenz sowie der Problemlöseaufgabe (Textilfabrik). Weiterhin waren die Korrelationen zwischen MMI und AC für identische Dimensionen meist höher als für unterschiedliche Dimensionen. Insgesamt konnten plausible Zusammenhänge zwischen MMI und den externen Verfahren und Konstrukten gefunden werden,

wodurch die entsprechenden einschlägigen Befunde der Interviewforschung (z.B. Mumford et al., 1996; Schuler & Moser, 1995) weitere Unterstützung erfahren.

In der letzten Fragestellung wurde der Einfluss der gemeinsamen Beobachtervarianz auf die interne Konstruktvalidität untersucht. Analog zu Kolk et al. (in press-a, Studie 2) berechneten wir zunächst eine MTMM-Matrix, in welcher jedoch nicht die Mittelwerte der Beurteilungen, sondern die Dimensionsratings der *einzelnen* Beobachter korreliert wurden. Wiederum wurden die mittleren konvergenten und diskriminanten Korrelationen berechnet. Hierbei ergaben sich keine signifikanten Unterschiede zwischen den Korrelationen mit und ohne gemeinsame Beobachtervarianz. Die gemeinsame Beobachtervarianz hat somit keinen Einfluss auf die interne Konstruktvalidität des strukturierten Interviews. Dieses Ergebnis wird auf die verhaltensverankerten Beurteilungsskalen zurückgeführt, welche spezifisch für jede Interviewfrage entwickelt wurden. Sie ermöglichen eine sehr reliable Beurteilung, wodurch Halo-Effekte vermutlich nur in geringem Ausmaß wirksam werden konnten.

Zusammenfassend lässt sich festhalten: Wie erwartet ist die interne Konstruktvalidität des strukturierten Interviews als eher gering einzuschätzen. Dies ist insofern überraschend, als das sich in einer explorativen Analyse einige plausible Korrelationen zwischen den einzelnen Interviewdimensionen und den verschiedenen externen Kriterien finden ließen. Die Unterscheidung zwischen interner und externer Konstruktvalidität erscheint daher sowohl aus logischer als auch empirischer Sicht sinnvoll. Anders als in der AC-Forschung, scheint der Einfluss der gemeinsamen Beobachtervarianz auf konvergente und diskriminante Validität des strukturierten Interviews vernachlässigbar. In den beiden folgenden Studien wurden weitere potentielle Einflussfaktoren auf die interne Konstruktvalidität des strukturierten Interviews untersucht.

6.3 Die Fähigkeit, Anforderungsdimensionen zu erkennen (Studie 2)

Im Rahmen der zweiten Studie wurden die Annahmen untersucht, dass die Fähigkeit, Anforderungsdimensionen zu erkennen („capability to discern dimensions“, CDD, Hartstein & Kleinmann, 2002) mit der Beurteilung durch die Beobachter korreliert und eine Moderatorvariable der konvergenten Validität im strukturierten Interview darstellt

(vgl. Kleinmann, 1993). Weiterhin sollten insbesondere die Annahmen überprüft werden, dass es sich bei CDD um eine personenbezogene Fähigkeit handelt, welche relativ unabhängig vom jeweiligen eignungsdiagnostischen Instrument erfasst werden kann. Die Datenerhebung erfolgte im Rahmen eines eintägigen Bewerbungstrainings mit insgesamt 95 Teilnehmern, wobei wiederum das MMI und das AC aus Studie 1 zum Einsatz kamen.

Die Erhebung der CDD-Werte ging folgendermaßen vonstatten: Nach jedem Übungsteil (z.B. Gruppendiskussion, Biographisches Interview) schrieben die Teilnehmer in eigenen Worten auf, worauf es ihrer Meinung nach bei der gerade eben durchgeführten Übung angekommen war (Hypothesen der Teilnehmer). Am Ende des Tages erhielten die Teilnehmer für jede Übung ein Blatt mit sechs potentiellen Dimensionen inklusive Verhaltensbeispielen ausgeteilt und erläutert, wobei neben den drei „echten“ Dimensionen auch die drei Distraktoren (irrelevante Dimensionen) zu finden waren. Die Teilnehmer ordneten dann anhand eines Ratings von 1 („meine Hypothese entspricht etwas der Anforderungsdimension“) bis 4 („meine Hypothese entspricht vollständig der Anforderungsdimension“) jede einzelne Hypothese genau einer Dimension zu. Hierbei bestand auch die Möglichkeit „keine der angegebenen Dimensionen“ anzugeben. Zur Ermittlung des CDD-Wertes wurden dann die Ratings derjenigen Hypothesen addiert, welche den korrekten Dimensionen zugeordnet waren. Hierbei wurde jeweils das höchste Rating pro Dimension und Übung genommen.

In Hypothese 1 wurde die Annahme überprüft, dass sich die Teilnehmer darin unterscheiden, in welchem Ausmaß sie die Anforderungsdimensionen im strukturierten Interview erkennen. Analog zu Kleinmann (1993) wurden daher Verteilung und innere Konsistenz (Cronbach´s Alpha) der CDD-Werte untersucht. Dabei zeigten sich zufriedenstellende innere Konsistenzen der CDD-Skala für das Gesamtinterview (MMI) sowie ausreichende bis befriedigende Werte für die Komponenten BI und SI. Lediglich in der Selbstpräsentation war die Reliabilität der CDD-Skala ungenügend. Weiterhin waren die CDD-Werte annähernd normalverteilt und erstreckten sich fast über die gesamte Spannweite der theoretisch möglichen Werte, so dass die erste Hypothese zusammenfassend als bestätigt angesehen werden kann.

In Hypothese 2 wurde der Zusammenhang zwischen der Leistungsbeurteilung und dem Ausmaß des Erkennens untersucht. Wie erwartet zeigten sich signifikante

Korrelationen zwischen den CDD-Werten und den zugehörigen Leistungsmaßen. Diese lagen im gleichen Größenbereich wie sie auch von Kleinmann (1993, 1997b) berichtet werden. Personen schneiden also sowohl im AC als auch im strukturierten Interview besser ab, wenn es ihnen gelingt die zugrunde liegenden Anforderungsdimensionen zu erkennen. Lediglich für die Selbstpräsentation konnte die Hypothese nicht bestätigt werden, was wahrscheinlich auf die geringe Reliabilität der zugehörigen CDD-Werte (vgl. Hypothese 1) zurückgeführt werden kann.

In Hypothese 3 wurde postuliert, dass die Erkennensfähigkeit neben interindividuellen Leistungsunterschieden (vgl. Hypothese 2) auch intraindividuelle Unterschiede in den Bewertungen erklärt. Hierzu wurden für jeden Teilnehmer zwei Mittelwerte berechnet (mittlere Bewertung für Fragen mit erkannter bzw. nicht erkannter Anforderungsdimension). Wie erwartet erzielten die Teilnehmer in den Fragen, deren Dimensionen sie erkannt hatten, signifikant bessere Bewertungen als in den Interviewkomponenten, deren Dimensionen sie nicht erkannt hatten. Dieses Ergebnis stützt die Hypothese, dass das Erkennen der Anforderungsdimensionen maßgebliche Auswirkung auf die gezeigte Leistung hat.

In Hypothese 4 wurde der Einfluss der Erkennensfähigkeit auf die konvergente Validität des MMI untersucht. Hierzu wurden zunächst alle möglichen Beurteilungspaare bezüglich identischer Dimensionen gebildet, welche dann wiederum in drei Gruppen aufgeteilt wurden (Teilnehmer die keine, eine oder beide Anforderungsdimensionen erkannt hatten). Wie erwartet ergaben sich in Gruppe 3 die höchsten Zusammenhänge. Die Teilnehmer in dieser Gruppe hatten bezüglich beider Fragen erkannt, welche Anforderungsdimensionen erfasst werden sollen. Daher war ein konsistentes Antwortverhalten möglich, welches entsprechend zu einer hohen Korrelation zwischen den Bewertungen in identischen Dimensionen führte. Das Ausmaß, in welchem die Anforderungsdimensionen erkannt werden, hat demnach direkten Einfluss auf die konvergente Validität des strukturierten Interviews. Anders als bei Kleinmann (1993) ergab sich jedoch auch in der Gruppe 1 (keine Anforderungsdimension erkannt) ein geringer Zusammenhang zwischen Bewertungen auf identischen Dimensionen.

Im Folgenden wurden verschiedene Hypothesen bezüglich der Zusammenhänge zwischen der im AC bzw. MMI ermittelten CDD-Werte sowie den verschiedenen Leistungsbeurteilungen überprüft. Wie erwartet ergaben sich signifikante Korrelationen zwischen den CDD-Werten aus AC und den CDD-Werten im Interview

sowie signifikante Zusammenhänge zwischen den CDD-Werten im Interview und den Leistungswerten im AC bzw. den CDD-Werten im AC und den Beurteilungen im Interview. Damit konnte gezeigt werden, dass die Leistungsbeurteilung in einem eignungsdiagnostischen Verfahren (AC bzw. MMI) durch die Erkennensfähigkeit (CDD-Werte) aus einem *anderen* Verfahren vorhergesagt werden kann und dass diese Vorhersage ähnlich gut ausfällt, wie wenn man die jeweils zum Verfahren gehörigen CDD-Werte verwendet. Dieses Ergebnis stützt die zentrale Annahme, dass die Erkennensfähigkeit eine personengebundene Fähigkeit ist, die relativ unabhängig vom zugrunde liegenden Testverfahren erhoben werden kann.

In der letzten Hypothese wurde postuliert, dass der Zusammenhang zwischen den Leistungsbeurteilungen im AC und im MMI teilweise durch die Erkennensfähigkeit der Teilnehmer erklärt werden kann. Wie erwartet kam es bei Herauspartialisierung der Erkennensfähigkeit zu einem Absinken der Korrelation zwischen MMI und AC, wobei in Ermangelung eines geeigneten Testverfahrens kein Signifikanztest durchgeführt werden konnte. Dieses Ergebnis stützt die Hypothese, dass eignungsdiagnostische Verfahren nicht nur die beabsichtigten Dimensionen (beispielsweise Zusammenarbeit oder Steuerung sozialer Prozesse) erfassen, sondern zum Teil auch solche Aspekte wie die Fähigkeit, relevante Anforderungsdimensionen zu erkennen (CDD).

Zusammenfassend lässt sich festhalten: Das Ausmaß des Erkennens der Anforderungsdimensionen hat sowohl im AC als auch im strukturierten Interview Einfluss auf die gezeigte Leistung und die konvergente Validität. Weiterhin sind Auswirkungen auf die Kriteriumsvalidität plausibel, die sich auch deskriptiv zeigen ließen.

6.4 Die Transparenz der Anforderungsdimensionen (Studie 3)

In der dritten Studie wurden die Forschungsergebnisse von Kleinmann et al. (1996) bzw. Kleinmann (1997b) auf das strukturierte Interview übertragen. So sollte gezeigt werden, dass die Bekanntgabe der Anforderungsdimensionen (Experimental-Bedingung Transparenz) ein konsistenteres Verhalten der Teilnehmer zur Folge hat und zur Erhöhung der internen Konstruktvalidität führt.

Zur Erhebung der Daten wurden zwei unabhängigen Studien A und B mit N = 123 (EG *N* = 59, KG *N* = 64) bzw. N = 176 (EG *N* = 68, KG *N* = 108) durchgeführt. Während in Studie A die Auswirkungen der Transparenz auf die Konstruktvalidität im Mittelpunkt des Interesses standen, konnten in Studie B zusätzlich die Folgen für die Kriteriumsvalidität untersucht werden.

In unserer ersten Hypothese postulierten wir einen Leistungsanstieg im MMI unter Transparenz, obwohl in der AC-Forschung (Kolk et al., 2000) kein solcher Effekt gefunden werden konnte. In der Tat konnten wir in beiden Studien eine entsprechende Verbesserung der Leistungsbeurteilungen nachweisen. Wir vermuten, dass die Beobachter im AC automatisch auf die anderen AC-Teilnehmer als Vergleichsmaßstab zurückgreifen, so dass die mittleren Beurteilungen in den beiden Bedingungen nur wenig differieren. Im MMI werden die Teilnehmerantworten hingegen aufgrund von verhaltensverankerten Beurteilungsskalen bewertet, die spezifisch für jede Frage entwickelt worden. Dadurch wird ein Vergleich der Leistungen zwischen den Teilnehmern unnötig. Hinzu kommt, dass das Interview eine Einzelübung ist, wodurch der Vergleich mit anderen Teilnehmern erschwert wird. Lediglich für das BI in Studie A ergab sich kein signifikanter Anstieg der Leistung.

In der zweiten Hypothese wurden mit Hilfe der konfirmatorischen Faktorenanalyse (CFA) die Folgen transparenter Dimensionen auf die interne Konstruktvalidität des MMI untersucht. Wie erwartet zeigten sich unter Intransparenz vermehrt Konvergenzprobleme. Daher konnten in Studie A nur die „correlated-uniqueness“-(CU)-Modelle der Transparenz-Bedingung auf konvergente und diskriminante Validität geprüft werden. Hierbei zeigte sich eine zufriedenstellende Konstruktvalidität des transparent durchgeführten Interviews. In Studie B konvergierten hingegen alle getesteten CU-Modelle, so dass die interne Konstruktvalidität des MMI sowohl für Transparenz- als auch Intransparenz-Bedingung mit Hilfe von hierarchischen χ^2-Tests überprüft werden konnte. Hypothesenkonform erklärte in der Intransparenz-Bedingung ein Modell mit nur einem Dimensions-Faktor die Daten am besten, während in der Transparenz-Bedingung ein Modell mit drei Dimensions-Faktoren den besten Fit erzielte. Damit erweist sich die Bekanntgabe der Dimensionen sowohl im AC als auch beim strukturierten Interview als effektiver Moderator der internen Konstruktvalidität.

Die dritte Hypothese besagte, dass die Bekanntgabe der Anforderungsdimensionen bei den Teilnehmern zu einer Reduzierung irrelevanter Verhaltensweisen führt, wodurch es in der Folge zu einem Anstieg der Beurteilerübereinstimmung kommt. Zusammenfassend lassen sich in beiden Studien Hinweise auf entsprechende Effekte finden; insbesondere zeigten sie sich für Beurteilungen, welche nicht mit Verhaltensankern versehen sind. Dies ist nicht weiter verwunderlich, da die Reliabilität verhaltensverankerter Beurteilungsskalen auch ohne Transparenz-Manipulation schon sehr hoch ist und daher nur unter großen Mühen weiter gesteigert werden kann.

In der vierten und letzten Hypothese wurde der Zusammenhang zwischen MMI (Intransparent/Transparent) und einem intransparentem Kriterium (AC) untersucht. Auch diese Fragestellung sollte wichtige Hinweise auf die im strukturierten Interview erfassten Konstrukte liefern. Ein Absinken der Korrelation unter der Transparenz-Bedingung wäre als Beleg aufzufassen, dass durch die experimentelle Manipulation gemeinsame Varianz in Prädiktor und Kriterium verloren geht. Dies wäre ein gewichtiger Anhaltspunkt dafür, dass die Fähigkeit, Anforderungsdimensionen zu erkennen (vgl. Kapitel 3), ein im strukturierten Interview gemessenes Konstrukt darstellt, das bedeutsam für ihre prognostische Validität ist. Bei deskriptiver Betrachtung zeigten sich hypothesenkonform höhere Kriteriumsvaliditäten für das intransparente MMI. Die Unterschiede wurden jedoch nicht signifikant, obwohl die Zusammenhänge zum Teil fast halbiert wurden.

Zusammenfassend lässt sich festhalten: Die Bekanntgabe der Anforderungs-dimensionen (Transparenz) erweist sich als effektiver Moderator der internen Konstruktvalidität im strukturierten Interview, mit tendenziell positiven Auswirkungen auf die Beurteilerübereinstimmung. Wie erwartet zeigen sich des weiteren tendenziell negative Auswirkungen auf die Kriteriumsvalidität, die jedoch weiter untersucht werden müssen.

Im letzten Abschnitt folgt eine zusammenfassende Bewertung der durchgeführten Studien und erzielten Resultate. Außerdem werden generelle Grenzen unseres Forschungsansatzes diskutiert und Hinweise auf zukünftige Fragestellungen gegeben.

Literatur

Amthauer, R., Brocke, B., Liepmann, D. & Beauducel, A. (1999). *Intelligenz-Struktur-Test 2000*. Göttingen: Testzentrale.

Arthur, W., Woehr, D. J. & Maldegen, R. (2000). Convergent and discriminant validity of assessment center dimensions: A conceptual and empirical reexamination of the assessment center construct-related validity paradox. *Journal of Management, 26*, 813-835.

Becker, B. & Staufenbiel, T. (2002). *Entwicklung eines videogestützten Verfahrens zur Messung der sozialen Wahrnehmung.* Paper presented at the 43. Kongress der Deutschen Gesellschaft für Psychologie, Berlin.

Becker, T. E. & Cote, J. A. (1994). Additive and multiplicative method effects in applied psychological research: An empirical assessment of three models. *Journal of Management, 20*, 625-641.

Binning, J. F. & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology, 74*, 478-494.

Borchert, N. (2001). *Die Fähigkeit, relevante Anforderungsdimensionen zu erkennen: Eine Moderatorvariable im Multimodalen Interview.* Unpublished diploma thesis, Philipps-Universität, Marburg.

Bortz, J. (1999). *Statistik für Sozialwissenschaftler* (5th ed.). Berlin: Springer.

Brannick, M. T., Michaels, C. E. & Baker, D. P. (1989). Construct validity of in-basket scores. *Journal of Applied Psychology, 74*, 957-963.

Bungard, W. (1987). Zur Problematik von Reaktivitätseffekten bei der Durchführung eines Assessment Centers. In H. Schuler & W. Stehle (Eds.), *Assessment Center als Methode der Personalentwicklung* (pp. 99-125). Stuttgart: Verlag für Angewandte Psychologie.

Bycio, P., Alvares, K. M. & Hahn, J. (1987). Situational specificity in assessment center ratings: A confirmatory factor analysis. *Journal of Applied Psychology, 72*, 463-474.

Byrne, B. M. (1994). *Structural equation modeling with EQS and EQS/Windows: basic concepts, applications, and programming*. Thousand Oaks, CA: Sage Publications.

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Campion, M. A., Campion, J. E. & Hudson, J. P. (1994). Structured interviewing: A note on incremental validity and alternative question types. *Journal of Applied Psychology, 79*, 998-1102.

Campion, M. A., Palmer, D. K. & Campion, J. E. (1997). A review of structure in the selection interview. *Personnel Psychology, 50*, 655-702.

Campion, M. A., Pursell, E. D. & Brown, B. K. (1988). Structured interviewing: Raising the psychometric properties of the employment interview. *Personnel Psychology, 41*, 25-42.

Conway, J. M. (1996). Analysis and design of multitrait-multirater performance appraisal studies. *Journal of Management, 22*, 139-162.

Conway, J. M., Jako, R. A. & Goodman, D. F. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology, 80*, 565-579.

Conway, J. M. & Peneno, G. M. (1999). Comparing structured interview question types: Construct validity and applicant reactions. *Journal of Business and Psychology, 13*, 485-506.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Cronshaw, S. F. & Wiesner, W. H. (1989). The validity of the employment interview: Models for research and practice. In R. W. Eder & G. R. Ferris (Eds.), *The employment interview: Theory, research, and practice* (pp. 269-281). Newbury Park, CA: Sage Publications.

Crowne, D. P. & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24*, 349-354.

Deller, J. (1991). *Situatives Interview: Entwicklung und Evaluierung.* Unpublished diploma thesis, Christian-Albrechts-Universität, Kiel.

Deller, J. & Kendelbacher, I. (1998). Potentialeinschätzung von oberen Führungskräften im Daimler-Benz-Konzern. In M. Kleinmann & B. Strauss (Eds.), *Potentialfeststellung und Personalentwicklung* (pp. 133-149). Göttingen: Verlag für Angewandte Psychologie.

Deutscher Sparkassen- und Giroverband. (1988). *Handbuch: Standardisierungshilfen für Einstellungsgespräche. Ausbildungsberufe Bankkaufmann/ Sparkassenkaufmann.* Stuttgart: Deutscher Sparkassenverlag.

Diemand, A. & Schuler, H. (1991). Sozial erwünschtes Verhalten in eignungsdiagnostischen Situationen. In H. Schuler & U. Funke (Eds.), *Eignungsdiagnostik in Forschung und Praxis. Psychologische Information für Auswahl, Beratung und Förderung von Mitarbeitern* (pp. 242-248). Stuttgart: Verlag für Angewandte Psychologie.

Diemand, A. & Schuler, H. (1998). Wirksamkeit von Selbstdarstellungsvariablen im Rahmen der prognostischen Validierung eines Potentialanalyseverfahrens Effectiveness of self-presentation variables in the context of the predictive validation of an assessment center. *Zeitschrift für Arbeits- und Organisationspsychologie, 42*, 134-146.

Donahue, L. M., Truxillo, D. M., Cornwell, J. M. & Gerrity, M. J. (1997). Assessment center construct validity and behavioral checklists: Some additional findings. *Journal of Social Behavior and Personality, 12*, 85-108.

Eder, R. W., Kacmar, K. M. & Ferris, G. R. (1989). Employment interview research: History and synthesis. In R. W. Eder & G. R. Ferris (Eds.), *The Employment Interview: Theory, Research, and Practice* (pp. 17-31). Newbury Park, CA: Sage Publications.

Feild, H. S. & Gatewood, R. D. (1989). Development of a selection interview: A job content strategy. In R. W. Eder & G. R. Ferris (Eds.), *The employment interview: Theory, research, and practice* (pp. 145-157). Newbury Park, CA: Sage Publications.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51*, 327-358.

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

Gage, N. L. & Berliner, D. C. (1996). *Pädagogische Psychologie* (5th ed.). Weinheim: Psychologie Verlags Union.

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C. & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72*, 493-511.

Gaugler, B. B. & Thornton, G. C. (1989). Number of assessment center dimensions as a determinant of assessor accuracy. *Journal of Applied Psychology, 74*, 611-618.

Gill, R. W. (1982). A trainability concept for management potential and an empirical study of its relationship with intelligence for two managerial skills. *Journal of Occupational Psychology, 55*, 139-147.

Haaland, S. & Christiansen, N. D. (2002). Implications of trait-activation theory for evaluating the construct validity of assessment center ratings. *Personnel Psychology, 55*, 137-163.

Harris, M. M. (1999). What is being measured? In R. W. Eder & M. M. Harris (Eds.), *The Employment Interview Handbook* (pp. 143-157). Thousand Oaks, CA: Sage Publications.

Harris, M. M., Becker, A. S. & Smith, D. E. (1993). Does the assessment center scoring method affect the cross-situational consistency of ratings? *Journal of Applied Psychology, 78*, 675-678.

Hartstein, T. (2003). *Das Mosaik der Konstruktvalidität - Untersuchungen zum Erkennen relevanter Anforderungsdimensionen im Assessment-Center.* Unpublished dissertation, Philipps-Universität, Marburg.

Hartstein, T. & Kleinmann, M. (2002). *Das Erkennen von Anforderungsdimensionen in Assessment Centern und der Zusammenhang mit Leistungs- und Persönlichkeitsmaßen.* Paper presented at the 43. Kongress der Deutschen Gesellschaft für Psychologie, Berlin.

Hasselmann, D. (1993). Eignungsdiagnostische Validität des computersimulierten Szenarios Textilfabrik. In A. Gebert & U. Winterfeld (Eds.), *Arbeits-, Betriebs- und Organisationspsychologie vor Ort. Bericht über die 34. Fachtagung der Sektion Arbeits-, Betriebs- und Organisationspsychologie im Berufsverband Deutscher Psychologen e.V., Bad Lauterberg 1992* (pp. 541-550). Bonn: Deutscher Psychologen Verlag.

Hasselmann, D. & Strauß, B. (1995). Herausforderung Komplexität. Computersimulierte Problemlöseaufgaben für Management-Diagnostik und Training. Baustein 2 (Textilfabrik). Hamburg: Windmühle Verlag.

Höft, S. & Schuler, H. (2001). The conceptual basis of assessment centre ratings. *International Journal of Selection and Assessment, 9*, 114-123.

Hoyt, C. (1941). Test reliability estimated by analysis of variance. *Psychometrika, 6*, 153-160.

Huffcutt, A. I. & Arthur, W. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry-level jobs. *Journal of Applied Psychology, 79*, 184-190.

Huffcutt, A. I., Conway, J. M., Roth, P. L. & Klehe, U.-C. (2002). Comparison of the situational and behavior description interview formats. Manuscript submitted for publication.

Huffcutt, A. I., Conway, J. M., Roth, P. L. & Stone, N. J. (2001). Identification and meta-analytic assessment of psychological constructs measured in employment interviews. *Journal of Applied Psychology, 86*, 897-913.

Huffcutt, A. I., Roth, P. L. & McDaniel, M. A. (1996). A meta-analytic investigation of cognitive ability in employment interview evaluations: Moderating characteristics and implications for incremental validity. *Journal of Applied Psychology, 81*, 459-473.

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

Janz, T. (1982). Initial comparisons of patterned behavior description interviews versus unstructured interviews. *Journal of Applied Psychology, 67*, 577-580.

Janz, T. (1989). The patterned behavior description interview: The best prophet of the future is the past. In R. W. Eder & G. R. Ferris (Eds.), *The employment interview: Theory, research, and practice* (pp. 158-168). Newbury Park, CA: Sage Publications.

Jeserich, W. (1981). *Mitarbeiter auswählen und fördern. Assessment-Center-Verfahren*. München: Hanser.

Joiner, D. A. (2000). Guidelines and ethical considerations for assessment center operations: International task force on assessment center guidelines. *Public Personnel Management, 29*, 315-331.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrica, 52*, 99-111.

Kacmar, K. M., Ratcliff, S. L. & Ferris, G. R. (1989). Employment interview research: Internal and external validity. In R. W. Eder & G. R. Ferris (Eds.), *The employment interview: Theory, research, and practice* (pp. 32-41). Newbury Park, CA: Sage Publications.

Kenny, D. A. & Kashy, D. A. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin, 112*, 165-172.

Klehe, U.-C. (2000). *Die Crux mit der Konstruktvalidität.* Unpublished diploma thesis, Philipps-Universität, Marburg.

Kleinmann, M. (1991). Reaktivität von Assessment Centern. In H. Schuler & U. Funke (Eds.), *Eignungsdiagnostik in Forschung und Praxis. Psychologische Information für Auswahl, Beratung und Förderung von Mitarbeitern* (pp. 159-162). Stuttgart: Verlag für Angewandte Psychologie.

Kleinmann, M. (1993). Are rating dimensions in assessment centers transparent for participants? Consequences for criterion and construct validity. *Journal of Applied Psychology, 78*, 988-993.

Kleinmann, M. (1997a). *Assessment Center: Stand der Forschung - Konsequenzen für die Praxis*. Göttingen: Verlag für Angewandte Psychologie.

Kleinmann, M. (1997b). Transparenz der Anforderungsdimensionen: Ein Moderator der Konstrukt- und Kriteriumsvalidität des Assessment-Centers. *Zeitschrift für Arbeits- und Organisationspsychologie, 41*, 171-181.

Kleinmann, M., Andres, J., Fedtke, C., Godbersen, F. & Köller, O. (1994). Der Einfluss unterschiedlicher Auswertungsmethoden auf die Konstruktvalidität von Assessment-Centern. *Zeitschrift für Experimentelle und Angewandte Psychologie, 41*, 184-210.

Kleinmann, M., Exler, C., Kuptsch, C. & Köller, O. (1995). Unabhängigkeit und Beobachtbarkeit von Anforderungsdimensionen im Assessment Center als Moderatoren der Konstruktvalidität. *Zeitschrift für Arbeits- und Organisationspsychologie, 39*, 22-28.

Kleinmann, M. & Köller, O. (1997). Construct validity of assessment centers: Appropriate use of confirmatory factor analysis and suitable construction principles. *Journal of Social Behavior and Personality, 12*, 65-84.

Kleinmann, M., Kuptsch, C. & Köller, O. (1996). Transparency: A necessary requirement for the construct validity of assessment centres. *Applied Psychology: An International Review, 45*, 67-84.

Klimoski, R. & Brickner, M. (1987). Why do assessment centers work? The puzzle of assessment center validity. *Personnel Psychology, 40*, 243-260.

Kolk, N. J. (2001). *Assessment centers: Understanding and improving construct-related validity*. Enschede: Kurt Lewin Institut.

Kolk, N. J., Born, M. P. & Van der Flier, H. (2000). *The transparent assessment centre: The effects of revealing dimensions to candidates.* Paper presented at the 15th annual conference of the Society for Industrial and Organizational Psychology, New Orleans.

Kolk, N. J., Born, M. P. & Van der Flier, H. (2001). A meta-analysis of assessment center construct validity. Manuscript submitted for publication.

Kolk, N. J., Born, M. P. & Van der Flier, H. (in press-a). Impact of common rater variance on construct validity of assessment center dimension judgements. *Human Performance*.

Kolk, N. J., Born, M. P. & Van der Flier, H. (in press-b). The transparent assessment center: The effects of revealing dimensions to applicants. *Applied Psychology: An International Review*.

Kudisch, J. D., Ladd, R. T. & Dobbins, G. H. (1997). New evidence on the construct validity of diagnostic assessment centers: The findings may no be so troubling after all. *Journal of Social Behavior and Personality, 12*, 129-144.

Kurecka, P. M., Austin, J. M., Johnson, W. & Mendoza, J. L. (1982). Full and errant coaching effects on assigned role leaderless group discussion performance. *Personnel Psychology, 35*, 805-812.

Lance, C. E., Noble, C. L. & Scullen, S. E. (2002). A critique of the correlated trait-correlated method and correlated uniqueness models for multitrait-multimethod data. *Psychological Methods, 7*, 228-244.

Landy, F. J. (2002). Persönliche Kommunikation.

Langdale, J. A. & Weitz, J. (1973). Estimating the influence of job information on interviewer agreement. *Journal of Applied Psychology, 57*, 23-27.

Latham, G. P. (1989). The reliability, validity, and practicality of the situational interview. In R. W. Eder & G. R. Ferris (Eds.), *The employment interview: Theory, research, and practice* (pp. 169-182). Newbury Park, CA: Sage Publications.

Latham, G. P. & Finnegan, B. J. (1993). Perceived practicality of unstructured, patterned, and situational interviews. In H. Schuler & J. L. Farr (Eds.), *Personnel selection and assessment: Individual and organizational perspectives. Series in applied psychology* (pp. 41-55). Hillsdale, NJ: Lawrence Erlbaum Associates.

Latham, G. P. & Saari, L. M. (1984). Do people do what they say? Further studies on the situational interview. *Journal of Applied Psychology, 69*, 569-573.

Latham, G. P., Saari, L. M., Pursell, E. D. & Campion, M. A. (1980). The situational interview. *Journal of Applied Psychology, 65*, 422-427.

Latham, G. P. & Sue Chan, C. (1999). A meta-analysis of the situational interview: An enumerative review of reasons for its validity. *Canadian Psychology, 40*, 56-67.

Lievens, F. (1998). Factors which improve the construct validity of assessment centers: A review. *International Journal of Selection and Assessment, 6*, 141-152.

Lievens, F. (2001). Assessors and use of assessment centre dimensions: A fresh look at a troubling issue. *Journal of Organizational Behavior, 22*, 203-221.

Lievens, F. & Keer, E. V. (2001). The construct validity of a Belgian assessment centre: A comparison of different models. *Journal of Occupational and Organizational Psychology, 74*, 373-378.

Lober, K., Kleinmann, M., Borchert, N. & Richter, G. (2002). *Messen Einstellungsinterviews das, was sie messen wollen?* Paper presented at the 43. Kongress der Deutschen Gesellschaft für Psychologie, Berlin.

Locke, E. A. & Latham, G. P. (1990). *A theory of goal setting and task performance*. New York: Prentice Hall.

Lück, H. E. & Timaeus, E. (1969). Skalen zur Messung Manifester Angst (MAS) und sozialer Wünschbarkeit (SDS-E und SDS-CM). *Diagnostica, 15*, 134-141.

Marcus, B. (2000). *Kontraproduktives Verhalten im Betrieb: Eine individuumsbezogene Perspektive*. Göttingen: Verlag für Angewandte Psychologie.

Marcus, B., Funke, U. & Schuler, H. (1997). Integrity Tests als spezielle Gruppe eignungsdiagnostischer Verfahren: Literaturüberblick und metaanalytische Befunde zur Konstruktvalidität. *Zeitschrift für Arbeits- und Organisationspsychologie, 41*, 2-17.

Marcus, B., & Schuler, H. (1998). *Fragebogen zu Einstellungen und Selbsteinschätzungen.* Unpublished test, Psychologisches Institut, Universität Hohenheim.

Marsh, H. W. (1989). Confirmatory factor analyses of multitrait-multimethod data: Many problems and a few solutions. *Applied Psychological Measurement, 13*, 335-361.

Marsh, H. W. & Grayson, D. (1995). Latent variable models of multitrait-multimethod data. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 177-198). Thosand Oaks, CA: Sage Publications.

Maruyama, G. M. (1998). *Basics of structural equation modeling*. Thousands Oaks, CA: Sage Publications.

Maurer, S. D., Sue-Chan, C. & Latham, G. P. (1999). The Situational Interview. In R. W. Eder & M. M. Harris (Eds.), *The Employment Interview Handbook* (pp. 159-177). Thousand Oaks, CA: Sage Publications.

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L. & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology, 79*, 599-616.

McDaniel, M. A., Whetzel, D. L., Schmidt, F. O., Hunter, J. E., Maurer, S. D. & Russel, J. (1986). *The validity of employment interviews: A review and meta-analysis*. Washington DC: US Office of Personnel.

Michel, L. & Conrad, W. (1982). Theoretische Grundlagen psychometrischer Tests. In K. J. M. Groffmann (Ed.), *Grundlagen psychologischer Diagnostik* (pp. 1-129). Göttingen: Hogrefe.

Mielke, R. & Kilian, R. (1990). Wenn Teilskalen sich nicht zu dem ergänzen, was die Gesamtskala erfassen soll: Untersuchungen zum Self-Monitoring-Konzept. *Zeitschrift für Sozialpsychologie, 21*, 126-135.

Moser, K., Diemand, A. & Schuler, H. (1996). Inkonsistenz und Soziale Fertigkeiten als zwei Komponenten von Self-Monitoring. *Diagnostica, 42*, 268-283.

Moses, J. L. & Ritchie, R. J. (1976). Supervisory relationships training: A behavioral evaluation of a behavior modeling program. *Personnel Psychology, 29*, 337-343.

Motowidlo, S. J., Carter, G. W., Dunnette, M. D. & Tippins, N. (1992). Studies of the structured behavioral interview. *Journal of Applied Psychology, 77*, 571-587.

Mucha, D., & Waldeyer, K. (2002). *Kann Time Discounting Zeitmanagement-Probleme erklären? Ein Überprüfung im Rahmen von Assessment Center Bewerbungstrainings.* Unpublished diploma thesis, Philipps-Universität, Marburg.

Mumford, M. D., Costanza, D. P., Connelly, M. S. & Johnson, J. F. (1996). Item generation procedures and background data scales: Implications for construct and criterion-related validity. *Personnel Psychology, 49*, 361-398.

Ones, D. S., Viswesvaran, C. & Schmidt, F. L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology, 78*, 679-703.

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist, 17*, 776-783.

Petty, M. M. (1974). A multivariate analysis of the effects of experience and training upon performance in a leaderless group discussion. *Personnel Psychology, 27*, 271-282.

Posthuma, R. A., Morgeson, F. P. & Campion, M. A. (2002). Beyond employment interview validity: A comprehensive narrative review of recent research and trends over time. *Personnel Psychology, 55*, 1-81.

Pulakos, E. D. & Schmitt, N. (1995). Experience-based and situational interview questions: Studies of validity. *Personnel Psychology, 48*, 289-308.

Reilly, R. R. & Chao, G. R. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology, 35*, 1-62.

Richter, G. & Kleinmann, M. (2002). *Personenmerkmale und Konstruktvalidität im Multimodalen Interview.* Paper presented at the 43. Kongress der Deutschen Gesellschaft für Psychologie, Berlin.

Robertson, I., Gratton, L. & Sharpley, D. (1987). The psychometric properties and design of managerial assessment centres: Dimensions into exercises won't go. *Journal of Occupational Psychology, 60*, 187-195.

Ryan, A. M., McFarland, L., Baron, H. & Page, R. (1999). An international look at selection practices: Nation and culture as explanations for variability in practice. *Personnel Psychology, 52*, 359-391.

Sackett, P. R. (1987). Assessment centers and content validity: Some neglected issues. *Personnel Psychology, 40*, 13-25.

Sackett, P. R. & Dreher, G. F. (1982). Constructs and assessment center dimensions: Some troubling empirical findings. *Journal of Applied Psychology, 67*, 401-410.

Sackett, P. R., Zedeck, S. & Fogli, L. (1988). Relations between measures of typical and maximum job performance. *Journal of Applied Psychology, 73*, 482-486.

Sagie, A. & Magnezy, R. (1997). Assessor type, number of distinguishable dimension categories, and assessment centre construct validity. *Journal of Occupational and Organizational Psychology, 70*, 103-108.

Salgado, J. F. & Moscoso, S. (2002). Comprehensive meta-analysis of the construct validity of the employment interview. *European Journal of Work and Organizational Psychology, 11*, 299-324.

Schleicher, D. J., Day, D. V., Mayes, B. T., & Riggio, R. E. (2002). A new frame for frame-of-reference training: Enhancing the construct validity of assessment centers. *Journal of Applied Psychology, 87*, 735-746.

Schmidt, F. L. & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262-274.

Schmidt, F. L. & Rader, M. (1999). Exploring the boundary conditions for interview validity: Meta-analytic validity findings for a new interview type. *Personnel Psychology, 52*, 445-464.

Schmitt, N. (1976). Social and situational determinants of interview decisions: Implications for the employment interview. *Personnel Psychology, 29*, 79-101.

Scholz, G. & Schuler, H. (1993). Das nomologische Netzwerk des Assessment Centers: Eine Metaanalyse. *Zeitschrift für Arbeits- und Organisationspsychologie, 37*, 73-85.

Schuler, H. (1989). Construct validity of a multimodal employment interview. In B. J. Fallon & H. P. Pfister & J. Brebner (Eds.), *Advances in industrial organizational psychology* (pp. 343-354). Amsterdamm: North-Holland.

Schuler, H. (1992). Das Multimodale Einstellungsinterview. *Diagnostica, 38*, 281-300.

Schuler, H. (1993). Social validity of selection situations: A concept and some empirical results. In H. Schuler & J. L. Farr (Eds.), *Personnel selection and assessment: Individual and organizational perspectives. Series in applied psychology* (pp. 11-26). Hillsdale, NJ: Lawrence Erlbaum Associates.

Schuler, H. (1996). *Psychologische Personalauswahl. Einführung in die Berufseignungsdiagnostik*. Göttingen: Verlag für Angewandte Psychologie.

Schuler, H. (1999). Auswahl von Gründungsunternehmern mittels Interview - Ein Erfahrungsbericht. In K. Moser & B. Batinic & J. Zempel (Eds.), *Unternehmerisch erfolgreiches Handeln* (pp. 145-153). Göttingen: Verlag für Angewandte Psychologie.

Schuler, H. (2002). *Das Einstellungsinterview*. Göttingen: Hogrefe.

Schuler, H., Frier, D. & Kauffmann, M. (1993). *Personalauswahl im europäischen Vergleich*. Göttingen: Verlag für Angewandte Psychologie.

Schuler, H. & Fruhner, R. (1993). Effects of assessment center participation on self-esteem and on evaluation of the selection situation. In H. Schuler & J. L. Farr & M. Smith (Eds.), *Personnel selection and assessment. Individual and organizational perspectives* (pp. 109-124). Hillsdale: Erlbaum.

Schuler, H. & Funke, U. (1989). The interview as a multimodal procedure. In R. W. Eder & G. R. Ferris (Eds.), *The employment interview: Theory, research, and practice* (pp. 183-192). Newbury Park, CA: Sage Publications.

Schuler, H., Funke, U., Moser, K., Donat, M. & Barthelme, D. (1995). *Personalauswahl in Forschung und Entwicklung - Eignung und Leistung von Wissenschaftlern und Ingenieuren*. Göttingen: Hogrefe.

Schuler, H. & Moser, K. (1995). Die Validität des Multimodalen Interviews. *Zeitschrift für Arbeits- und Organisationspsychologie, 39*, 2-12.

Schuler, H., Moser, K., Diemand, A. & Funke, U. (1995). Validität eines Einstellungsinterviews zur Prognose des Ausbildungserfolgs. *Zeitschrift für Pädagogische Psychologie, 9*, 45-54.

Schuler, H. & Rolfs, H. (2000). Hohenheimer Gründerdiagnose: Konzeption zur eignungsdiagnostischen Untersuchung potenzieller Unternehmensgründer. In G. F. Müller (Ed.), *Existenzgründung und unternehmerisches Handeln: Forschung und Förderung* (pp. 55-73). Landau: Verlag Empirische Pädagogik.

Schulz, C., Schuler, H. & Stehle, W. (1985). Die Verwendung eignungsdiagnostischer Methoden in deutschen Unternehmen. In H. Schuler & W. Stehle (Eds.), *Organisationspsychologie und Unternehmenspraxis: Perspektiven der Kooperation* (pp. 126-123). Stuttgart: Verlag für Angewandte Psychologie.

Silverman, W. H., Dalessio, A., Woods, S. B. & Johnson, R. L. (1986). Influence of assessment center methods on assessors' ratings. *Personnel Psychology, 39*, 565-578.

Smith-Jentsch, K. A., Salas, E. & Brannick, M. T. (2001). To transfer or not to transfer? Investigating the combined effects of trainee characteristics, team leader support, and team climate. *Journal of Applied Psychology, 86*, 279-292.

Spychalski, A. C., Quinones, M. A., Gaugler, B. B. & Pohley, K. (1997). A survey of assessment center practices in organizations in the United States. *Personnel Psychology, 50*, 71-90.

Stahl, G. K. (1995). Ein strukturiertes Auswahlinterview für den Auslandseinsatz. *Zeitschrift für Arbeits- und Organisationspsychologie, 39*, 84-90.

Taylor, P. J. & Small, B. (2000). *A meta-analytic comparison of situational and behavioral description interview questions.* Paper presented at the 15th Conference of the Society for Industrial and Organizational Psychology, New Orleans.

Terpstra, D. E. & Rozell, E. J. (1997). Why some potentially effective staffing practices are seldom used. *Public Personnel Management, 26*, 483-495.

Tomás, J. M., Hontangas, P. M. & Oliver, A. (2000). Linear confirmatory factor models to evaluate multitrait-multimethod matrices: The effects of number of indicators and correlation among methods. *Multivariate Behavioral Research, 35*, 469-499.

van der Zee, K. I., Bakker, A. B. & Bakker, P. (2002). Why are structured interviews so rarely used in personnel selection? *Journal of Applied Psychology, 87*, 176-184.

Wanous, J. P. (1978). Realistic job previews: Can a procedure to reduce turnover also influence the relationship between abilities and performance? *Personnel Psychology, 31*, 249-258.

Weekley, J. A. & Gier, J. A. (1987). Reliability and validity of the situational interview for a sales position. *Journal of Applied Psychology, 72*, 484-487.

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement, 9*, 1-26.

Wiesner, W. H. & Cronshaw, S. F. (1988). A meta-analytic investigation of the impact of interview format and degree of structure on the validity of the employment interview. *Journal of Occupational Psychology, 61*, 275-290.

Woehr, D. J. & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67*, 189-205.

Wright, P. M., Lichtenfels, P. A. & Pursell, E. D. (1989). The structured interview: Additional studies and a meta-analysis. *Journal of Occupational Psychology, 62*, 191-199.