Konstruktvalidität und Assessment Center. Ein empirischer Beitrag.

Diplomarbeit

für die Zulassung zur Diplomprüfung
im Studiengang Psychologie des
Fachbereichs Psychologie der Universität Hamburg

Erster Prüfer: Prof. Dr. Heinrich Berbalk
Zweiter Prüfer: Prof. Dr. Reinhold Schwab

FB Psychologie
Klassifikation: 844 Organisationspsychologie

vorgelegt von: Kristof Kupka
Thorner Straße 42
21339 Lüneburg

Hamburg, den 09.12.1999

Inhaltsverzeichnis

1 Einleitung

1.1 Bedeutung des Themas und Ziel der Arbeit

„An den Streß wird sich Dieter S. wohl noch lange erinnern. Der Diplomvolkswirt hatte sich bei einer großen Wirtschaftprüfungsgesellschaft beworben und war daraufhin zum ‚Recruiting Day´ eingeladen worden" (Schwertfeger, 1999). Wie Dieter S. ergeht es mittlerweile vielen Stellensuchenden, die sich bei großen Unternehmen bewerben.
Um möglichst kompetente Mitarbeiter[1] zu rekrutieren, setzen Firmen weltweit in zunehmenden Maße Assessment Center ein. Warum? Der Einsatz von Mitarbeitern, die sich nicht nur fachlich, sondern auch hinsichtlich der sogenannten „weichen Faktoren" wie z.B. Team- und Kommunikationsfähigkeit auszeichnen, ist für den Erfolg von zunehmender Bedeutung. Der Vorsitzende des Arbeitskreises Assessment Center e.V. Leiter (1996) beschreibt den gemeinsamen Nenner vieler Diskussionen rund um die verschiedenen Aspekte des Assessment Centers so: „Our business is people ... and human potential!" (S.9).

Wie erfährt nun aber ein Unternehmen, ob Bewerber kompetent sind? Mit welchen Instrumenten lassen sich Aussagen über Bewerber und interne Nachwuchskräfte machen? Eines der bekanntesten Verfahren der Personalauswahl und Personalentwicklung ist das Assessment Center. Die Bedeutung des Verfahrens beschreiben die renommierten Organisationspsychologen Thornton und Byham (1982) folgendermaßen: „In our opinion, assessment centers are one of the two or three major developments in the field of personnel psychology in the last 25 years" (S. xi).

Der hohe Aufwand, der für ein derartiges Verfahren aufgebracht wird, muß sich aber auch lohnen. Daher ist zur Überprüfung der Qualität von Assessment Centern in den letzten Jahren einige Forschungsanstrengung unternommen worden. Bei der Betrachtung dieser Untersuchungen läßt sich feststellen, daß sich die meisten Studien mit Aspekten der prognostischen Validität beschäftigt haben. In jüngster Zeit, spätestens seit der Publikation von Sackett und Dreher (1982), hat sich jedoch eine rege Diskussion um die Konstruktvalidität entwickelt, also der Frage danach, was genau ein Assessment Center mißt. Sackett und Dreher (1982) fanden heraus, daß die Beobachter nicht die postulierten Dimensionen (Personenmerkmale) bewerteten, sondern Pauschalurteile für eine bestimmte Übung gaben. Für diese scheinbar „niederschmetternden" Ergebnisse der Konstruktvalidität des Assessment Centers wird seitdem nach Erklärungen gesucht (Obermann, 1992). Neueste Studien

[1] Der Einfachheit und Übersichtlichkeit halber werden in dieser Arbeit nur die männlichen Begriffsformen verwendet. Diese sollen vom Sinngehalt die weiblichen stets mit einschließen.

von Kleinmann (1997) und Guldin und Schuler (1997) haben erstmals herausgefunden, daß auch Dimensionsfaktoren[2] den Bewertungen der Beobachter zugrunde lagen. Die z.T. kontroversen Ergebnisse zur Frage, was im Assessment Center bewertet wird, lassen vermuten, daß hierzu weiterhin großer Forschungsbedarf besteht. Die 1989 von Schuler aufgestellte Forderung scheint von aktueller Bedeutung zu sein: „Um die [Assessment Center] Methode wesentlich zu verbessern, werden wir um forcierte Konstruktaufklärung nicht herumkommen" (S. 242).
Der Nachweis des Einflusses von Dimensionsbewertungen zur Erklärung der Verhaltensvarianz, die anhand von Daten aus Assessment Centern der Wirtschaft überprüft wurden, steht noch aus. Diese Arbeit stellt einen Versuch dar, diese Lücke zu schließen. Es soll ein empirischer Beitrag zur Diskussion um die Konstruktvalidität geleistet werden. Dabei werden die Ergebnisse der Forschung in bezug auf angemessene Analyseverfahren berücksichtigt.

1.2 Zusammenfassung

Gegenstand dieser Arbeit ist die Untersuchung der Konstruktvalidität von Assessment Centern. Es wird der Fragestellung nachgegangen, was in Assessment Centern gemessen wird, Dimensionen oder Übungen. Dazu wurde ein Auswahlverfahren eines deutschen Großunternehmens überprüft.

Zu Beginn werden der theoretische Hintergrund von Assessment Centern, die historische Entwicklung, die Anwendungsgebiete und die bisherige Forschung skizziert. Dabei wird neben der Darstellung kontextrelevanter Informationen über Assessment Center Verfahren, insbesondere der Forschungsstand zur Interrater-Reliabilität und Konstruktvalidität, vorgestellt.

Es folgt eine ausführliche Beschreibung des untersuchten Assessment Centers, wobei die einzelnen Übungen des Verfahrens erklärt werden. Über die Datenerhebung mit Hilfe von Beurteilungsbögen und die verschiedenen Teilnehmergruppen des Assessment Centers wird informiert. Besondere Gewichtung erfährt im weiteren die Darstellung der Analyseverfahren, die im Rahmen dieser Arbeit angewendet werden. Damit der aktuellen Diskussion um angemessene Verfahren der Bestimmung der Konstruktvalidität Rechnung getragen wird, sollen hier die drei gängigsten Analyseverfahren berücksichtigt werden:

- Analyse der Multitrait-Multimethod-Matrix nach Campbell und Fiske (1959)
- Hauptkomponentenanalyse
- konfirmatorische Faktorenanalyse mit Hilfe von LISREL (Jöreskog & Sörbom, 1989)

[2] Unter (Assessment Center) Dimensionen sind in dieser Arbeit die übergeordneten Personenmerkmale (z.B. Entscheidungsfähigkeit, Kreativität) zu verstehen. Im dritten Kapitel werden die hier verwendeten Assessment Center Begriffe näher definiert.

Abgeleitet aus den Konstruktionsprinzipien von Assessment Centern, soll die Hypothese überprüft werden, ob Beobachter Personenmerkmale in verschiedenen Übungen bewerten. Darüberhinaus soll die Beobachter-Übereinstimmung untersucht werden. Dafür werden die Daten von N = 317 Kandidaten analysiert, die im Zeitraum von Januar 1996 bis Oktober 1998 an dem zweitägigen Assessment Center teilnahmen. Die Bewerber wurden von insgesamt 51 Beobachtern in acht verschiedenen Übungen bewertet.

Die Auswertung erfolgt in drei Schritten, zu Beginn werden die Ergebnisse zur Interrater-Reliabilität des Verfahrens und anschließend die Ergebnisse zur Konstruktvalidität vorgestellt. Am Ende findet sich die Überprüfung der Hypothesen.

Die Ergebnisse zur Interrater-Reliabilität liegen im erwarteten Bereich. Das bedeutet, daß die hier ermittelten Korrelationskoeffizienten (Spannweite von r = 0.33 bis r = 0.76) vergleichbar mit den Ergebnissen anderen Untersuchungen sind (vgl. Borman, 1982; Jones, 1981; Lammers, 1992; Scholz, 1994). Die Höhe der Korrelationen weist insgesamt auf eine befriedigende, wenngleich z.T. niedrige Beobachter-Übereinstimmung hin.

Zur Konstruktvalidität zeigt sich, daß die Analyse der MTMM-Matrix und die Hauptkomponentenanalyse die Ergebnisse der klassischen Untersuchung von Sackett und Dreher (1982) bestätigen. Auch die konfirmatorischen Faktorenanalyse mit Hilfe von LISREL kann – entgegen der Erwartung – den vermuteten Einfluß von Personenmerkmalen (Dimensionen) auf die Bewertung der Kandidaten nicht ermitteln. Im Gegenteil, die Ergebnisse zeigen deutlich, daß Übungsfaktoren die Daten besser repräsentieren. In dem untersuchten Assessment Center werden somit Übungen und nicht Personenmerkmale gemessen.

Die Analyse der Daten weist außerdem daraufhin, daß der Übungstyp entscheidend die Bewertungen der Beobachter beeinflußt. Dieser Zusammenhang und die Ergebnisse zur Interrater-Reliabilität und Konstruktvalidität werden abschließend diskutiert.

2 Theoretischer Hintergrund, Historische Entwicklung, Stand der Forschung

Diese Arbeit will einen empirischen Beitrag zu der Frage leisten, was in Assessment Centern gemessen wird. Das untersuchte Verfahren ist ein Bewerberauswahlverfahren zur Rekrutierung von selbständig arbeitenden Partnern eines Großunternehmens. Zur Selektion dieser Bewerber setzt das Unternehmen in Deutschland ein Assessment Center Verfahren ein. Bevor jedoch näher auf die Struktur und Gestaltung des Verfahrens eingegangen wird (Abschnitt 3.1), soll ein Überblick über das zweite Kapitel gegeben werden.

In diesem Kapitel wird als erstes der Begriff Assessment Center definiert. Die Assessment Center Methode soll im weiteren hinsichtlich wesentlicher Merkmale vorgestellt werden. Neben Anwendungsgebieten von Assessment Centern und den damit verbundenen Zielsetzungen wird im Anschluß ein geschichtlicher Abriß der Entwicklung von Assessment Centern skizziert. Der Abschnitt „Forschung zu Assessment Center Verfahren" beschreibt den Stand der wissenschaftlichen Untersuchungen und beleuchtet insbesondere die Forschung zur Interrater-Reliabilität und Konstruktvalidität.

2.1 Begriff, Definition und typischer Ablauf

Übersetzt ins Deutsche bedeutet „assessment" Festsetzung, Beurteilung, Einschätzung. Unter „Assessment Center" ist somit der Ort der Einschätzung zu verstehen. Jeserich (1995) weist darauf hin, daß korrekterweise von der „Assessment Center-Methode" oder dem „Assessment Center-Verfahren" zu sprechen wäre, sich jedoch der kürzere Begriff durchgesetzt habe. Der Begriff „Assessment Center" wurde von dem in den 30er Jahren tätigen Persönlichkeitsforscher Murray geprägt (Domsch & Jochum, 1989). Dieser Terminus ist im deutschsprachigen Raum zeitweise verworfen worden. Eingedeutschte Bezeichnungen fanden jedoch keine breite Zustimmung, so daß im allgemeinen am amerikanischen Ausdruck festgehalten wird (Schuler, 1987).

Nach Kleinmann (1997) hat sich im deutschsprachigen Raum die Definition für Assessment Center von Jeserich (1981) durchgesetzt. Danach versteht Jeserich (1981) unter der Assessment Center Methode „ein systematisches Verfahren zur qualifizierten Feststellung von Verhaltensleistungen bzw. Verhaltensdefiziten, das von mehreren Beobachtern gleichzeitig für mehrere Teilnehmer in bezug auf vorher definierte Anforderungen angewandt wird" (S. 33-34). Diese Definition ist an die amerikanische Begriffsbestimmung von 1980 der *„Task Force on Assessment Center Standards*", einer Gruppe aus Praktikern und Wissenschaftlern, angelehnt (vgl. Scholz, 1994).

Trotz der allgemeinen Verbreitung des Begriffs Assessment Center benennen viele Unternehmen ihr Verfahren entsprechend des Einsatzbereichs mit Bezeichnungen wie z.B. Beurteilungsseminar, Mitarbeiterentwicklungsprogramm, Analyse-Entwicklungs-Center, Workshop zur Karriereentwicklung (vgl. Kompa, 1989; Obermann, 1992). Die Gestaltungsformen von Assessment Centern sind ebenfalls vielfältig. Kompa (1989) behauptet, daß „nur mit Vorbehalt von *der* [Hervorhebung im Original; Anm. d. Verf.] Praxis von ACs gesprochen werden kann" (S. 27).
Unter Assessment Center ist also kein einzelnes, konkretes Verfahren zu verstehen, sondern eher ein Sammelbegriff für eine bestimmte diagnostische Vorgehensweise. Die Anzahl der Beobachter und Teilnehmer, die Art der Bewertungsfindung, die verwendeten Übungen, die Dauer, die Rotationstechnik und die Anzahl und Art der Dimensionen kann variieren (Kleinmann, 1997). Auf die unterschiedlichen Anwendungsgebiete von Assessment Centern wird in Abschnitt 2.1.1 näher eingegangen.
Was ist also das gemeinsame an den Verfahren, die sich unter dem Begriff Assessment Center zusammenfassen lassen? Mit ihren vier Charakteristika des Verfahrens beschreiben Fisseni und Fennekels (1995) die Grundprinzipien der Assessment Center Methode.

- *Anforderungsnähe*: Die Leistungen, die in den Übungen eines Assessment Centers abgerufen werden, sind den Anforderungen der Zielposition möglichst ähnlich.
- *Beobachtungsnähe*: Die Übungssituationen sind so zu konstruieren, daß sie leicht zu beobachten sind. (Man kann auch von Verhaltensnähe sprechen.
- *Verfahrensvielfalt*: Die Anforderungen der Zielposition werden in möglichst unterschiedlichen Übungen, also möglichst unterschiedlichen Verfahren, abgebildet.
- *Beobachtervielfalt*: Das Verhalten, das die Teilnehmer in den Übungen zeigen, wird von mehreren Beobachtern erfaßt.
  (Fisseni & Fennekels, 1995, S.15)

Eine weitere Gemeinsamkeit ist, daß den meisten Assessment Centern eine bestimmte zeitliche Struktur zu Grunde liegt. Jeserich (1981) skizziert einen solchen typischen Ablauf, der in der Abbildung 2.1 dokumentiert ist.

Die Abbildung 2.1 zeigt, daß sich ein Assessment Center im Grunde aus drei Phasen zusammensetzt: Vorbereitung, Durchführung und Abschluß und Feedback. Jeserich (1981) unterteilt diese Phasen wiederum in jeweils fünf Bausteine, die aufeinander folgen. Auf die einzelnen Bausteine eines Assessment Centers soll hier nicht weiter eingegangen werden; sie werden im Rahmen der Beschreibung des untersuchten Verfahrens (Abschnitt 3.1) genauer vorgestellt.

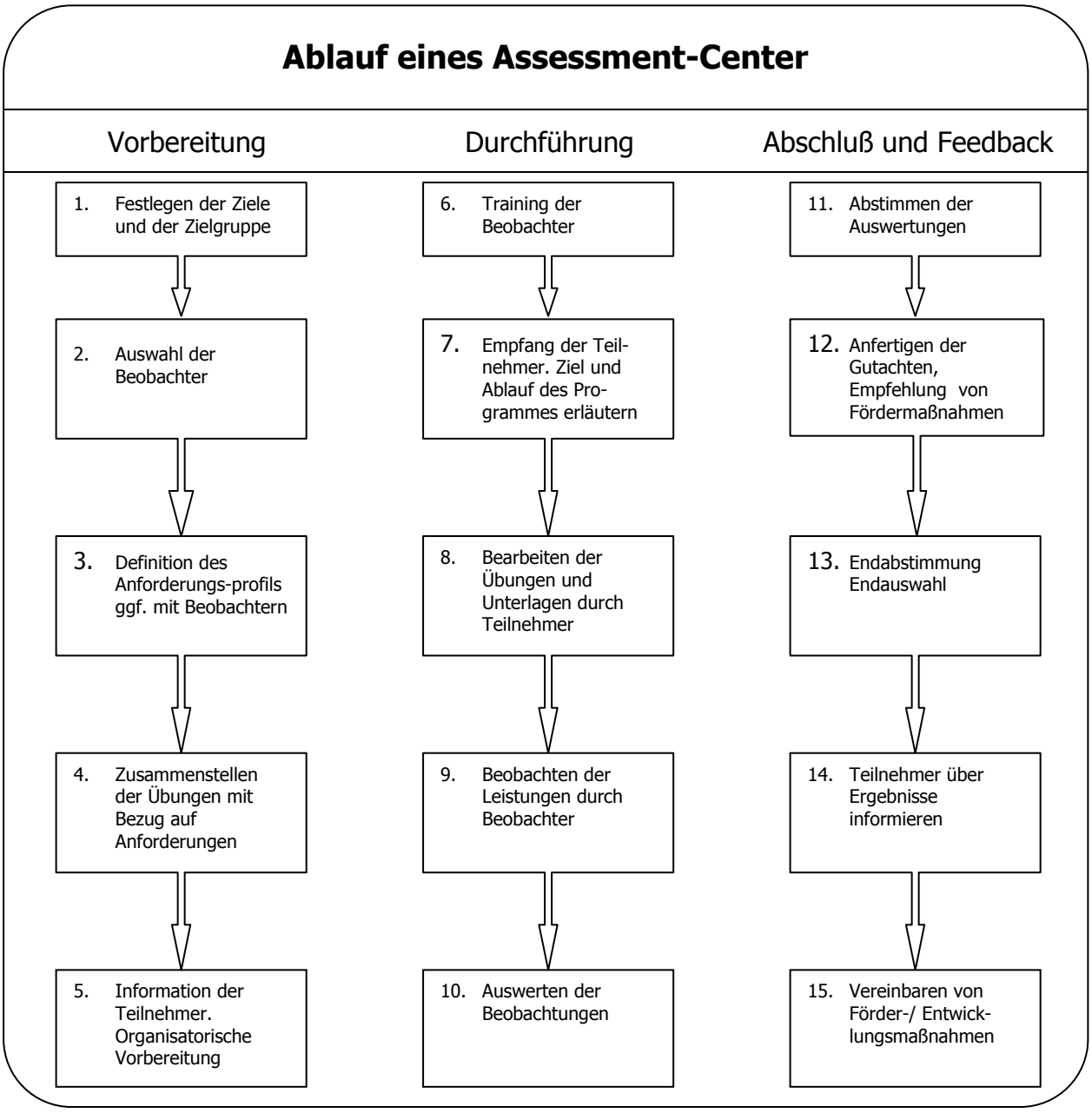


Abbildung 2.1: typischer Ablauf eines Assessment Centers (Jeserich, 1981, S.35)

2.2 Historische Entwicklung

Bereits in der Antike lassen sich erste Vorläufer diagnostischer Methoden finden. Schuler und Moser (1995) beschreiben umfangreiche Prüfungen der Tauglichkeit von Bediensteten im chinesischen Kaiserreich des Jahres 1000 v.Chr. Um in den Beamtenstatus aufzusteigen, mußten Bewerber verschiedene Einzelübungen wie Reiten, Bogenschießen und Arithmetik absolvieren.

Eine systematische Vorgehensweise, Personen hinsichtlich ihrer Eignung zu testen, findet sich aber erst zu Beginn des 20. Jahrhunderts (Obermann, 1992).
Nachfolgend soll der Entwicklungsprozeß des Verfahrens in militärischen Organisationen einerseits und in der Wirtschaft andererseits skizziert werden.

2.2.1 Entwicklung im militärischen Kontext

„Als Vorbild des heutigen AC gilt allgemein das Offiziersauswahlverfahren der deutschen Wehrmacht , das Mitte der 20er Jahre entwickelt und bis zum Ende des 2. Weltkrieges als Regelform der Auswahl eingesetzt wurde" (Domsch & Jochum, 1989, S.1). Dieses Verfahren soll nun näher vorgestellt werden, bevor im weiteren auf die Entwicklung des Verfahrens in England und Amerika eingegangen wird.

Seit dem 1. Weltkrieg wurde mit Hilfe der sogenannten Psychotechnik die Spezialistenauswahl von Kraftfahrern, Funkern und Flugzeugführern getroffen. Inhalt dieser Untersuchungen waren v.a. gut beobachtbare und meßbare Reaktionen. Diese psychotechnische Spezialprüfung erwies sich bald als unzureichend, so daß für die Auswahl von Offiziersanwärtern eine „charakterologische Komplexprüfung" eingeführt wurde (Simoneit, 1933). In diesem Zusammenhang gründete der Berliner Professor J.B. Rieffert im Auftrag des Reichsministeriums ein psychologisches Forschungszentrum an der Universität, das in psychologischen Stationen die Prüfung durchführen ließ (Domsch & Jochum, 1989).
Simoneit, der spätere Amtsnachfolger von Rieffert, beschreibt 1933 den Grundgedanken dieser neuen „Auslese" von Personen: „Es wird hierbei versucht, praktische Menschenkenntnis zu objektivieren. Dabei muß beobachtet werden, welche Methoden im praktischen Leben angewandt werden und welche Gedankenreihen zu praktisch brauchbaren Menschen-Erkenntnissen führen" (Simoneit, 1933, S. 43).

Die gesamte Veranstaltung, die als Vorbild heutiger Assessment Center gewertet wird (vgl. Domsch & Jochum, 1989; Schuler & Moser, 1995; Obermann, 1992), dauerte drei Tage lang. Zwei Gruppen von Offiziersanwärtern wurden von einem Auswahlgremium von vier bis sechs Beobachtern beurteilt. Dieses Gremium setzte sich aus Offizieren, Mitgliedern der Prüfstelle sowie Psychologen und Psychiatern (Sanitätsoffizieren) zusammen (Jeserich, 1981). Eine abschließende Beurteiler-Konferenz entschied über die Gesamtergebnisse, wobei die endgültige Bewertung dem Prüfungsoffizier oblag (Obermann 1992).

Die Prüfung war nach Simoneit (1933) in vier Elemente unterteilt:
- Lebenslaufanalyse: Daten, die sich auf die seelische und geistige Entwicklung der Offiziersanwärter ausgewirkt haben könnten, wurden in diesem Rahmen analysiert.
- Ausdrucksanalyse: Hierbei wurden Mimik und Gestik sowie sprachlicher und schriftlicher Ausdruck bewertet.

- Geistesanalyse: Mit Hilfe einer Aufgabenreihe und Explorationstechniken testete das Auswahlgremium in diesem Zusammenhang „Intelligenzgrad", „Denkantrieb", „Denkrichtung" und „Denkmethoden".
- Handlungsanalyse: Innerhalb der Handlungsanalyse mußten die Offiziersanwärter eine Reihe von Übungen durchlaufen, wie die „Befehlsreihe", die „Führerprobe" und das „Schlußkolloquium". Eine Reaktionsprüfung, während der das reaktive Handeln an Apparaten getestet wurde, komplettierte die Handlungsanalyse.

In der Übung „Befehlsreihe" (innerhalb der Handlungsanalyse) wurden die Prüflinge instruiert, einige Befehle auszuführen. Die Ausgestaltung blieb jedoch offen, so daß spontanes Handeln überprüft werden konnte. Die „Führerprobe" testete Führungsqualitäten und die Gesamtpersönlichkeit im Rahmen einer gestellten Situation, dabei wurden den Offiziersanwärtern Soldaten unterstellt, mit denen sie Aufgaben zu bewältigen hatten. Das „Schlußkolloquium" bestand aus einer kontroversen Gruppendiskussion, während der das Verhalten der Anwärter und ihre Persönlichkeit abschließend eingeschätzt wurden. Insgesamt sollte bei der Handlungsanalyse die Willensseite der Person getestet werden, während bei den ersten drei Analysen eher die unwillkürliche, unbewußte Seite der Prüflinge untersucht wurde (Simoneit, 1933).

Das Besondere an der geschilderten Veranstaltung ist, daß sie schon damals die Grundprinzipien eines heutigen Assessment Centers verfolgte, wie
- Prüfung in Gruppen durch Gruppen
- Vielfalt der Methoden
- Trennung von Beobachtung und Beurteilung
- Einsatz situativer Elemente
  (vgl. Jeserich, 1981; Obermann, 1992).

Die damaligen Prüfungen "Befehlsreihe" und "Schlußkolloquium" finden sich in vergleichbarer Art in heutigen Assessment Centern als Bausteine "Postkorb" und "führerlose Gruppendiskussion" wieder (Domsch & Jochum, 1989).
Auch Thornton und Byham (1982) betonen, daß viele Aspekte heutiger Veranstaltungen letztlich auf diese Quelle zurückgehen, wie z.B. der Einsatz von mehreren Beobachtern, die Anwendung von komplexen Tests und v.a. der Betonung der Verhaltensbeobachtung.
Auf Probleme der Validitätsprüfung dieses Verfahrens weisen Domsch und Jochum (1989) hin: „Die von ihm [Simoneit] 1954 veröffentlichten Validitätsangaben erschienen zu spät und waren zu spärlich" (S. 6/7).

Die Teilnahme an diesem Auswahlverfahren wurde ab 1927 für alle Anwärter auf den Offiziersgrad Voraussetzung. In den folgenden Jahren mit Beginn des Nationalsozialismus in Deutschland wurde diese Methode in der Reichswehr jedoch

schrittweise wieder abgeschafft. Jeserich (1981) nennt als Grund für das Scheitern des Verfahrens die Diskrepanz zwischen der demokratischen Leitidee der Methode einerseits und der Parteiideologie des Nationalsozialismus sowie der alten preußischen Offizierstradition andererseits. Erst 1957 setzte die Bundeswehr wieder ein Assessment Center Verfahren ein.

Die Entwicklung der Methode außerhalb Deutschlands begann während des Zweiten Weltkrieges in Großbritannien auf Initiative des britischen Militärattachés in Berlin, der das Verfahren aus Deutschland übernahm. Erste positive Erfahrungen bei der Pilotenauswahl führten zur Einführung dieses neuen Verfahrens zur Offiziersauswahl in Großbritannien. In diesem Zusammenhang wurden sogenannte War Office Selection Boards (WOSB) gegründet, die zwei Zielen dienen sollten: Militärische Moral während der Krisensituation des Krieges festigen und ausreichend qualifizierten Führungsnachwuchs bereitstellen (Thornton & Byham, 1982). 100.000 Prüfungen wurden in den ersten drei Jahren abgehalten (Morris, 1949 in Domsch & Jochum, 1989).

Das britische Verfahren stellte eine Weiterentwicklung der deutschen Methode dar mit der stärkeren Gewichtung von situativen, führerlosen Gruppentests und -diskussionen sowie der Konzeption einer Führungstheorie (Domsch & Jochum, 1989). Während der 3-4 tägigen Veranstaltung wurden die Gruppen von acht Teilnehmern von einem Team beobachtet und bewertet. Dem Team von Offizieren, einem Psychiater und einem Psychologen stand ein Präsident (Oberst) vor, der die Entscheidungsgewalt inne hatte. In Gruppen- und Einzeltests sowie individuellen Interviews wurden soziale und kognitive Fähigkeiten beobachtet und bewertet (Domsch & Jochum, 1989).

Die positiven Ergebnisse der Validitätsstudien von Vernon und Parry (1949 in Jeserich, 1981) und Morris (1949, in Domsch & Jochum, 1989) - einem ehemaligen Mitglied des psychologischen Forschungsstabs - unterstützten letztlich auch die Verbreitung dieses Verfahrens in anderen Länder des Commonwealth wie Kanada und Australien Anfang der vierziger Jahre (Jeserich, 1981).

Die Erfahrungen der WOSBs wurden nach Beendigung des Krieges auf sogenannte Civil Sevice Selection Boards zur Auswahl von Mitarbeitern für den Öffentlichen Dienst weitergegeben (Domsch & Jochum, 1989).
Der entscheidende Beitrag der Briten an der Entwicklung der Assessment Center Methode liegt nach Thornton und Byham (1982) in dem Entwurf einer sozialpsychologischen Führungstheorie und der Einführung der ersten Validitätsstudien.

Die Entwicklung der Assessment Center Methode in den USA wurde durch das Office of Strategic Services (OSS) - einer Vorläuferorganisation der CIA - geprägt (Schuler & Moser, 1995).

Ziel war es, qualifizierter als bisher Geheimdienstagenten auszuwählen (Jeserich, 1995). Robert Tyron - ein Abteilungsleiter in Diensten der OSS - gründete 1943 eine Arbeitsgruppe, die in kürzester Zeit ein Assessment Center nach Vorbild des britischen Verfahrens entwickelte und durchführte (MacKinnon, 1977). Mitglied dieser Arbeitsgruppe war auch Prof. Murray, dem die Entwicklung des Terminus "Assessment Center" zugeschrieben wird (s.o.). Murray soll nach MacKinnon (1977) auch entscheidend die Grundkonzeption des Assessment Center Verfahrens des OSS bestimmt haben. Das Auswahlverfahren beinhaltete standardisierte Tests, projektive Verfahren, Einzel- und Gruppenaufgaben, biographische und soziometrische Fragebögen und die Simulation einer Belastungssituation (Obermann, 1992). Besonderheit dieses Assessment Centers gegenüber den deutschen und britischen Verfahren war die gemeinsame Entscheidung über das Gesamtergebnis durch das Beurteiler-Gremium, das sich aus Psychologen, Psychiatern und Vertretern aus sozialwissenschaftlichen Bereichen zusammensetzte (Domsch & Jochum, 1989).
Eine deutlich andere Einschätzung der geschichtlichen Entwicklung der Assessment Center Methode liefert Kompa (1989): „Die Adaptierung des ACs durch das Office of Strategic Services gegen Ende des Zweiten Weltkriegs ist ein Paradebeispiel dafür, wie unter Zeitdruck ein pragmatisches Programm zur Auswahl von Saboteuren, Agenten und Propagandaexperten erstellt wurde, das sich aus einer bunten Mischung intuitiv begründeter Verfahrenselemente zusammensetzte" (S.24).

2.2.2. Entwicklung im wirtschaftlichen Kontext

Entscheidend für die Ausbreitung der Assessment Center Methode im Wirtschaftsbereich war die Veröffentlichung des Verfahrens des Office of Strategic Services (OSS) durch die damalige Arbeitsgruppe (vgl. Domsch & Jochum, 1989; Jeserich, 1981). Darin wurden neben den Beschreibungen der Methode auch Empfehlungen für die praktische Anwendung gegeben (MacKinnon, 1977). Bray und Grant (1966) richteten danach das erste Assessment Center im wirtschaftlichen Kontext beim amerikanischen Kommunikationsunternehmen AT&T ein. Als Vorbild diente das OSS-Verfahren, das für diesen Rahmen leicht geändert wurde (Domsch & Jochum, 1989). In der viel zitierten „Management Progress Study", einer Langzeitstudie (1956-1960), wurden die Ergebnisse von 422 Nachwuchs-führungskräften des Unternehmens untersucht, die in diesem Zeitraum das Assessment Center durchlaufen hatten (vgl. Bray, 1964; Bray & Grant, 1966).

Das ursprüngliche Ziel der zu Forschungszwecken durchgeführten Untersuchung war nach Bray (1964) allgemein gehalten. Es sollten Berufserfolgsfaktoren und Charakteristika des Entwicklungsprozesses von Managern gefunden und untersucht werden. Um dem Anspruch der Wissenschaftlichkeit der Studie gerecht zu werden, wurden während der Untersuchungsphase weder die Teilnehmer noch das Unternehmen von den Ergebnissen in Kenntnis gesetzt. Das Verfahren ging über dreieinhalb Tage und wurde mit einer Gruppengröße von zwölf Teilnehmern und neun Beobachtern durchgeführt.

Es bestand aus folgenden Komponenten:
- Interviews
- Objektive und projektive Tests
- In-Basket (Postkorbübung)
- Miniatur-Unternehmensspiel
- Gruppendiskussionen
- Papier-Bleistift-Tests und Fragebogen
- Verschiedenes (z.B. Selbstbeurteilung)
  (vgl. Bray & Grant, 1966)

Abschließend sollten die Beurteiler die Kandidaten nach umfangreicher Diskussion hinsichtlich 25 Dimensionen bewerten. Außerdem sollten sie einschätzen, ob die Teilnehmer im Laufe von zehn Jahren ins mittlere Management aufsteigen würden (Schuler & Moser, 1995). Später wurden zur Überprüfung der Vorhersagequalität des Verfahrens diese Beurteilungen mit dem tatsächlich erreichten Karriereerfolg verglichen.

Die von Bray und Grant (1966) publizierten Ergebnisse zeigten, daß das Assessment Center prognostisch valide Aussagen produzierte. Genauer wird auf die Ergebnisse der Studie in Abschnitt 2.4 eingegangen. Aus den Daten der Management-Progress-Studie ermittelten Bray und Grant (1966) auch Faktoren, die den Berufserfolg bedingten. Dazu gehörten

Administrative Fähigkeiten
Zwischenmenschliche Fähigkeiten
Kontrolle der Gefühle
Intellektuelle Fähigkeit
Arbeitsorientierte Motivation und
Passivität (Bray & Grant, 1966).

In der Diskussion um die Bedeutung dieser Studie betonte Bray (1964), daß die Anwendung von Assessment Centern als Auswahl- und Entwicklungsinstrument von Führungskräften eine der wichtigsten Auswirkungen für die Praxis war. Domsch und Jochum (1989) sahen im Schaffen von „nachweisbaren Fakten" das für die Assessment Center Methode „bahnbrechende Verdienst" der Studie.

Das erste Assessment Center, das nicht als Forschungsstudie konzipiert war, wurde dann letztlich 1958 bei der AT&T Tochterfirma Michigan Bell durchgeführt (Byham, 1970). Es dauerte aber noch einige Jahre bis sich dieses neue Instrument in der Wirtschaft richtig durchsetzen konnte (Jeserich, 1981). 1969 hatten nach Byham (1977) erst zwölf Organisationen in Amerika Assessment Center angewendet. Erst Anfang der siebziger Jahre schritt die Verbreitung von Assessment Centern in den Vereinigten Staaten rasch voran (Scholz, 1994). Finkle schätzte bereits 1976, daß es mehr als 1000 Organisationen gab, die Assessment Center durchführten.

In Deutschland fand die Verbreitung des Verfahrens in der Wirtschaft erst später statt. Zuerst führten deutsche Niederlassungen amerikanischer Firmen die Methode ein - v.a. zur Auswahl von Hochschulabsolventen. Nach Jeserich (1996) hat Simpfendörfer das erste Nachkriegs - Assessment Center im Dezember 1969 bei IBM durchgeführt. Seit 1970 verwenden BAT und IBM Assessment Center (Jeserich, 1981). Die Zahl der Unternehmen, die Assessment Center durchführen, ist seitdem stetig gestiegen. Entscheidenden Anteil an der Popularität des Verfahrens trägt der Arbeitskreis Assessment Center, der sich bereits 1977 zusammenschloß (Obermann, 1992). Vertreter verschiedener deutscher Wirtschaftsunternehmen bilden darin einen Kreis, der sich regelmäßig über Erfahrungen und neue Entwicklung der Assessment Center Methode austauscht. Nachdem 1979 beim ersten „Kongreß Assessment Center" des Arbeitskreises in Köln erst 14 Unternehmen bekannt waren, die dieses Instrument in Deutschland verwendeten, führten nach einer Umfrage des „Instituts für Qualitative Personalarbeit" 1988 fast 90 Organisationen Assessment Center durch (Jeserich, 1989). Eine Liste von 132 Anwendern, durch die auch die wachsende Bedeutung der Methode verdeutlicht wird, lieferte Obermann (1992). Jochmann (1999) konstatiert dazu: „Insgesamt hat sich das Verfahren trotz allen Aufs und Abs durchgesetzt – von den 50 größten Konzernen Deutschlands wenden über 80% das Verfahren" (S. V) an.

Es gibt auch Kritik an der Assessment Center Methode. Sarges (1996) unterscheidet dabei zwischen methodischer und ideologischer. Danach weist die methodische Kritik auf die Schwierigkeiten im Zusammenhang mit mangelnder Validität des Instruments hin, die im Rahmen der Forschung (Abschnitt 2.4) diskutiert wird. Die ideologische Kritik hingegen thematisiert die Diskrepanz zwischen der - gemessen an dem hohen Aufwand - geringen Effizienz und der hohen Beliebtheit bei Unternehmen. Auf eine weitergehende Darstellung der ideologischen Kritik wurde im Rahmen dieser Arbeit verzichtet; die Lektüre der interessanten Arbeiten von Kompa (1989) und Neuberger (1989) sei aber ausdrücklich empfohlen.

2.3 Anwendungsgebiete und Nutzen des Assessment Center Verfahrens

Die ursprüngliche Zielsetzung der Assessment Center Methode als Personalauswahl- und Förderungsinstrument ist im Laufe der Jahre erweitert worden, wobei noch heute diese beiden Funktionen im Vordergrund stehen. Thornton und Byham (1982) schätzen, daß ungefähr 95% der Assessment Center diesen beiden Kategorien zuzurechnen sind.
In der Literatur wird eine Reihe von weiteren potentiellen Anwendungsbereichen für Assessment Center genannt. Die folgende Liste nennt die wichtigsten Zielsetzungen:

- Interne Personalauswahl
- Auswahl externer Bewerber
- Laufbahnplanung
- Ausbildungsberatung
- Potentialsuche und –beratung

- Trainingsbedarfsanalyse
- Teamentwicklung
- Berufsberatung
- Erfolgskontrolle
- Arbeitsplatzgestaltung
  (vgl. Schuler, 1987; Kleinmann, 1997)

Neben diesen direkten Zielen werden in der Literatur immer häufiger auch die Vorteile von Zusatznutzen des Einsatzes von Assessment Centern beschrieben. Jeserich (1995) betont die Organisationsentwicklungswirkung des Verfahrens und die damit verbundene Möglichkeit der „Erhöhung der sozialen Kompetenz" für Beobachter und Teilnehmer. Darunter versteht Jeserich (1995) die Chance für Beobachter, mit Hilfe kompetenter Unterstützung wichtige Personalentscheidungen vorbereiten und treffen zu können sowie die Chance für Teilnehmer, ihre Stärken und Schwächen hinsichtlich eines bestimmten Anforderungsprofils einer Zielposition kennenlernen zu können. Einen Nutzen für die Gesamtorganisation sehen Obermann (1992) und Schuler (1987) in dem möglichen Beitrag eines Assessment Centers zur Entwicklung der Unternehmenskultur.

In diesem Zusammenhang weist Schuler (1987) auf weitere „latente" Funktionen der Assessment Center Technik hin, wie z.B. dem Gewinn eines Überblicks über Organisationseinheiten, Programme und Führungsstile im Unternehmen und über den Leistungsstand des Nachwuchses und der Unterstützung des Selbstverständnisses von Führungskräften, die als Beobachter fungieren. Jung und Leiter (1989) sehen Vorteile für das Unternehmen darin, daß firmeninterne Beobachter für Führungsprobleme, Führungsverhalten sowie Beobachtung und Bewertung von Mitarbeitern durch die Teilnahme am Assessment Center sensibilisiert werden. Kleinmann (1997) nennt die Sensibilisierung der Beobachter und Teilnehmer bereits eines der potentiellen Ziele des Einsatzes von Assessment Center Verfahren. Lattmann (1989) sieht in Assessment Centern eine besonders wertvolle Bereicherung des sozialen Systems eines Unternehmens aufgrund von „mittelbaren Nutzenwirkungen", wie der recht hohen sozialen Validität (siehe auch 2.4.2), der fördernden Wirkung auf das Betriebsklima und der Möglichkeit für Denkanstöße für Beobachter (Lattmann, 1989).

Bereits 1970 wurden die Zusatznutzen von Assessment Centern von Byham diskutiert. Bei weitem der wichtigste Zusatznutzen ist nach Byham (1970) das Beobachter-Training und damit die Möglichkeit für eine Führungsperson, außerhalb des täglichen Geschäfts Verhalten von Bewerbern zu beobachten und anschließend ihre Eindrücke zu diskutieren. Danach profitieren die Beobachter in soweit, als daß sie ihre Erfahrungen aus dem Assessment Center leicht auf ihren Job transferieren und ihre Fähigkeiten hinsichtlich des Führens von Interviews und Diskussionen erweitern können. Nicht nur im Nutzen für den einzelnen sieht Byham (1970) Vorteile des Verfahrens, sondern durch die Definition von Arbeitszielen und Anforderungsprofilen für Führungspersonen unterstützen Assessment Center auch die Organisationsentwicklung.

Diese Zusatznutzen nennt Byham auch als letztlich ausschlaggebend für die Überlegenheit der Assessment Center Methode gegenüber anderen Verfahren.

Lorenzo (1984) hat in einer Studie die Auswirkungen einer Assessment-Center-Teilnahme auf Beobachter untersucht. Dabei zeigte sich, daß die Manager, die bereits drei Monate auf Assessment Center Veranstaltungen als Beobachter fungiert hatten, den Führungskräften, die nicht am Assessment Center teilgenommen hatten, in wesentlichen Aspekten überlegen waren. Die erfahrenen Manager bewiesen größere Fähigkeiten hinsichtlich des Führens von Bewerberinterviews; sie konnten ihre beobachteten Eindrücke von Kandidaten in Diskussionen besser darstellen und schriftliche Beurteilungen konkreter verfassen. Außerdem waren die Beurteilungen von auf Video festgehaltenen Bewerberinterviews laut Lorenzo (1984) von höherer psychometrischer Qualität.

2.4 Forschung zu Assessment Center Verfahren

Nach einem kurzen Überblick über das Ausmaß der Forschung zu Assessment Centern in den letzten Jahren sollen in diesem Abschnitt Untersuchungen zur Reliabilität und Validität von Assessment Centern vorgestellt werden. Besondere Aufmerksamkeit genießen in diesem Zusammenhang Studien zur Interrater-Reliabilität und Konstruktvalidität. Dabei wird kein Anspruch auf Vollständigkeit erhoben. Das bedeutet, daß hier nur die Untersuchungen berücksichtigt werden, die im Rahmen dieser Arbeit sinnvoll erscheinen.

Zum Ausmaß der Forschung zu Assessment Centern hat Kleinmann (1997) eine Recherche der in den Datenbanken PSYCLIT und PSYINDEX veröffentlichten Artikel initiiert. Danach sind von 1974 bis zum Zeitpunkt der Schrift 358 Arbeiten zum Thema Assessment Center publiziert worden, wovon sich 116 mit Validität und davon 22 mit Konstruktvalidität beschäftigen. Kleinmann (1997) weist darauf hin, daß das Forschungsinteresse an Assessment Center Verfahren insgesamt zunimmt, jedoch der Themenkomplex Konstruktvalidität quantitativ wenig bedacht wird.

2.4.1 Reliabilität der Assessment Center Verfahren

Definition und Arten der Reliabilität

Die Reliabilität oder Zuverlässigkeit beschreibt den Grad der Meßgenauigkeit eines Instrumentes (Bortz & Döring, 1995). Lienert (1969) versteht unter der Zuverlässigkeit eines Tests „den Grad der Genauigkeit, mit dem er ein bestimmtes Persönlichkeits- oder Verhaltensmerkmal mißt, gleichgültig, ob er dieses Merkmal auch zu messen beansprucht" (S.14).
Zur Reliabilitätsmessung von Assessment Centern unterscheiden Hinrichs und Haanperä (1976) zwischen drei Varianten: Der Retest-Reliabilität, die die Stabilität wiederholter Messungen zu unterschiedlichen Zeitpunkten untersucht, der internen

Konsistenz, die auf der Annahme basiert, daß die Dimension „A" das gleiche in Übung 1 wie in Übung 2 bedeutet und der Interrater-Reliabilität, die den Grad der Übereinstimmung zwischen mehreren Beurteilern beleuchtet. Forschung zur Retest-Reliabilität ist nach Obermann (1992) in der Literatur sehr wenig behandelt worden und soll hier auch nur kurz skizziert werden. Moses (1973, in Obermann, 1992) verglich dabei die Bewertungen eines eintägigen und eines zweitägigen Assessment Centers nach mehr als einem Monat miteinander. Danach ergab sich ein Zusammenhang von r = 0.73.

Scholz (1994) nennt in einer Übersicht eine weitere Studie zur Retest-Reliabilität eines Assessment Centers. Danach haben McConnell und Parker (1972, in Scholz, 1994) Koeffizienten von r = 0.74 ermittelt. Die Bestimmung der Retest-Reliabilität ist nach Obermann (1992) aufgrund von möglichen Lerneffekten problematisch. Lammers (1992) betont, daß es meistens nicht möglich sei, Assessment Center mit der gleichen Besetzung zu wiederholen. Auf Untersuchungen zur inneren Konsistenz soll hier nicht näher eingegangen werden, da sie in ihrer Vorgehensweise Studien zur Konstruktvalidität entsprechen, die im nächsten Abschnitt vorgestellt werden sollen (vgl. Obermann, 1992).

Forschung zur Interrater-Reliabilität

Interrater-Reliabilitätskoeffizienten resultieren i.d.R. aus zwei möglichen Vergleichen: Es können einerseits die Gesamtbewertungen der Übungen („overall ratings") und andererseits die Dimensions-Bewertungen verschiedener Beobachter korreliert werden (vgl. Thornton & Byham, 1982). Bevor jedoch Ergebnisse der Studien zu diesen beiden Bereichen dargestellt werden, soll kurz ein weiterer Aspekt der Forschung zur Interrater-Reliabilität beleuchtet werden.
Dabei war der mögliche Einfluß eines Informationsaustausches sowie des Trainingsstandes der Beobachter auf die Höhe der Reliabilitätskoeffizienten Gegenstand einiger Untersuchungen. Schmitt (1977) und Jones (1981) fanden hierzu heraus, daß sich die Werte der Beobachter-Übereinstimmung erhöhen, wenn der Bewertung eine Aussprache vorausging. Jones (1981) berichtet von Interrater-Reliabilitäten von r = 0.65 bis 0.73 bei Bewertungen vor einem Informationsaustausch und von Korrelationen von r = 0.67 bis 0.86 bei Bewertungen nach einem Austausch. Scholz (1994) ermittelte in einer ähnlichen Studie Korrelationen von r = 0.45 bis 0.54 vor Aussprache der Beobachter und r = 0.66 bis 0.76 nach Aussprache. Richards und Jaffee (1972) haben untrainierte und trainierte Beobachter hinsichtlich der Beurteiler-Übereinstimmung verglichen. Bei untrainierten Beobachtern zeigten sich Reliabilitätskoeffizienten von r = 0.46 und r = 0.58 der Bewertungen in zwei Dimensionen, während bei trainierten Beobachtern Werte von r = 0.78 und r = 0.90 ermittelt wurden. Die Studien verdeutlichen zwei Einflußgrößen der Interrater-Reliabilität:

- Trainierte Beobachter scheinen reliabler zu bewerten als untrainierte.
- Beobachter bewerten reliabler, wenn sie sich vor der Beurteilung über die Performance der Kandidaten austauschen.

Interrater-Reliabilität der Gesamtbewertungen der Übungen

Scholz (1994) gibt in einer Übersicht Interrater-Reliabilitäten aus zwölf verschiedenen Studien wieder. Diese liegen bei Gesamtbewertungen auf Aufgabenebene zwischen r = 0.43 und 0.95 vor einer Aussprache der Beobachter und bei r = 0.67 bis 0.99 bei Bewertungen nach einem Informationsaustausch. Im Rahmen der ersten Validitätsstudien zu Assessment Center Verfahren wurden auch Reliabilitätskoeffizienten bestimmt. Bray und Grant (1966) fanden bei der bereits zitierten AT&T-Studie Reliabilitäten für verschiedene Übungen zwischen r = 0.60 und r = 0.92. In einer Übersicht berichtet Huck (1977) von Reliabilitäten von r = 0.68 bis 0.99. Howard (1974, in Obermann, 1992) beschreibt in einer Zusammenfassung mehrerer Studien Reliabilitäten von r = 0.60 bis 0.98. Diese Ergebnisse zur Beobachter-Übereinstimmung werden von den meisten Autoren (vgl. Obermann, 1992; Huck, 1977; Thornton & Byham,1982; Scholz, 1994) als Indiz dafür gedeutet, daß die Interrater-Reliabilität der Übungsurteile in Assessment Centern insgesamt gut ist.

In der Praxis der meisten Assessment Center beurteilen die Beobachter jedoch nicht die Übungen als ganzes, sondern geben innerhalb der Übungen Bewertungen hinsichtlich bestimmter vorgegebener Dimensionen (Thornton & Byham, 1982). Die Ergebnisse von Studien zur Interrater-Reliabilität, die die Bewertung der einzelnen Dimensionen untersuchen, sollen daher im folgenden näher betrachtet werden.

Interrater-Reliabilität der Dimensions-Bewertungen

In der Literatur werden unterschiedliche Einschätzungen der Interrater-Reliabilität auf Basis von Dimensions-Bewertungen gegeben. Während Thornton und Byham (1982) 23 Studien mit z.T. sehr hohen Interrater-Reliabilitäten auflisten, betonen Scholz (1994) und Lammers (1992), daß es nur wenige Untersuchungen zur Interrater-Reliabilität von Dimensions-Bewertungen gibt. Auch die Höhe der Korrelationen der Beobachter-Übereinstimmung ist bei Lammers (1992) und in den von Scholz (1994) zitierten Studien meist niedriger. Im weiteren sollen die Übersichten von Scholz (1994) und Thornton und Byham (1982) und einige mir wichtig erscheinende Untersuchungen kurz wiedergegeben werden.

Thornton und Byham (1982) ermittelten in ihrer Zusammenfassung von 23 Studien sehr hohe Interrater-Reliabilitäten für die Dimensionen Organisation und Planung, Entscheidungsfähigkeit, Initiative, Kommunikationsfähigkeit und Führung von mindestens 0.80. Niedrigere Werte ($0.50 < r < 0.70$) wurden für andere Dimensionen wie Fähigkeiten der Streßbewältigung oder Unabhängigkeit gefunden (Thornton & Byham, 1982). In den meisten dieser Studien ging den Bewertungen ein Informationsaustausch der Beobachter voraus. Die Autoren folgerten aus diesen Daten, daß Assessment-Center-Beurteiler fähig sind, in den meisten Dimensionen zuverlässige Bewertungen, gemessen an der Beobachter-Übereinstimmung, abzugeben (Thornton & Byham, 1982).

Das Vorgehen von Thornton und Byham (1982), Interrater-Reliabilitäten aus verschiedenen Assessment Centern für bestimmte Dimensionen zu ermitteln, stieß bei einigen Autoren auf Kritik (vgl. Fennekels, 1987; Lammers, 1992). Danach sei zu bezweifeln, ob die Dimensionen in zwei verschiedenen Studien exakt gleich operationalisiert worden sind.

In einer Übersicht nennt Scholz (1994) vier Studien zur Interrater-Reliabilität, die sich mit Dimensions-Bewertungen beschäftigen. Dabei wurde in Gruppendiskussionen eine niedrigere Interrater-Reliabilität (Spannweite von r = 0.38 bis 0.67) ermittelt als in Einzelübungen (Interrater-Reliabilität zwischen 0.65 und 0.95). Diese Tendenz zeigte sich auch in der eigenen Untersuchung von Scholz (1994). Die Interrater-Reliabilität in Gruppendiskussionen lag zwischen r = 0.43 und r = 0.55; in Einzelübungen hingegen bei r = 0.63 bis 0.76. Lammers (1992) fand in einer Studie z.T. sehr niedrige Interrater-Reliabilitäten mit einer Spannweite von r = 0.10 bis r = 0.53 je nachdem, welche Beobachter-Kombination und welche Dimension gegeben waren.

Borman (1982) untersuchte die Interrater-Reliabilität eines Assessment Centers der amerikanischen Streitkräfte. In diesem Assessment Center wurde jeder Kandidat in jeder Übung von zwei Beobachtern bewertet. Die sich daraus ergebenden Korrelationen hatten eine Spannweite von r = 0.44 bis 0.92 mit einem Median von r = 0.76 (Borman, 1976).

In einer weiteren Studie ermittelte Fennekels (1987) Interrater-Reliabilitäten zwischen r = 0.56 und r = 0.87. Er nimmt an, daß sich durch Lerneffekte die Beobachter-Übereinstimmung im Laufe des Assessment Center steigert.
Nach Jeserich (1981) ergaben sich in der ersten deutschen Untersuchung zum Thema der Beurteiler-Übereinstimmung (Jeserich, 1980 in Jeserich, 1981) Korrelationskoeffizienten von r = 0.84 für trainierte Psychologen und r = 0.81 für trainierte Linienmanager. Diese Werte beziehen sich auf eine Rangreihenbildung bei einer allerdings sehr kleinen Stichprobe von acht Versuchspersonen (Jeserich, 1981).

Insgesamt wird deutlich, daß die Forschung zur Beobachter-Übereinstimmung im Assessment Center bisher uneinheitliche Ergebnisse produziert hat. Obwohl die meisten Autoren die Interrater-Reliabilität von Assessment Center Bewertungen als gut ansehen, zeigen die Ergebnisse z.T. anderes. Die Spannbreite der Korrelationen ist sehr groß und liegt für alle hier zitierten Untersuchungen bei r = 0.10 bis r = 0.95. Das heißt, daß in einigen Studien Beobachter nur sehr geringe Übereinstimmung in ihren Bewertungen zeigten, während in anderen Studien für bestimmte Dimensionen fast perfekte Interrater-Reliabilität ermittelt wurde.

Auch die Einschätzung, wie umfangreich die Interrater-Reliabilität der Dimensions-Bewertungen bisher beforscht wurde, wird von den Autoren unterschiedlich gesehen (s.o.). Unbeantwortet ist auch, ob es bestimmte Dimensionen oder Übungen gibt, die höhere Interrater-Reliabilitäten bedingen. Es läßt sich also postulieren, daß Forschungsbedarf hinsichtlich der Bestimmung der Beobachter-Übereinstimmung im Assessment Center besteht.

2.4.2 Validität der Assessment Center Verfahren

Definition

Die Validität eines Tests sagt aus, wie gut der Test in der Lage ist, genau das zu messen, was er zu messen vorgibt (Bortz & Döring, 1995). Lienert (1969) beschreibt Validität folgendermaßen:

> Die Validität eines Testes gibt den Grad der Genauigkeit an, mit dem dieser Test dasjenige Persönlichkeitsmerkmal oder diejenige Verhaltensweise, das (die) er messen soll oder zu messen vorgibt, tatsächlich mißt. Ein Test ist demnach vollkommen valide, wenn seine Ergebnisse einen unmittelbaren und fehlerfreien Rückschluß auf den Ausprägungsgrad des zu erfassenden Persönlichkeits- oder Verhaltensmerkmal zulassen, wenn also der individuelle Testpunktwert eines Pb diesen auf der Merkmalsskala eindeutig lokalisiert (S. 16).

Es wird allgemein zwischen drei Arten von Validität unterschieden:
- Inhaltsvalidität
- Kriteriumsvalidität
- Konstruktvalidität
  (vgl. Bortz, 1993; Lienert, 1969)

Kritisch über diese Dreiteilung äußert sich Ghiselli (1964 in Schuler, 1989), wonach es nicht drei verschiedene Arten oder Typen von Validität gibt, sondern nur unterschiedliche Facetten einer Gesamtvalidität. Anastasi (1986 in Schuler, 1989) erklärt, daß die anderen beiden Arten dem Konzept der Konstruktvalidität untergeordnet bzw. nur spezielle Varianten seien. In dieser Arbeit soll jedoch an der üblichen Aufgliederung des Validitätsbegriffs festgehalten werden unter Berücksichtigung, daß die unterschiedlichen Typen nicht unabhängig voneinander sind.

Einen wichtigen Diskussionsbeitrag zur Validität leisteten Schuler und Stehle (1983). Die drei ursprünglichen Elemente erweitern sie um den Aspekt der „sozialen Validität", der nachfolgend in Grundzügen beschrieben werden soll. Vorher wird jedoch ein kurzer Überblick über den folgenden Abschnitt gegeben: Die Forschung zu den vier Aspekten der Validität von Assessment Centern ist Gegenstand der weiteren Ausführungen, wobei Untersuchungen zur Konstruktvalidität ausführlicher betrachtet werden.

Zur sozialen Validität von Assessment Centern

Das von Schuler und Stehle (1983) eingeführte und von Schuler (1990) weiterentwickelte Konzept der sozialen Validität steht als Sammelbegriff für das, „was die eignungdiagnostische Situation zu einer akzeptablen sozialen Situation macht" (Schuler & Stehle, 1983, S. 35).

Darunter sind im wesentlichen vier Aspekte zu verstehen:
1. Mitteilung relevanter Informationen über die wichtigen Charakteristika von Arbeitsplatz und Organisation
2. Partizipation an der Entwicklung und Anwendung eignungsdiagnostischer Instrumente
3. Transparenz des Verfahrens und der Schlußfolgerungen
4. Feedback in rücksichtsvoller, verständlicher Form
   (vgl. Schuler, 1990)

Soziale Validität überprüft v.a. folgende Fragestellung: Ist das Verfahren für die *Bewerber* akzeptabel? Fruhner, Schuler, Funke und Moser (1991) untersuchten am Beispiel der Auswahlverfahren Test und Vorstellungsgespräch, wie Kandidaten die Situation erlebten und bewerteten. Die Autoren fanden heraus, daß nicht die subjektive Belastung der Bewerber entscheidend für das Erleben einer Auswahlsituation ist, sondern Situationsparameter des Modells der sozialen Validität. Einige Studien überprüften die Bewertung von Assessment Centern in Abhängigkeit vom Abschneiden der Bewerber. Dabei zeigten die meisten Untersuchungen, daß es kaum signifikante Unterschiede zwischen erfolgreichen und nicht erfolgreichen Bewerbern hinsichtlich der Akzeptanz des durchlaufenen Assessment Centers gibt (vgl. Harburger, 1992; Schröder, 1997). Die Ergebnisse weiterer Studien (vgl. Sichler, 1989; Holling & Leippold, 1991) bestätigten, daß Bewerber das Assessment Center als sozial valide im Sinne von Schuler und Stehle (1983) empfinden.
Insgesamt kommen somit die meisten Autoren zu der Einschätzung, daß Assessment Center sozial valide sind (vgl. Fruhner et al. , 1991; Sichler, 1989; Schuler, 1987).

Zur Inhaltsvalidität von Assessment Centern

Nach Bortz und Döring (1995) liegt Inhaltsvalidität vor, „wenn der Inhalt der Test-Items das zu messende Konstrukt in seinen wichtigsten Aspekten erschöpfend erfaßt" (S. 185). Laut Lienert (1969) wird die Inhaltsvalidität „in der Regel durch ein Rating von Experten" (S.17) bestimmt. Somit basiert auch die Höhe der Inhaltsvalidität in der Regel auf subjektiver Einschätzung, da sie nicht numerisch ermittelt wird (vgl. Bortz & Döring, 1995).

Bei der Inhaltsvalidität von Assessment Centern geht es nach Schuler (1989) v.a. darum, wie repräsentativ die Verhaltensstichprobe ist. Das bedeutet, es soll überprüft werden, ob die Übungen „berufsrelevant" sind. Dabei wird das durch die einzelnen Übungen abgefragte Verhalten in Bezug auf zukünftige Tätigkeiten in einem bestimmten Beruf gesetzt. Die simple Schlußfolgerung, daß Assessment Center wegen des Anwendens praxisnaher Fallstudien und Rollenübungen inhaltsvalide seien, stellt jedoch nach Obermann (1992) eine grobe Vereinfachung dar. Einige Verhaltensbereiche wie z.B. Aspekte von Monotonieresistenz, Motivation und Ausdauer, könnten mit Hilfe einer kurzen Verhaltensstichprobe, wie sie durch das Assessment Center abgebildet wird, nur unzureichend beleuchtet werden (Schuler, 1987).

Abschließend läßt sich der Einschätzung von Schuler (1987) zustimmen, daß - trotz der offensichtlichen hohen Relevanz der Übungen für die spätere Berufstätigkeit - die Inhaltsvalidität von Assessment Centern nicht bewiesen ist, die meisten Autoren jedoch die inhaltliche Validität als nicht bedeutend ansehen. Daher soll in dieser Arbeit auf eine weiterführende Darstellung dieses Aspekts der Validität verzichtet und sich der Kriteriumsvalidität zugewandt werden.

Zur Kriteriumsvalidität von Assessment Centern

Die überwiegende Mehrzahl der Assessment Center werden für Personalauswahl und Personalentwicklungsentscheidungen verwandt (s.o.). Für die Personalauswahl ist die prognostische Validität besonders wichtig, also die Frage, wie gut ein Assessment Center späteren Berufserfolg vorhersagt (vgl. Kleinmann, 1997). Die Kriteriumsvalidität (auch kriterienbezogene oder empirische Validität) läßt sich in konkurrente und prognostische Validität aufgliedern, die sich jedoch nur im Faktor Zeit unterscheiden. Dabei wird differenziert zwischen zeitgleicher Messung, die die konkurrente Validität beschreibt und zeitversetzter Messung, die die prädiktive oder prognostische Validität bestimmt (vgl. Schuler, 1989; Kleinmann 1997). Die Operationalisierung des Validitätsbegriffs erfolgt gewöhnlich durch die Ermittlung des Zusammenhangs zwischen dem Prädiktor (hier: Ergebnisse aus dem Assessment Center) und einem Kriterium, z.B. dem beruflichen Erfolg. Dabei werden die Zusammenhänge zwischen dem Prädiktor und den Vorhersagen, die damit gemacht werden sollen, meistens durch Korrelationen wiedergegeben (Lienert, 1969).

Im Rahmen der Forschung zu Assessment Centern interessierte v.a. die prognostische Validität. Zu diesem Bereich gab es bisher auch die größte Forschungsanstrengung (vgl. Kleinmann, 1997).

Die oben bereits vorgestellte Management-Progress-Studie bei AT&T (Bray, 1964; Bray & Grant, 1966) brachte hierzu die ersten umfangreichen Ergebnisse hervor. Die Beurteiler sollten im Rahmen der Assessment Center Veranstaltungen die Bewerber dahingehend einschätzen, ob sie in den nächsten Jahren ins mittlere Management aufsteigen würden. Diese Einschätzungen wurden mit den tatsächlich erreichten Positionen der Assessment Center Teilnehmer verglichen (Thornton & Byham, 1982). Die folgenden Tabelle gibt die Trefferquoten und Validitätskoeffizienten wieder:

Tabelle 2.1 Trefferquoten und Validitätskoeffizienten der Einschätzungen der Management Progress-Studie (Daten aus Bray & Grant, 1966; Thornton & Byham, 1982)

Prädiktor		Kriterium			
Einschätzung der Beurteiler:		nach 8 Jahren		nach 16 Jahren	
Kandidat wird in 10 Jahren im		mit	ohne	mit	ohne
mittleren Management arbeiten	N	College	College	College	College
Ja	103	64%	40%	89%	63%
Nein/fraglich	106	32%	9%	66%	18%
Validitätskoeffizient		0.46	0.46	0.33	0.40

Die Daten zeigen, daß nach achtjähriger Tätigkeit knapp zwei Drittel (64%) der Kandidaten mit College-Abschluß, denen eine Position im mittleren Management prognostiziert wurde, diese auch tatsächlich erreicht haben. Nachwuchskräfte, bei denen erwartet wurde, daß sie nicht ins mittlere Management befördert würden, erreichten auch nur zu einem Drittel (32%) diese Ebene. Bei den Kandidaten ohne College-Abschluß ist dieses Verhältnis noch deutlicher (40% gegenüber 9%). Daraus läßt sich für beide Gruppen ein Validitätskoeffizient von r = 0.46 ermitteln. Rückt man die Daten nach sechzehn Jahren in den Blickpunkt, fällt auf, daß - entgegen der Prognose - die meisten der damaligen Nachwuchskräfte ins mittlere Management aufgestiegen sind. Diesen Umstand kritisieren Thornton und Byham (1982) und auch Kleinmann (1997), der darauf hinweist, daß die Varianz des Kriteriums geringer ist und somit auch die Validitätskoeffizienten negativ beeinflußt (r = 0.33 bzw. r = 0.40). Dies ist jedoch nur ein Teilaspekt der Studie.

Insgesamt werten die meisten Autoren (u.a. Domsch & Jochum, 1989; Kleinmann, 1997; Schuler, 1989; Thornton und Byham, 1982) die Ergebnisse als positiv hinsichtlich der Prognosequalität des Verfahrens; die Management-Progress-Studie wird allgemein als „eindrucksvoller Nachweis für die Validität des Assessment-Centers" (Kleinmann 1997, S.21) eingeschätzt.

Darüber hinaus gibt es eine Vielzahl weiterer Untersuchungen zur prognostischen Validität. Die bisher umfangreichste Zusammenfassung dieser Ergebnisse liefert die Meta-Analyse von Thornton, Gaugler, Rosenthal und Bentson (1987). Die Autoren ermittelten eine korrigierte mittlere Validität (bester Schätzwert) von r = 0.37, mit einer Varianz von 0.017. In diese Untersuchung gingen 50 Validitätsstudien mit insgesamt 107 Validitätskoeffizienten und einer Streubreite von r = -0.25 und r = +0.78 ein. Maukisch (1986) kommt in einer weiteren Meta-Analyse zu einem Koeffizienten von r = 0.40 für die Vorhersagequalität von Assessment Centern.
Probleme bei der Bestimmung der prognostischen Validität bereiten vor allem die Auswahl eines geeigneten Erfolgskriteriums (vgl. Thornton & Byham, 1982; Jeserich, 1981; Klimoski & Strickland, 1977) sowie der geringe zeitliche Abstand zwischen Assessment Center und Überprüfung der Ergebnisse (Jeserich, 1981).

Insgesamt kommen die meisten Autoren jedoch zu der Einschätzung, daß das Assessment Center Verfahren als prognostisch valide anzusehen ist (u.a. Schuler, 1989; Thornton & Byham, 1982; Kleinmann, 1997) und im Vergleich zu anderen Methoden gut abschneidet (Obermann, 1992; Kleinmann, 1997; Maukisch, 1986). Obermann (1992) hält dazu fest, „daß das Assessment Center gegenüber alternativen Methoden, wie Interviews, zumindest in der Prognose späterer Vorgesetztenurteile oder Aufstiegsmaße, trotz einiger methodischer Fragezeichen, überlegen ist" (S.261). Ähnlich äußert sich Maukisch (1986): „Assessment Center Systeme nehmen einen relativ gesicherten und günstigen Platz unter den Prädiktorklassen ein" (S. 90). Es kann als gesichert gelten, daß Assessment Center gute Vorhersagequalität für Berufserfolg unabhängig von Schulbildung, Geschlecht, ethnischer Herkunft oder Vorerfahrung besitzen (Klimoski & Brickner, 1987).

Was aber nun in einem Assessment Center gemessen wird, wird durch die prognostische Validität nicht beleuchtet. Dieser Frage gehen Studien zur Konstruktvalidität nach, die im nächsten Abschnitt vorgestellt werden.

2.4.2.1 Konstruktvalidität

Begriff und Definition

Der Begriff „Konstruktvalidität" wurde erstmals umfassend in den fünfziger Jahren von Cronbach und Meehl (1955 in Scholz, 1994) diskutiert. Danach sind Konstrukte nicht beobachtbare Eigenschaften von Personen (latente Variablen). Sie stehen zu anderen latenten aber auch zu beobachtbaren Variablen in Beziehung und werden durch dieses sogenannte „nomologische Netzwerk" implizit definiert. Zur Überprüfung eines nomologischen Netzwerks eines Konstruktes werden die beobachtbaren Variablen untersucht. Verfahren zur Überprüfung der Konstruktvalidität versuchen somit, dieses Netzwerk aufzuklären (vgl. Kleinmann, 1997). „Ein Test ist konstruktvalide, wenn aus dem zu messenden Zielkonstrukt Hypothesen ableitbar sind, die anhand der Testwerte bestätigt werden können" (Bortz und Döring, 1995, S. 186). Es wird dabei unterschieden zwischen konvergenter und diskriminanter Validität. Konvergente Validität ergibt sich, wenn mehrere Methoden dasselbe Konstrukt mit hoher Übereinstimmung (Konvergenz) messen. Unter diskriminanter Validität ist ein möglichst geringer Zusammenhang von Messungen verschiedener Konstrukte zu verstehen. Zum Nachweis von Konstruktvalidität muß demnach konvergente und diskriminante Validität vorhanden sein (vgl. Bortz & Döring, 1995).

Die Konstruktvalidität überprüft also die Frage, was ein Test mißt (Kleinmann (1997).

Bortz und Döring (1995) betonen die besondere Bedeutung der Konstruktvalidität in den Sozialwissenschaften. Auch Kleinmann (1997) kommt zu der Einschätzung, daß die Konstruktvalidität die zentrale diagnostische Frage ist. Übertragen auf Assessment Center Verfahren heißt das folgendes:

Da Assessment Center so konstruiert sind, daß Personenmerkmale in berufsrelevanten Übungen erfaßt werden sollen, ist zu überprüfen, ob diese Zusammenhänge wirklich gegeben sind. Werden die Teilnehmer eines Assessment Centers von den Beobachtern hinsichtlich der vorher festgelegten Dimensionen (Personenmerkmale), wie Durchsetzungsvermögen, Kreativität usw., gemessen? Können Beobachter überhaupt Personenmerkmale von Bewerbern beobachten und verläßlich bewerten? Sind die Assessment Center Dimensionen „Konstrukte" im Sinne von Cronbach und Meehl (1955, in Scholz, 1994), die verläßlich erfaßbar und voneinander abgrenzbar sind? (Schuler, 1989)

Bevor jedoch auf Studien zur Bestimmung der Konstruktvalidität eingegangen wird, soll ein weiterer Aspekt der Assessment Center Forschung vorgestellt werden. Einige Autoren diskutieren den Zusammenhang und die Bedeutung der prognostischen versus der Konstruktvalidität.

Dabei stellen Klimoski und Brickner (1987) eine Diskrepanz zwischen der recht gut erwiesenen prädiktiven Validität und der mangelnden Konstruktvalidität (s.u.) fest. Warum funktionieren Assessment Center, obwohl sie nicht das messen, was sie beabsichtigen? Klimoski und Strickland (1977) vermuten eine „indirekte Kriterienkontamination". Danach bewerten die Assessment Center Beobachter die Teilnehmer nicht objektiv nach den postulierten Dimensionen, sondern in hohem Maße auch nach impliziten Kriterien. Diese impliziten Kriterien setzen sich aus subjektiven Kriterien zukünftiger Vorgesetzter und unternehmensinternen Werten zusammen. Positiv bewertete Bewerber werden daher auch im Unternehmen später positiv bewertet. Das würde bedeutet, daß die hohe prognostische Validität von Assessment Centern durch diese impliziten Kriterien bedingt wäre. Diese These konnte in einer Studie von Kleinmann (1997) unterstützt werden.

Eine weitere mögliche Erklärung der Diskrepanz zwischen gut belegter prognostischer Validität und mangelnder Konstruktvalidität (s.u.) bietet Kleinmann (1993). Danach ist die „praktische Intelligenz" der Teilnehmer entscheidend. Bewerber, die erkennen, welche Dimensionen in einer Übung relevant sind, schneiden im Assessment Center als auch im späteren Berufsleben besser ab. Die Konstruktvalidität eines Assessment Centers würde dabei jedoch gering sein. Auch diese These konnte Kleinmann (1997) in einer Studie bestätigen.

Forschung zur Konstruktvalidität von Assessment Centern

Wie ermittelt man nun aber Konstruktvalidität? Diese Frage wird in der Literatur teilweise kontrovers diskutiert, worauf am Ende dieses Abschnittes näher eingegangen werden soll. In den Studien zur Bestimmung der Konstruktvalidität wurden eine Reihe von Verfahren angewendet. Die drei gängigsten Methoden sind die Analyse der Korrelationen der Multitrait-Multimethod-Matrix (MTMM-Matrix) nach Campbell und Fiske (1959), die explorative Faktorenanalyse mittels der Hauptkomponentenanalyse und die konfirmatorische Faktorenanalyse (Jöreskog & Sörbom, 1989).

Die Analyse der MTMM-Matrix soll im Rahmen dieses Abschnitts nur soweit beschrieben werden, wie es zum Verständnis der vorgestellten Studien nötig ist. Eine detailliertere Beschreibung dieser und der beiden faktorenanalytischen Auswertungsmethoden erfolgt im Rahmen des nächsten Kapitels.

Eine Methode zur Bestimmung der Konstruktvalidität entwickelten Campbell und Fiske (1959) mit ihrer „Multitrait-Multimethod-Matrix" (MTMM-Matrix), die voraussetzt, daß mehrere Traits (Personenmerkmale) mittels mehrerer Erfassungsmethoden erhoben werden. Unterschieden wird dabei in konvergente und diskriminante Validität (s.o.). Mit Hilfe der Analyse der Multitrait-Multimethod-Matrix läßt sich überprüfen, wie gut verschiedene Methoden dasselbe Konstrukt messen (konvergente Validität) und mit welcher Übereinstimmung verschiedene Konstrukte durch eine Methode differenziert werden (diskriminante Validität) (vgl. Bortz & Döring, 1995). Auf Assessment Center Verfahren angewendet, spiegeln die Übungen die verschiedenen Methoden wider und die unterschiedlichen Dimensionen die Traits.

Zahlreiche Untersuchungen wandten die Methode von Campbell und Fiske (1959) zur Überprüfung der Konstruktvalidität von Assessment Centern an. Die wahrscheinlich am häufigsten zitierte Arbeit dieser Art lieferten Sackett und Dreher (1982) mit ihrem Artikel über „some troubling empirical findings" der Konstruktvalidität. Dabei wurden die Assessment Center-Bewertungen von über 500 Teilnehmern aus drei verschiedenen Organisationen untersucht. Mithilfe der Analyse der Korrelationen zwischen Dimensionen innerhalb einer Übung (diskriminante Validität), der Korrelationen zwischen Bewertungen einer Dimension in verschiedenen Übungen (konvergente Validität) und der Hauptkomponentenanalyse (s. Abschnitt 3.4) fanden Sackett und Dreher (1982) heraus, daß die untersuchten Assessment Center nicht konstruktvalide im testtheoretischen Sinne sind.

Die durchschnittlichen Korrelationen zwischen Dimensionen (Traits), die in verschiedenen Übungen bewertet wurden, waren bei zwei Organisationen sehr gering ($r = 0.074$ bzw. $r = 0.109$). Wäre konvergente Validität gegeben, müßten jedoch Messungen derselben Dimension in verschiedenen Übungen hoch miteinander korrelieren. Dies war hier aber nicht der Fall, so daß den untersuchten Assessment Centern konvergente Validität abzusprechen ist.

Die konvergenten Korrelationskoeffizienten waren in der Untersuchung von Sackett und Dreher (1982) zudem auch deutlich kleiner als die durchschnittlichen Korrelationskoeffizienten von verschiedenen Dimensionen innerhalb einer Übung ($r = 0.638$ bzw. $r = 0.395$). Wäre diskriminante Validität gegeben, müßten die Korrelationskoeffizienten von verschiedenen Dimensionen innerhalb einer Übung niedrig und deutlich kleiner als die konvergenten Korrelationskoeffizienten sein. Das Gegenteil war hier der Fall, so daß auch nicht von diskriminanter Validität gesprochen werden kann. Nur die Korrelationen der Daten der dritten Organisation wichen hiervon etwas ab.

Die Analyse der Korrelationen der MTMM-Matrix ergab somit in der Untersuchung von Sackett und Dreher (1982), daß bei Assessment Centern von zwei Organisationen weder konvergente noch diskriminante Validität gegeben war.

Die Autoren ermittelten außerdem mit Hilfe des faktorenanalytischen Modells der Hauptkomponentenanalyse, daß den Beobachter-Ratings Übungsfaktoren und nicht Dimensionsfaktoren zugrunde lagen. Insgesamt bewerteten die Beobachter demnach nicht die einzelnen Dimensionen (Personenmerkmale), sondern gaben für eine bestimmte Übung Pauschalurteile ab (Sackett & Dreher, 1982).

Diese klassische Untersuchung von Sackett und Dreher (1982) löste rege Forschungstätigkeit aus und wurde mehrfach wiederholt. Die Studien von Robertson, Gratton und Sharpley (1987), Russell (1987), Turnage und Muchinsky (1982) und im deutschsprachigen Raum Hoenle (1995) und Neubauer (1989) erzielten ähnliche Ergebnisse, was dazu führte, daß das Vorhandensein von Konstruktvalidität in Assessment Centern insgesamt in Frage gestellt wurde (Kompa, 1989). Es galt demnach als gesichert, daß Assessment Center Verfahren dem testtheoretischen Anspruch der Konstruktvalidität nicht standhielten. Es werden also keine situationsübergreifenden Persönlichkeitsmerkmale des potentiellen Führungs-nachwuchses gemessen. Konvergente als auch diskriminante Validität ist nicht gegeben. Beurteiler sind anscheinend nicht in der Lage, in einer Übung zwischen unterschiedlichen Dimensionen zu differenzieren. Für diese scheinbar „niederschmetternden" Ergebnisse (Obermann, 1992, S. 238) der Konstruktvalidität des Assessment Centers wird seitdem nach Erklärungen gesucht.

Dazu wurde in einer Reihe von Studien die Fragestellung untersucht, welche Variablen die Konstruktvalidität von Assessment Centern moderieren bzw. welche Faktoren für das Nichterfassen der intendierten Konstrukte verantwortlich sind. Betrachtet man den Aufbau und Ablauf eines Assessment Centers könnten verschiedene Faktoren die mangelnde Konstruktvalidität beeinflussen.
Denkbare Einflußgrößen wären u.a. folgende:

- Assessment Center Entwickler
- Beobachter
- Teilnehmer
- Dimensionen
- Übungen

Zu allen Bereichen sind Untersuchungen publiziert worden, von denen nachfolgend die wichtigsten kurz wiedergegeben werden sollen.

Einflußgröße Assessment Center

Neubauer (1989) und Maukisch (1989) sehen als Ursache für die unbefriedigenden Ergebnisse zur Konstruktvalidität Fehler bei den Assessment Center Entwicklern. Nach Neubauer (1989) beabsichtigen Entwickler, die Dimensionen als „Dreh- und Angelpunkt der Assessment Center-Beurteilungen" (S.201) darzustellen. In der

Realität hingegen beurteilten Assessment Center-Beobachter demnach keine abstrakten Fähigkeiten oder Eigenschaften, sondern das Gesamtverhalten einer Person in einer bestimmten Übung.

Neubauer (1989) analysierte Daten eines Assessment Center zur Auswahl von Führungskräften. Dabei ergaben sich die gleichen Ergebnisse wie auch in der Untersuchung von Sackett und Dreher (1982). Neubauer (1989) sah es damit als erwiesen, daß „Faktorenanalysen über die Einzelurteile STETS [Hervorhebung im Original; Anm. d. Verf.] Übungsfaktoren und keine Merkmalsfaktoren" (S.203) ergeben.

Die Einordnung der zu beobachtenden Dimensionen in einem Assessment Center wurde von einigen Autoren z.T. kontrovers diskutiert. Diese als State-Trait bekannte Debatte versuchte aufzuklären, ob die Dimensionen eher reine Verhaltensmerkmale oder übergeordnete Personeneigenschaften darstellten (vgl. Kleinmann, 1997). Während Neubauer (1980) und Jeserich (1981) die Verhaltensorientierung von Assessment Centern betonten, kritisierte Maukisch (1989), daß die Autoren von Lehrbüchern zu Assessment Centern „in einer eigentümlichen Ambivalenz zwischen dem Ansatz der kriterienorientierten und der normorientierten Diagnostik ... verharren" (S. 261). Die Ursachen mangelnder Konstruktvalidität liegen demnach bei den Assessment Center-Konstukteuren, die sich nicht konsequent vom problematischen Traitansatz loslösten (Maukisch, 1989). Guldin und Schuler (1997) argumentierten, daß die Assessment Center Beurteilungs-Dimensionen nicht einheitlich als Traits zu verstehen sind, sondern daß es z.T. erhebliche Unterschiede zwischen den Dimensionen gebe.

Einflußgröße Beobachter

Nach Thornton und Byham (1982) und auch Jeserich (1981) sind Beobachterfehler eine mögliche Quelle mangelnder diskriminanter und konvergenter Validität. Danach würde ein sorgfältiges Beobachtertraining eventuelle Fehler minimieren.

Den Zusammenhang zwischen Interrater-Reliabilität und Konstruktvalidität haben einige Autoren diskutiert. Lammers (1992) untersuchte dabei Daten eines experimentellen Assessment Centers. Er konnte zeigen, daß mangelnde Interrater-Reliabilität die Konstruktvalidität negativ beeinflußt. Die Praxis von Assessment Centern sieht i.d.R. so aus, daß jeder Beobachter zwei Teilnehmer beurteilt. Dabei ergeben sich aufgrund der Beobachter-Rotation die Korrelationen zwischen unterschiedlichen Dimensionen innerhalb einer Übung aus den Bewertungen eines Beobachters. Die Korrelationen zwischen denselben Dimensionen in unterschiedlichen Übungen werden jedoch aus den Bewertungen verschiedener Beobachter ermittelt.

„Unter der Annahme, alle Beobachter seien in ihrem Urteilsverhalten nahezu gleich, erwachsen aus der Beobachtervariation keine Probleme. Sind sie es nicht, verringern sich durch diese Tatsache eventuell die gemittelten Korrelationen so stark, daß sie kleiner werden als die Korrelationen für jede Aufgabe gemittelt über alle Dimensionen" (Lammers, 1992, S. 67).

Auch Thornton und Byham (1982) und Kleinmann (1997) diskutierten dieses Problem. Kleinmann (1997) schlug dazu den Verzicht eines Rotationssystems vor, um mögliche Effekte zu umgehen. In einer Studie konnte Kleinmann (1997) diese Annahme tendenziell bestätigen.

Silverman, Dalessio, Woods und Johnson (1986) zeigten in einer Untersuchung, daß unterschiedliche Beurteilungsverfahren die konvergente und diskriminante Validität beeinflussen können. Ein aufgabenweises Vorgehen, bei dem nach jeder Übung alle Dimensionen von den Beobachtern beurteilt wurden, ergab niedrigere konvergente Validitäten als ein dimensionsgeleitetes Vorgehen, bei dem nach der letzten Übung jede Dimension über alle Aufgaben bewertet wurde. Kleinmann (1997) und Harris, Becker und Smith (1993 in Kleinmann, 1997) konnten in ähnlichen Studien diese Ergebnisse nicht wiederholen. Kleinmann (1997) schließt daraus, daß die von Silverman et al. (1986) gefundenen Ergebnisse auf die Anzahl der zu beobachtenden Dimensionen (s.u.) und nicht auf die Art des Beurteilungsverfahrens zurückzuführen sei.

Den Einfluß der Anzahl der zu beobachtenden Dimensionen auf die Konstruktvalidität überprüften Gaugler und Thornton (1989). Sie variierten die Anzahl der Dimensionen (3, 6 oder 9), die Beobachter innerhalb einer Übung bewerten mußten. Gaugler und Thornton (1989) ermittelten, daß Beobachter, die nur wenige Dimensionen zu beurteilen hatten, genauer und mit einer höheren konvergenten Validität bewerteten als Beobachter mit vielen Dimensionen. Die Autoren plädierten daher für eine Minimierung der Anzahl der Dimensionen pro Übung, um die Beobachter nicht kognitiv zu überfordern. Kleinmann (1997) unterstreicht, daß diese These durch weitere Forschung indirekt gestützt wird, da in Studien, bei denen pro Übung eine große Anzahl an Dimensionen zu beurteilen waren, keine konvergente Validität ermittelt wurde (Bycio, Alvares und Hahn, 1987; Sackett & Dreher, 1982). Während in den Studien, in denen konvergente Validität gefunden wurde, nur wenige Dimensionen pro Übung beurteilt wurden (Kleinmann, 1997). Lammers (1992) wiederholte die Studie von Gaugler und Thornton (1989) in ähnlicher Weise, konnte aber nicht die gleichen Ergebnisse ermitteln. Lammers (1992) bezweifelte abschließend den Nutzen einer Verringerung der Dimensionen.

Einflußgröße Teilnehmer

Kleinmann (1993) nahm an, daß Teilnehmer konsistenteres Verhalten in verschiedenen Übungen zeigen, wenn ihnen die beobachteten Dimensionen bekannt gegeben werden. In einer Studie konnte dieser Zusammenhang auch empirisch bestätigt werden (Kleinmann, 1997). Dabei verzeichneten die Bewertungen der Teilnehmer die höchste konvergente und diskriminante Validität, die angaben, sich nach den Dimensionen gerichtet zu haben. Den Einfluß des Teilnehmerverhaltens auf die Konstruktvalidität untersuchte auch Kuptsch (1994). Danach variiert die Verhaltenskonsistenz von Teilnehmern in einem Assessment Center interindividuell. Bewertungen von Teilnehmern, die ihr Verhalten variabel an die jeweilige Übung anpaßten, zeigten nach Kuptsch (1994) höhere konvergente Validität.

Einflußgröße Dimensionen

Die Zusammensetzung von Dimensionen innerhalb der Übungen untersuchte Kleinmann (1997). Dabei ergaben sich höhere diskriminante Validitäten, wenn Dimensionen verwendet wurden, die eine zeitgleiche Messung von unterschiedlichem dimensionsrelevantem Teilnehmerverhalten zuließen.

Für den Einfluß von Verhaltens-Checklisten auf die einzelnen Dimensionen interessierten sich Reilly, Henry und Smither (1990). Die Autoren konnten nachweisen, daß der Einsatz von Verhaltens-Checklisten die konvergente Validität steigerte. Die Autoren begründeten dies mit der verbesserten Möglichkeit für die Beobachter, das Teilnehmerverhalten eindeutig zuzuordnen.

Einflußgröße Übungen

Zusammenhänge zwischen Aufgabentypen untersuchte Templer (1995). Dabei wurden die Daten aus Assessment Centern zur Personalentwicklung hinsichtlich der Verfahrensklassen „Verhaltensbeobachtung", „Test zur Erfassung intellektueller Fähigkeiten" und „Organisationsaufgaben" analysiert. Es zeigte sich, daß Dimensionen innerhalb von situativen Aufgaben erwartungsgemäß hoch korrelierten. Templer (1995) empfahl abschließend, in Assessment Centern möglichst verschiedenartige Verfahrensklassen anzuwenden.

Ein weiterer Bereich der Assessment Center Forschung versuchte der Frage nachzugehen, was Assessment Center messen, wenn sie nicht die Dimensionen messen. Hierzu erschienen einige Studien, die den Zusammenhang zwischen Assessment Center Ergebnissen und Intelligenz- und Persönlichkeitstests untersuchten.

Scholz und Schuler (1993) unterstrichen in einer Meta-Analyse, die 55 Studien mit insgesamt über Zwanzigtausend Teilnehmern umfaßte, daß allgemeine Intelligenz der beste Prädiktor ($r = 0.33$) für das Abschneiden im Assessment Center ist. Auch andere Persönlichkeitskonstrukte wie soziale Kompetenz, Leistungsmotivation, Selbstvertrauen und Dominanz korrelierten demnach mittelhoch ($r = 0.23$ bis 0.31) mit dem Assessment Center Ergebnis (Scholz & Schuler, 1993).

Wie oben bereits angesprochen, hat sich spätestens seit der Arbeit von Sackett und Dreher (1982) eine Diskussion darüber entwickelt, mit welcher Methode die Konstruktvalidität von Assessment Centern zu bestimmen sei. Untersuchungen zur Konstruktvalidität produzierten unter Anwendung unterschiedlicher Methoden z.T. widersprüchliche Ergebnisse. Die mir am bedeutsamsten erscheinenden Studien sollen im folgenden genauer betrachtet werden. Im Zusammenhang mit der Assessment Center Forschung wurden eine Reihe von Verfahren angewendet. Der größte Teil läßt sich einteilen in die Analyse der MTMM-Matrix nach Campbell und Fiske (1959), die Varianzanalyse (vgl. Silverman et al., 1986), die explorative und konfirmatorische Faktorenanalyse.

Die meisten Untersuchungen, darunter auch die klassische Studie von Sackett und Dreher (1982) haben die Analyse der MTMM-Matrix und die explorative Faktorenanalyse nach der Hauptkomponentenanalyse verwendet (vgl. Hoenle, 1995; Neubauer, 1989; Robertson, Gratton & Sharpley, 1987; Russell, 1987; Turnage und Muchinsky, 1982). Ergebnisse dieser Studien ergaben anhand der Korrelationen der MTMM-Matrix, daß in den meisten untersuchten Assessment Centern konvergente und diskriminante Validität nicht vorhanden war oder es an ihr mangelte. Die Hauptkomponentenanalyse auf Grundlage der MTMM-Matrix produzierte in den Untersuchungen Übungsfaktoren und keine Dimensionsfaktoren. Dies führte dazu, daß Assessment Center insgesamt als nicht konstruktvalide eingeschätzt wurden (s.o.). Auf eine genauere Darstellung der Ergebnisse der einzelnen Studien soll in diesem Zusammenhang verzichtet werden, da sie letztlich keine neuen Resultate produzierten. Es wird daher auf die oben ausführlicher beschriebene Untersuchung von Sackett und Dreher (1982) verwiesen.

Einige Autoren (u.a. Bycio et al., 1987; Fennekels, 1987; Kleinmann, 1997) kritisieren den Einsatz der MTMM-Matrix und der Hauptkomponentenanalyse. Danach ist die konfirmatorische Faktorenanalyse mittels LISREL die einzig richtige Methode, die Konstruktvalidität von Assessment Centern zu bestimmen. Kleinmann (1997) weist auf Schwierigkeiten einer objektiven Bestimmung der konvergenten und diskriminanten Validität durch ausschließliche Verwendung der MTMM-Matrix hin. Schwierigkeiten liegen demnach darin, „daß die Auswertung auf der Grundlage der Korrelationen zwischen meßfehlerbehafteten, manifesten Variablen geschieht, die anschließende Interpretation allerdings Schlußfolgerungen über latente Variablen (Trait- und Methodenfaktoren) enthält" (Kleinmann, 1997, S. 46). Auch die Hauptkomponentenanalyse ist nach Kleinmann (1997) nicht geeignet, Konstruktvalidität von Assessment Centern zu ermitteln, weil sie keine statistische Überprüfung der gefunden Lösungen erlaubt. Der Autor nennt eine Reihe von weiteren Gründen, auf die hier nicht näher eingegangen werden soll. Der interessierte Leser sei zur weiteren vertiefenden Lektüre auf die Werke von Fennekels (1987) und Kleinmann (1997) verwiesen.

Einige neuere Studien (Bycio et al., 1987; Fennekels, 1987 und Kleinmann, 1997) verwendeten daher die konfirmatorische Faktorenanalyse mit dem Programm LISREL von Jöreskog und Sörbom (1989) zur Bestimmung der Konstruktvalidität.

In der Studie von Bycio et al. (1987) wurden die Bewertungen von fünf situativen Übungen, in denen jeweils acht Dimensionen zu beobachten waren, untersucht. Dabei ergaben sich analog zu den Ergebnissen von Sackett und Dreher (1982), daß die Bewertungen situationsspezifisch waren. Die Autoren schlossen daraus, daß in Assessment Centern vor allem die Performance in den einzelnen Übungen und nicht übergeordnete Managerfähigkeiten gemessen würden.

Fennekels (1987) überprüfte die Konstruktvalidität von Personalentwicklungs-seminaren eines deutschen Großunternehmens. Mit Hilfe der konfirmatorischen

Faktorenanalyse konnte er zeigen, daß die Bewertungen v.a. aus speziellen Übungsbedingungen resultierten. Zusätzlich beeinflußten nach Fennekels (1987) auch Methodenfehler, wie unkontrollierbare Interaktionsprozesse und der Zeitpunkt, die Bewertungen. Insgesamt produzierten diese beiden Untersuchungen demnach keine anderen Ergebnisse als die von Sackett und Dreher (1982).

Die Studien von Bycio et al. (1987) und Fennekels (1987), die Daten aus Assessment Centern aus der Praxis mit Hilfe der konfirmatorischen Faktorenanalyse überprüften, kritisierte Kleinmann (1997). Demnach könnten die Beobachter durch eine zu große Anzahl von Dimensionen überfordert gewesen sein, wie in der Untersuchung von Gaugler und Thornton (1989) nachgewiesen. Die Ergebnisse könnten auch durch mangelnde Interrater-Reliabilität trotz Einsatzes einer Beobachterrotation beeinflußt worden sein.

Kleinmann (1997) konnte in einer Laborstudie erstmals zeigen, daß sich bei der Analyse der Bewertungen neben Übungs- auch Dimensionsfaktoren abbildeten. Dazu wurden die Daten von 69 studentischen Teilnehmern eines eintägigen Assessment Centers mit Hilfe dreier Verfahren analysiert. Um Effekte mangelnder Beobachter-Übereinstimmung zu vermeiden, wurde auf eine Beobachterrotation verzichtet (vgl. Lammers, 1992). Außerdem mußten die Beobachter nur drei Dimensionen pro Übung bewerten, um die von Gaugler und Thornton (1989) postulierte kognitive Überforderung der Beobachter zu umgehen (s.o.). Während die Analyse der MTMM-Matrix und die Hauptkomponentenanalyse größtenteils die bekannten Ergebnisse produzierte, ergab die konfirmatorische Faktorenanalyse, daß „Übungseinflüsse *und* [Hervorhebung im Original; Anm. d. Verf.] die unterschiedlichen Konstrukte zur Erklärung der Verhaltensvarianz nötig sind" (Kleinmann, 1997, S. 44). Es konnten laut Kleinmann (1997) somit erstmals konvergente Validität und Dimensionsfaktoren von Assessment Center Bewertungen nachgewiesen werden.

Eine Studie von Kudisch, Ladd und Dobbins (1997) untersuchte ebenfalls Daten eines experimentellen Assessment Centers. Dabei wurden Bewertungen von 138 Studenten mit Hilfe der Multitrait-Multimethod-Methode und der konfirmatorischen Faktorenanalyse untersucht. Die Analyse der MTMM-Matrix reproduzierte die Ergebnisse der Untersuchung von Sackett und Dreher (1982), während die konfirmatorische Faktorenanalyse sowohl Übungs- als auch Dimensionsfaktoren generierte.

Einen weiteren Beitrag zur Methodendiskussion lieferten Guldin und Schuler (1997), die ein von Steyer (1987, 1988, in Guldin & Schuler, 1997) entwickeltes Evaluationsmodell zur Bestimmung von Spezifität und Konsistenz von Personenmerkmalen auf das Assessment Center Verfahren übertragen und angewendet haben. Auch in dieser Laborstudie zeigte sich entgegen der bisherigen Ergebnisse, daß die postulierten Dimensionen bzw. Personenmerkmale durchaus einen Teil der Verhaltensvarianz erklärten.

Die vorgestellten Untersuchungen zeigten alle, daß entgegen bisheriger Forschungsergebnisse, Beobachter doch die Personenmerkmale bewerteten. Ergebnisse dieser jüngeren Studien lassen abschließend vermuten, daß die oben zitierte These des völligen Fehlens von Konstruktvalidität im Assessment Center nicht aufrechtzuerhalten ist (vgl. Guldin & Schuler, 1997; Kleinmann, 1997; Kudisch et al., 1997). Es bleibt somit weiter ungeklärt, was im Assessment Center gemessen wird.

3 Methode

Gegenstand dieses Kapitels ist die Beschreibung des untersuchten Assessment Centers, wobei insbesondere auf die einzelnen Übungen eingegangen werden soll. Außerdem werden die Art der Datenerhebung und die Merkmale der Stichprobe beschrieben. Fragestellung und Hypothesen sollen spezifiziert werden, und abschließend werden die Methoden vorgestellt, mit Hilfe derer die Daten untersucht und ausgewertet werden.

Zuerst sollen jedoch die hier verwendeten Begriffe erläutert werden. In der Literatur findet sich geradezu eine Begriffsvielfalt zu den Bezeichnungen der verschiedenen Bausteine eines Assessment Centers. Uneinigkeit besteht v.a. im Hinblick auf die Benennung der Assessment Center Dimensionen. Im Rahmen dieser Arbeit soll folgendes gelten:
Es wird vereinfachend insgesamt von Übungen gesprochen, obwohl genaugenommen das Interview (s. u.) nicht darunter fallen würde. Unterscheiden möchte ich im weiteren zwischen (übergeordneten) Personenmerkmalen, Assessment Center Dimensionen und Einzeldimensionen, wobei Assessment Center Dimensionen und Dimensionen sind als Synonyme zu verstehen sind. Ziel von Assessment Centern ist es, übergeordnete Personenmerkmale (oder Eigenschaften) wie Entscheidungsfähigkeit, Kreativität usw. zu messen. Dafür werden in berufsrelevanten Übungen Bewertungen zu Einzeldimensionen gegeben (hier z.B. Analysefähigkeit in Gruppendiskussion „B"; Entscheidungsfähigkeit im Rollenspiel, vgl. Ergebnismatrix 3.1). Die Bewertungen der zusammengehörenden Einzeldimensionen bestimmen die jeweilige Assessment Center Dimension. Als Einzeldimension sind in dieser Arbeit also die Einzelbewertungen einer Dimension innerhalb einer Übung gemeint. Ob die Dimensionen wirklich - wie beabsichtigt - Personenmerkmale darstellen, gilt es zu prüfen. Ein Beispiel soll diese Zusammenhänge weiter erläutern:
Die Bewertungen der drei Einzeldimensionen Entscheidungsfähigkeit/ Rollenspiel, Entscheidungsfähigkeit/ Präsentations-Übung 1 und Entscheidungsfähigkeit/ Präsentations-Übung 2 (vgl. Ergebnismatrix 3.1) bestimmen die Assessment Center Dimension Entscheidungsfähigkeit, die wiederum das Personenmerkmal Entscheidungsfähigkeit messen soll. Der Frage, ob die im Assessment Center gemessene Entscheidungsfähigkeit wirklich so etwas wie Entscheidungsfähigkeit von Personen prüft, wird im Rahmen der Forschung zur Konstruktvalidität (s.o.) nachgegangen.

3.1 Beschreibung des untersuchten Assessment Centers

Im folgenden soll eine Kurzbeschreibung des im Rahmen dieser Arbeit untersuchten Verfahrens gegeben werden. Es wird dabei zuerst auf die Entwicklung des Assessment Centers eingegangen; im zweiten Teil wird die Struktur und der Ablauf dieses Personalauswahlverfahrens beschrieben.

3.1.1 Entwicklung des Assessment Centers

Dieses Assessment Center wurde von einer deutschen Unternehmensberatung für ein Großunternehmen entwickelt. Nach Aussage des Assessment Center Entwicklers wurde im Vorwege die Repetory Grid Technik zur Anforderungsanalyse eingesetzt. Nach zusätzlichen Interviews mit den zuständigen Personen des Unternehmens und Personen, die bereits erfolgreich auf der Zielposition gearbeitet haben, wurde eine Liste von Anforderungen an potentielle Bewerber erstellt. Diese Liste von Personenmerkmalen wurde nach einer letzten Absprache mit dem Unternehmen auf beobachtbare Dimensionen operationalisiert, die letztlich auch im Assessment Center abgefragt wurden. Die zwölf Assessment Center Dimensionen waren:

- Organisation
- Analysefähigkeit
- Entscheidungsfähigkeit
- Flexibilität
- Kontaktfähigkeit
- Einfühlungsvermögen
- Integration
- Durchsetzung
- Kreativität
- Ethische Grundhaltung
- Innovationsfähigkeit
- Ausstrahlung

Das Assessment Center zeichnete sich durch eine „unternehmenskulturspezifische" Auswahl der Dimensionen aus (vgl. Ethische Grundhaltung, Ausstrahlung). Es war so konzipiert, daß theoretisch alle Bewerberpaare (zur Besonderheit der Kandidaten siehe nächsten Abschnitt) eines Termins bestehen bzw. nicht bestehen konnten. Grundlage der Ratings sollte demnach ein Bewertungsmaßstab sein, der unabhängig von der Gruppe der Bewerber war und alleine aus den Anforderungen der Zielposition abgeleitet wurde.

3.1.2 Aufbau des Assessment Centers

Nach der Vorstellung, was Assessment Center Verfahren gemeinsam haben (s. erstes Kapitel), soll nun das untersuchte Assessment Center genauer beschrieben werden. Das Bewerberauswahlverfahren ist hier definiert als ein Verfahren, in dessen Rahmen

- eine Gruppe von Bewerbern
- anhand einer Reihe verschiedenartiger Übungen
- in möglichst realitätsnahen Situationen
- von mehreren geschulten Gutachtern
- nach festgelegten Regeln
- beobachtet und beurteilt wird (aus Beobachter – Handbuch).

Das untersuchte Assessment Center wurde als Personalauswahlverfahren mit dem Ziel eingesetzt, selbständig arbeitende Partner des Großunternehmens zu rekrutieren. Dabei wurden als Besonderheit Bewerberpaare (i.d.R. Lebenspartner) gesucht, die jedoch die Übungen des Assessment Centers alleine zu absolvieren hatten (zur detaillierteren Beschreibung der Stichprobe s. Abschnitt 3.2). Die Bewerber mußten innerhalb von zwei Tagen insgesamt acht Übungen oder Aufgaben absolvieren. Folgende Darstellung soll einen Überblick über den zeitlichen Ablauf und die Bestandteile des Verfahrens geben:

	1.Tag
10.00 - 10.45	Begrüßung, Vorstellung
10.45 - 11.30	Einführung in den Ablauf
11.40 - 12.20	Gruppendiskussion
12.30 - 13.30	Mittagspause
13.30 - 17.25	Einzelübungen
17.30 - 18.15	Gruppendiskussion
19.00 - 20.00	Gemeinsames Abendessen
20.00 - 21.30	Interviews
	2.Tag
08.30 - 11.30	Kreativitäts-Übungen
11.30 - 11.45	Abschlußrunde
11.45 - 15.30	Beobachter – Konferenz
15.30 - 17.00	Feedback – Gespräche

Der Aufbau der Veranstaltung war verschachtelt. Das bedeutet, daß alle Bewerber alle Übungen durchliefen, jedoch zu unterschiedlichen Zeiten. Die Beobachter- und die Teilnehmerkonstellationen wurden immer wieder gewechselt (rotiert), so daß insgesamt jeder Beobachter jeden Bewerber mehrfach beurteilen mußte.
Insgesamt wurde durch diese Struktur erreicht, daß jeder Bewerber pro Übung zwei Bewertungen von unterschiedlichen Beobachtern erhielt.

Die Übungen setzten sich zusammen aus der Vorbereitungszeit, die ohne Beobachtung erfolgte und der Beobachtungszeit, während der das Verhalten der Bewerber bewertet wurde. Die folgende Übersicht soll die Übungen in ihrer zeitliche Struktur aus Sicht der Kandidaten verdeutlichen:

Struktur der Übungen		
	Vorbereitungs-zeit	Beobachtungs-zeit
Interaktionsübungen		
• Gruppendiskussion „A"	15 Min.	25Min.
• Gruppendiskussion „B"	20 Min.	25 Min.
• Rollenspiel	15 Min.	25 Min.
Einzelübungen		
• Präsentations-Übung 1	30 Min.	25 Min.
• Präsentations-Übung 2	30 Min.	25 Min.
• Kreativitäts-Übung 1	20 Min.	65 Min.
• Kreativitäts-Übung 2	20 Min	65 Min
• Interview	35 Min.	40 Min.

Anmerkung: die Bezeichnung der Übungen wurde aus Datenschutzgründen geändert.

Insgesamt verbrachten die Bewerber fünf Stunden in Beobachtungssituationen und mehr als drei Stunden damit, sich auf die Übungen vorzubereiten. Die Übungen, in denen das Verhalten der Kandidaten beobachtet und bewertet werden sollte, sind hier in Einzel- und Interaktionsübungen unterteilt. Letzterer Aufgabentyp ist durch Gruppensituationen oder durch das Zusammenspiel mit anderen Personen bestimmt (Gruppendiskussionen und Rollenspiel), während in den Einzelübungen die Kandidaten i.d.R. etwas alleine präsentieren oder darstellen sollen. Die beiden Kreativitäts-Übungen werden im Rahmen dieser Arbeit den Einzelübungen zugezählt, obwohl sie auch eingeschränkte Gruppensituation beinhalten. Im folgenden sollen nun die einzelnen Bausteine des Verfahrens vorgestellt werden.

Vorstellung

Zu Beginn der Veranstaltung wurden alle Teilnehmer durch einen führenden Mitarbeiter des Unternehmens und der Unternehmensberatung begrüßt. Bereits in der Anfangsphase sollte die Atmosphäre aufgelockert werden. Nach einer ausführlichen Vorstellung der anwesenden Unternehmensvertreter stellten sich die Bewerber vor. Anschließend präsentierte ein leitender Mitarbeiter das Großunternehmen und berichtete über Ziele und Hintergründe des Personalauswahlverfahrens. Offene Fragen seitens der Bewerber wurden zum Abschluß der Vorstellung beantwortet.

Einführung in den Verfahrensablauf

Der Moderator erklärte innerhalb dieser Einführung den Bewerbern den genauen Ablauf des Assessment Centers. Die Bewerber wurden ermutigt, Fragen zu stellen und eingeladen, das Unternehmen und die anderen Teilnehmer kennenzulernen. Das Leitmotiv der ersten beiden Bausteine des Assessment Centers war es, den Kandidaten Angst und Aufregung zu nehmen und das Verfahren - insbesondere die Beobachtungssituation - transparent zu machen.

Gruppendiskussion „A"

Diese Diskussion war als sog. führerlose Gruppendiskussion konzipiert, was bedeutet, daß jedes Gruppenmitglied gleichermaßen für die Organisation und Gestaltung der Gruppenprozesse verantwortlich war. Jeder Teilnehmer erhielt die Instruktion, als Mitglied eines fiktiven Ausschusses neue Anforderungen an Führungspersonen zu entwickeln. In einer früheren Sitzung hatte dieser Ausschuß demnach bereits eine Liste von wichtigen Eigenschaften erarbeitet. Die Teilnehmer sollten nun als Vorbereitung auf die folgende Sitzung eine Rangreihe ihrer fünf wichtigsten Anforderungen zusammenstellen. In der Gruppe sollten die Teilnehmer daraufhin gemeinsam einen Katalog der fünf wichtigsten Anforderungen diskutieren. Dabei war jeder Teilnehmer angehalten, eine möglichst hohe Übereinstimmung zwischen der Rangreihe der Gruppe und seiner eigenen zu erreichen. Das bedeutete, daß die Kandidaten versuchten, die Gruppe für möglichst viele eigene Vorschläge zu gewinnen. Bewerberpaare wurden voneinander getrennt und in zwei verschiedene Gruppen eingeteilt. Einzeldimensionen, die in diesem Zusammenhang beobachtet und anschließend bewertet wurden, waren Entscheidungsfähigkeit, Flexibilität, Kontaktfähigkeit, Integration und Durchsetzung.

Gruppendiskussion „B"

Die Teilnehmer beschäftigten sich zunächst in 20-minütiger Einzelarbeit jeweils mit einer Reihe von Bewerbungen. In der anschließenden führerlosen Gruppendiskussion sollten sich die Kandidaten dann als fiktive Kollegen über einige offene Positionen mit dem Ziel austauschen, eine möglichst gute Zuordnung von Bewerber zu Posten zu erreichen. Die Übung war so konzipiert, daß denkbare optimale Lösungen in der Regel über einen Austausch der Bewerbungen unter den einzelnen Kollegen zu erzielen waren. Zum Abschluß der 25-minütigen Diskussion sollte die Gruppe das Ergebnis am Flip-Chart präsentieren. Den Beobachtern lag dabei eine mögliche Lösung vor, wie die Bewerbungen auf die offenen Stellen verteilt werden könnten. Zu beobachtende Einzeldimensionen in dieser Diskussion waren Analysefähigkeit, Flexibilität, Einfühlungsvermögen, Integration und Durchsetzung.

Präsentations-Übungen

Die beiden Präsentations-Übungen bestanden jeweils aus zwei separaten Aufgaben, die von den Bewerberpaaren getrennt bearbeitet und präsentiert wurden. Den Teilnehmern standen dabei Präsentationsmaterialien wie z.B. Overheadprojektor oder Flip-Chart zur Verfügung. Im Rahmen der ersten Übung bearbeiteten die

Teilnehmerpaare zwei administrative Aufgabenstellungen. Eine aus vielen Assessment Centern bekannte Postkorb-Aufgabe und ein Schichtplan sollten dabei innerhalb von 30 Minuten vorbereitet werden. Die erarbeiteten Ergebnisse wurden anschließend den Beobachtern in einer 25-minütigen Präsentation vorgestellt. Die Bewerber sollten im Rahmen der Präsentation ihre Sichtweisen zu den Problemstellungen darlegen und konkrete Lösungsvorschläge aufzeigen. Die Beobachter bewerteten das Verhalten hinsichtlich der Einzeldimensionen Organisation, Analysefähigkeit und Entscheidungsfähigkeit. Die Präsentations-Übung 2 bestand aus einer betriebswirtschaftlich ausgerichteten Aufgabenstellung und einer allgemeineren Aufgabe zum Thema „Entwicklung einer Marktposition". Zur Vorbereitung der Präsentation waren 30 Minuten vorgesehen. Im Hinblick auf den Umfang der Aufgabe und die bestehende Zeitrestriktion kam es in beiden Präsentations-Übungen v.a. darauf an, daß die Bewerber eine effektive Arbeitseinteilung fanden. Einzeldimensionen, die in dieser Übung beobachtet werden sollten, waren Organisation, Analysefähigkeit, Entscheidungsfähigkeit und Kreativität.

Rollenspiel

Im Rahmen dieser Übung fungierten junge Mitarbeiter des Unternehmens als Rollenspieler. Sie wurden in ihre Rolle standardisiert eingewiesen. Die Bewerber sollten in der Rolle eines selbständigen Unternehmers zwei Kurzgespräche mit Mitarbeitern führen, deren Verhalten Anlaß zur Kritik gab sowie ein weiteres Gespräch mit einem Kunden, der eine Reklamation vorbrachte. Die Bewerber erhielten zur Vorbereitung des 25-minütigen Rollenspiels Informationen über die zwei Mitarbeiter und den Kunden. Fokus der Beobachtungen war in dieser Übung, wie die Teilnehmer auf nicht vorhersehbare zwischenmenschliche Konflikte reagierten und auf was für ein Verhaltensrepertoire sie zurückgriffen. Dabei sollten sechs Einzeldimensionen bewertet werden: Organisation, Entscheidungsfähigkeit, Flexibilität, Kontaktfähigkeit, Einfühlungsvermögen und Durchsetzung.

Interview

Im Interview wurden die Teilnehmerpaare von zwei Beobachtern zu ihrem bisherigen Lebensweg, ihren Stärken und Schwächen, persönlichen Werten und Einstellungen befragt. Das Interview war kein sogenanntes Streßinterview, sondern als Dialog gedacht. Die Teilnehmer hatten die Möglichkeit, sich 35 Minuten anhand vorformulierter Fragen vorzubereiten. Diese Fragen dienten dann den Interviewern als Leitfaden, auf den sie sich jedoch nicht beschränken mußten. Die Interviewer waren vielmehr aufgefordert auch Fragen zu stellen, die sich aus dem Dialog ergaben. Im Anschluß bewerteten die Interviewer die Paare analog zu den anderen Übungen auf einem Beurteilungsbogen hinsichtlich der Einzeldimensionen Innovationsfähigkeit, Ethische Grundhaltung und Ausstrahlung. Das Interview diente der Vertiefung der Eindrücke aus den bisherigen Übungen und der Einschätzung der Gesamtpersönlichkeit. Eine weitere Funktion war darin zu sehen, daß durch die Gesprächssituation den Bewerbern die Möglichkeit gegeben wurde, sich als Person individuell darzustellen. Das Unternehmen konnte sich somit stellvertretend durch die Interviewer ein ganz persönliches Bild von den Bewerbern machen.

Außerdem war es erwünscht, daß die Teilnehmerpaare ein Feedback zu der Veranstaltung und ihrer Teilnahme gaben.

Kreativitäts-Übungen

Der 2. Tag des Assessment Centers bestand neben dem abschließenden Feedbackgespräch im wesentlichen aus zwei Bausteinen: den Kreativitäts-Übungen und der Beobachter-Konferenz. Die Kreativitäts-Übungen wiesen einige Besonderheiten auf: Die Übungseinheiten unterteilten sich jeweils in eine 35-minütige Gruppenarbeit und eine 20-minütige Präsentation der in der Gruppe erarbeiteten Ergebnisse im Plenum. Die eigentliche Beobachtungssituation war die Vorbereitung der Präsentation durch die Gruppe. Dafür erhielten die Gruppen eine gemeinsame Instruktion. Die Paare wurden wiederum in verschiede Gruppen aufgeteilt. In diesem Rahmen wurden die Einzeldimensionen Kreativität, Innovationsfähigkeit und Ausstrahlung beobachtet.

Beobachterkonferenz / Urteilsbildung

Die Beobachterkonferenz, die von den Moderatoren geleitetet wurde, war das beschließende Gremium aller Beobachter zur Urteilsbildung. Die Performance jedes Teilnehmers wurde dabei anhand einer Ergebnismatrix der Beurteilungen diskutiert, die im folgenden dargestellt ist.

Tabelle 3.1: Ergebnismatrix

Übung / Dimension	GA	GB	RO	P1	P2	K1	K2	IN
Organisation	■	■	■			■	■	■
Analysefähigkeit	■		■			■	■	■
Entscheidungsfähigkeit		■				■	■	■
Flexibilität				■	■	■	■	■
Kontaktfähigkeit		■		■	■	■	■	■
Einfühlungsvermögen	■			■	■	■	■	■
Integration			■	■	■	■	■	■
Durchsetzung				■	■	■	■	■
Kreativität	■	■	■	■				■
Innovationsfähigkeit	■	■	■	■	■			
Ethische Grundhaltung	■	■		■	■	■	■	
Ausstrahlung	■	■	■	■	■			

Anmerkung: GA: Gruppendiskussion „A"; GB: Gruppendiskussion „B"; RS: Rollenspiel; P1: Präsentations-Übung 1; P2: Präsentations-Übung 2; K1: Kreativitäts-Übung 1; K2: Kreativitäts-Übung 2; IN: Interview.

Die freien Felder bedeuten, daß die jeweilige Einzeldimension nur in der jeweiligen Übung beobachtet wird. Die schattierten Felder deuten darauf hin, daß es diese Dimensions-Übungssituation nicht gibt. Insgesamt erhalten die Kandidaten des Assessment Centers also 32 Bewertungen (32 freie Felder) in acht verschiedenen Beobachtungssituationen (Übungen).

Die Vorgehensweise der Beobachterkonferenz war in diesem Assessment Center übungsorientiert; das heißt, die Bewertungen des jeweiligen Kandidaten wurden nacheinander für jede Übung diskutiert. Dabei sollten die Beobachter anhand ihrer Notizen ihre Beobachtungen und daraus resultierenden Bewertungen erörtern. Da jeder Kandidat pro Einzeldimension von zwei unterschiedlichen Beobachtern bewertet wurde (s.o.), sollten die Beobachter bei deutlichen Abweichungen über die Urteile diskutieren und möglichst zu einem Konsens kommen. Ein abschließendes Protokoll dokumentierte die beobachteten Stärken und Schwächen für jeden Teilnehmer in jeder Übung. Dieses Stärken- / Schwächen-Protokoll diente den Beobachtern als Grundlage für die Feedbackgespräche (s.u.). Eine Besonderheit war, daß die Stärken und Schwächen nicht auf die in den Übungen abgefragten Dimensionen beschränkt blieben, sondern auch allgemeine Eindrücke und Bereiche einbezogen werden konnten. Für jedes Bewerberpaar wurde abschließend das Gesamtergebnis diskutiert und eine Bewertung gegeben. Dabei kamen vier verschiedene Abschlußbewertungen in Betracht: erfolgreich bestanden; bestanden mit Einschränkungen; nicht bestanden - jedoch Zusammenarbeit eventuell möglich und nicht bestanden.

Feedbackgespräche

Die Rückmeldung der Assessment Center Ergebnisse an die Bewerberpaare war gedacht als offenes Gespräch und nicht als „Urteilsverkündung". Die Beobachter wurden im Rahmen des Beobachter-Trainings besonders auf die Wichtigkeit der Feedbackgespräche hingewiesen. Ihnen lag dabei eine Vorgehenshilfe zur organisatorischen und inhaltlichen Gestaltung des Gesprächs vor. Darin wurde u.a. die Chance betont, die die Gespräche für die Teilnehmer auf Grundlage der Fremdeinschätzungen bieten. Grundlage der Gespräche bildeten die Stärken / Schwächen – Protokolle (s.o.) der einzelnen Teilnehmer. Mit erfolgreichen Bewerberpaaren wurde im Anschluß das weitere Vorgehen besprochen.

Rahmen

Der Rahmen der Veranstaltung war ein Teil des Gesamtkonzepts und sollte v.a. die Funktion erfüllen, eine für alle Teilnehmer angenehme Atmosphäre zu schaffen. Schwerpunkt war das gegenseitige Kennenlernen. Das Verfahren ging über zwei Tage und wurde in Tagungshotels durchgeführt. Die Kosten für Übernachtung und Verpflegung im Tagungshotel sowie An- / Abreise wurden vom Unternehmen getragen. Die Mittag- und Abendessen waren ausdrücklich als gemeinsame Veranstaltung aller Teilnehmer konzipiert, um einen informellen Austausch zwischen allen Beteiligten anzuregen. Gewollter „Nebenertrag" der Veranstaltung waren erhoffte positive Effekte für das Unternehmen durch die Beteiligung von Führungskräften.

Leitidee der gesamten Veranstaltung war es, neben der Bewerberauswahl eine soziale Zusammenarbeit zu gewährleisten, da die Bewerber in den Augen des Unternehmens nicht nur Kandidaten auf eine Zielposition, sondern auch potentielle Kunden und Multiplikatoren von Erfahrung waren, die sie durch den nahen Kontakt mit dem Unternehmen machen konnten.

3.2 Datenerhebung

In diesem Kapitel sollen zunächst die Merkmale der Stichprobe beschrieben werden. Im Anschluß daran werden die verwendeten Beurteilungsbögen vorgestellt.

3.2.1 Beschreibung der Teilnehmer des Assessment Centers

Drei verschiedene Personengruppen haben an diesem Assessment Center teilgenommen: die Kandidaten, die Beobachter und die Moderatoren. In der Regel wurde die Veranstaltung in einem Setting von zwölf Kandidaten und acht Beobachtern durchgeführt. Dabei übernahmen jeweils zwei Mitarbeiter der Unternehmensberatung die Moderation. Bei der Datenerhebung waren die Beobachter und die Kandidaten des Assessment Centers von Bedeutung, die daher im folgenden genauer beschrieben werden sollen.

Die Beobachter des Assessment Center wurden zu Beginn der Durchführungen im Frühjahr 1995 in einem speziellen Training auf ihre Aufgaben vorbereitet. Insgesamt wurden ca. 40 Personen zu Beobachtern geschult, davon waren in etwa die Hälfte die zukünftigen direkten Vorgesetzten der Kandidaten. Die restlichen Beobachter rekrutierten sich aus Mitarbeitern der Personalabteilung. Ziel der Schulung war es, die Beobachter mit der Struktur und dem Ablauf der Veranstaltung vertraut zu machen. Die Übungen und die aus dem Anforderungsprofil abgeleiteten operationalisierten Dimensionen wurden vorgestellt und inhaltlich erarbeitet. Darüber hinaus wurden die Beobachter über die aus der psychologischen Grundlagenforschung bekannten Beurteilungstendenzen informiert, wie z.B. Halo-Effekt, Tendenz zur Mitte, Milde- / Strenge- Effekte. Neue Beobachter durchliefen nach einer theoretischen Einführung die erste Assessment Center Teilnahme als „Probelauf". Erst ab der zweiten Teilnahme wurden sie in der Bewertung der Bewerber berücksichtigt. Als Unterstützung während der Assessment Center Durchführungen erhielten alle Beobachter zu Beginn jeder Veranstaltung ein Handbuch, in dem Informationen über Ablauf und Inhalt der Übungen enthalten waren.

Bei den Teilnehmern des Assessment Centers handelte es sich um Personen, die sich auf die Zielposition, ein selbständiger Partner des ausschreibenden Unternehmens zu werden, beworben hatten. Besonderheit dieses Assessment Centers war, daß sich die Ausschreibung der Stellen an Paare (i. d. R. Lebenspartner) richtete. Die Bewerber durchliefen dabei die Übungen des Assessment Centers einzeln, wurden jedoch abschließend als Paar beurteilt. Hintergrund dieser besonderen Strategie der Paarbewerbung war der Gedanke, die hohen Anforderungen einer selbständigen

Tätigkeit auf zwei Personen zu verteilen und eine hohe Akzeptanz der Arbeit durch beide Lebenspartner zu garantieren. Dem Assessment Center ging ein Auswahlverfahren der Kandidaten voraus. Es wurden je nach Bedarf Anzeigen geschaltet, auf die sich potentielle Kandidaten bewarben. Nach einer Vorauswahl anhand der Bewerbungsunterlagen interviewten leitende Mitarbeiter die Kandidatenpaare. Die erfolgreichen Paare erhielten daraufhin eine Einladung zur Teilnahme am Assessment Center. Die folgende Tabelle soll die Stichprobe weiter spezifizieren.

Tabelle 3.2: Merkmale der Stichprobe

Zeitraum		Januar 1996 bis Oktober 1998
Anzahl der Durchführungen		27 Assessment Center
Anzahl der Beobachter	Insgesamt	51 Personen
	Frauen	10 Personen (20 %)
	Männer	41 Personen (80 %)
Anzahl der Teilnehmer	Insgesamt	317 Personen
	Paare (männlich– weiblich)	276 Personen (87 %)
	Paare (weiblich – weiblich)	6 Personen (2 %)
	Paare (männlich – männlich)	8 Personen (3 %)
	Einzelpersonen (weiblich)	4 Personen (1 %)
	Einzelpersonen (männlich)	23 Personen (7 %)

Die Konzeption dieses Assessment Centers als Auswahlverfahren für Paare spiegelt sich auch in der Struktur der Teilnehmer wieder. Über 90 % aller Kandidaten haben als Paar teilgenommen.

3.2.2 Beschreibung der Beurteilungsbögen

Die Beobachter erhielten zu jeder Übung eine Übersicht, in der der zeitliche Rahmen und der Inhalt der Übung beschrieben waren. Die zu beobachtenden Dimensionen wurden aufgelistet und mit Verhaltensbeispielen spezifiziert und erklärt. Diese konkreten Beispiele für Verhaltensbeobachtungen sollten die Beobachter unterstützen, dimensionsrelevantes Verhalten zu erkennen.
Während der Übungsdurchführung machten sich die Beurteiler Notizen, anhand derer sie nach Abschluß der Übung die Performance jedes Teilnehmers auf einer 6-stufigen Skala bewerteten. Den Beobachtern lag dazu ein auf jede Übung speziell zugeschnittener Beurteilungsbogen vor, der nur die zu bewertenden Dimensionen auflistete. Die Beobachtung und Bewertung erfolgte individuell ohne Kontakt bzw. Diskussion zwischen den Beobachtern. Das Verhalten der Bewerber wurde auf einer sechs-stufigen Ratingsskala bewertet.

Die Ausprägungen lauteten:

6	übertrifft die Anforderungen bei weitem
5	übertrifft die Anforderungen
4	erfüllt die Anforderungen voll
3	erfüllt die Anforderungen mit Abstrichen
2	erfüllt die Anforderungen nur zum Teil
1	erfüllt die Anforderungen nicht

3.3 Fragestellung und Hypothesen

Ableitung der Fragestellung

Wie oben beschrieben, wird bei der Bestimmung der Konstruktvalidität der Frage nachgegangen, was in Assessment Centern gemessen wird. Die Forschung hat hierzu z.T. kontroverse Ergebnisse produziert. Während Kleinmann (1997), Kudisch et al. (1997) und Guldin und Schuler (1997) in jüngster Zeit zeigen konnten, daß neben Übungen auch Dimensionen zur Erklärung der Verhaltensvarianz nötig sind, konnten alle anderen mir bekannten Untersuchungen keine Dimensionsfaktoren extrahieren (vgl. u.a. Bycio et al., 1987, Sackett & Dreher, 1982).

Diese Ergebnisse zeigen, daß weiterhin Forschungsbedarf zur Konstruktvalidität von Assessment Centern besteht. Die von Schuler 1989 aufgestellte Forderung scheint von aktueller Bedeutung zu sein: „Um die [Assessment Center] Methode wesentlich zu verbessern, werden wir um forcierte Konstruktaufklärung nicht herumkommen" (S. 242).

Die Betrachtung der Ergebnisse der Laborstudien von Kleinmann (1997), Kudisch et al. (1997) und Guldin und Schuler (1997) lassen vermuten, daß der Nachweis des Einflusses von Dimensionsbewertungen zur Erklärung der Verhaltensvarianz anhand von Daten aus Assessment Centern der Wirtschaft noch aussteht. Die allgemeine Fragestellung lautet daher: Was wird innerhalb des untersuchten Assessment Centers gemessen, Übungen oder Dimensionen?

Hypothesen

Abgeleitet aus den Konstruktionsprinzipien von Assessment Centern, wonach Personenmerkmale in berufsrelevanten Übungen bewertet werden sollen, wird folgendes zur Konstruktvalidität angenommen:

Beobachter, die ausreichend trainiert wurden, können im Assessment Center Personenmerkmale in verschiedenen Übungen bewerten (Hypothese 1).

Diese Hypothese soll in bezug auf eine faktorenanalytische Auswertung weiter spezifiziert werden:

Für ein Assessment Center, bei dem die Beobachter ausreichend trainiert wurden, wird angenommen, daß ein signifikanter Teil der Verhaltensvarianz durch Dimensionsfaktoren erklärt wird (Hypothese 1a).

Es soll außerdem die Reliabilität des Verfahrens untersucht werden. Ausgehend von den Ergebnissen zur Beobachter-Übereinstimmung (s. Abschnitt 2.4.1) ist folgendes zu vermuten:

Es wird angenommen, daß Beobachter, die ausreichend trainiert wurden, Kandidaten mit genügend hoher Übereinstimmung bewerten (Hypothese 2).

Dabei werden jedoch wegen des Fehlens eines Informationsaustausches zwischen den Beobachtern vor der Bewertung nicht so hohe Werte erwartet wie in Studien (vgl. Schmitt, 1977 und Jones, 1981), wo der Bewertung eine Diskussion vorherging (zum Einfluß eines Informationsaustauschs auf die Interrater-Reliabilität siehe 2.4.1).

3.4 Auswertung

Zur Überprüfung der Fragestellung - unter Berücksichtigung der Diskussion um ein angemessenes Analyseverfahren (s. Abschnitt 2.4.2) - werden in dieser Arbeit verschiedene Methoden eingesetzt, die nun vorgestellt werden.

3.4.1 Analyse der Multitrait-Multimethod-Matrix

Die Analyse der MTMM-Matrix zur Bestimmung der Konstruktvalidität wird im Rahmen der Assessment Center-Forschung häufig verwandt (vgl. Schuler, 1989; Kleinmann, 1997) und soll daher im folgenden kurz erläutert werden.

Auf Assessment Center Verfahren angewendet, stehen die Übungen für die verschiedenen Methoden und die unterschiedlichen Assessment Center Dimensionen für die Traits. Zur gleichzeitigen Überprüfung der konvergenten und der diskriminanten Validität (s.o.) dient die MTMM-Matrix. Darin werden die Korrelationen zwischen mehreren Konstrukten (Multi-Trait) wiedergegeben, die mittels mehrerer Erhebungsmethoden (Multi-Method) erhoben wurden. Zur Veranschaulichung ist im folgenden eine MTMM-Matrix dargestellt.

Tabelle 3.3: MTMM-Matrix

		Methode 1			Methode 2			Methode 3		
	Krite-rium	A1	B1	C1	A2	B2	C2	A3	B3	C3
Methode 1	A1	1,0								
	B1		1,0							
	C1			1,0						
Methode 2	A2				1,0					
	B2					1,0				
	C2						1,0			
Methode 3	A3							1,0		
	B3								1,0	
	C3									1,0

(vgl. Campbell & Fiske, 1959)

Campbell und Fiske (1959) konstatieren, daß unterschiedliche Traits (hier A, B, C) durch unterschiedliche Methoden (Methode 1 bis 3) erfaßt werden. In diesem Beispiel werden die drei Traits (A, B, C) in jeder Methode gemessen. Die Bezeichnung A1 spiegelt somit die Bewertung des Traits A in Methode 1 wider. Die sich ergebenden Korrelationen kann man in drei verschiedene Typen aufteilen, die hier hell, dunkel und schraffiert markiert wurden.

Im folgenden – wie im Ergebnisteil - soll die Terminologie von Campbell und Fiske (1959) verwendet werden, da sich trotz reichlicher Suche leider keine geeigneten deutschen Kurzbegriffe finden ließen. Dabei steht „Trait" für Personenmerkmal, „Method" für Übung sowie „mono" für gleiche (Übung oder Personenmerkmal) und „hetero" für unterschiedliche (Übung oder Personenmerkmal).

Die in der Abbildung dunkel markierten Korrelationen ▬ stellen nach Campbell und Fiske (1959) den Zusammenhang eines Merkmal in verschiedenen Übungen (Monotrait-Heteromethod) dar. Diese Koeffizienten spiegeln die konvergente Validität wider.

Die in der Abbildung hell markierten Korrelationsmöglichkeiten ▬ spiegeln die Heterotrait-heteromethod Korrelationen wider, d.h. unterschiedliche Merkmale, die in verschiedenen Übungen erhoben wurden.

Die in der Abbildung schraffierten Korrelationsmöglichkeiten ▭ stehen für die Korrelationen zwischen Merkmalen innerhalb einer Methode (Heterotrait-monomethod).

Zur Überprüfung der Konstruktvalidität eines Verfahrens mit der MTMM-Matrix postulieren Campbell und Fiske (1959) vier Kriterien:

1. Um von konvergenter Validität sprechen zu können, sollen die Monotrait-Heteromethod Korrelationen signifikant größer als Null sein (konvergente Validitätskoeffizienten).
2. Die Heterotrait-Heteromethod Korrelationen sollen signifikant kleiner sein als die Monotrait-Heteromethod Korrelationen.
3. Die Monotrait-Heteromethod Korrelationen sollen signifikant größer sein als die Heterotrait-Monomethod Korrelationen.
   Die Kriterien zwei und drei entscheiden über die diskriminante Validität.
4. Es soll sich für alle Heterotrait-Monomethod Korrelationen möglichst das gleiche Muster ergeben wie für die Heterotrait-Heteromethod Korrelationen. Das heißt, daß die Rangreihe der Trait-Interkorrelationen in allen Teilmatrizen identisch sein sollte. So sollte z.B. die Korrelation zwischen Trait A und Trait B immer am größten sein, gefolgt von der Korrelation zwischen Trait A und C und Trait B und C (s.a. Bortz & Döring, 1995).

3.4.2 Hauptkomponentenanalyse

Für die Auswertung der MTMM-Matrix sind einige faktorenanalytische Techniken vorgeschlagen worden (vgl. Kleinmann, 1997). Eine Möglichkeit stellt dabei Hauptkomponentenanalyse dar, die auch Sackett und Dreher (1982) in ihrer klassischen Untersuchung verwendeten. Diese explorative Faktorenanalyse hat zum Ziel, „einem größeren Variablensatz eine ordnende Struktur zu unterlegen" (Bortz, 1993, S. 472). Auf Basis der MTMM-Korrelationen wird dabei eine Hauptkomponentenanalyse mit quadrierten multiplen Korrelationen als Kommunalitätenschätzung und anschließender Varimax-Rotation durchgeführt (vgl. Sackett & Dreher, 1982).

Auf eine detaillierte Ausführung soll im Rahmen dieser Arbeit verzichtet werden und auf einschlägige Literatur zum Thema verwiesen werden (u.a. Bortz, 1993; Pawlik, 1968; Revenstorf, 1980), da angenommen werden kann, daß Faktorenanalysen im sozialwissenschaftlichen Kontext mittlerweise bekannt und anerkannt sind (vgl. Bortz, 1993).

3.4.3 Konfirmatorische Faktorenanalyse

Die konfirmatorische Faktorenanalyse auf Basis der Korrelationen (bzw. Kovarianzen) der MTMM-Matrix ist ein relativ junges Verfahren. Es beruht auf der Maximum-Likelihood Methode und kann als ein Teilmodell der linearen Strukturgleichungsmodelle verstanden werden (Bortz, 1993). Dabei werden im Vorwege Hypothesen darüber aufgestellt, welche manifesten oder latenten Variablen (s.u.) durch welche anderen kausal beeinflußt sein könnten. Dieses theoretische

Modell wird dann mit Hilfe der konfirmatorischen Faktorenanalyse anhand der empirischen Datenstruktur getestet. Es läßt sich somit überprüfen, ob die Abweichungen der empirisch ermittelten Faktorladungen, Meßfehlervarianzen und Kovarianzen der Faktoren von den hypothetisch angenommenen Parametern zufällig oder statistisch bedeutsam sind (vgl. Revenstorf, 1980).
Die konfirmatorische Faktorenanalyse läßt sich am besten mit dem Programm LISREL von Jöreskog und Sörbom (1989) durchführen, das daher im folgenden beschrieben werden soll. Das in den siebziger Jahren entwickelte LISREL – Modell (steht für LInear Structural RELationships) ist das bekannteste Verfahren zur konkreten Schätzung von Modellparametern. Es werden dabei lineare strukturale Beziehungen zwischen Variablen untersucht. Es soll zunächst eine kurze Einführung in den LISREL-Ansatz gegeben werden. Dabei wird die Beschreibung auf die Grundbegriffe und Anwendungen beschränkt, die für die Auswertung des vorliegenden Datenmaterials notwendig sind. LISREL soll hier nur zur Überprüfung der Konstruktvalidität des Assessment Centers eingesetzt werden.

Allgemein wird zwischen manifesten (beobachtbaren) und latenten (nicht direkt meßbaren) Variablen unterschieden. Die Ausprägungen auf einer manifesten Variable x ist bedingt durch latente Variablen (ξ; lies: xi) und dem Meßfehler (δ; lies: delta). In der Terminologie von LISREL, die hier verwendet wird, werden die latenten Variablen mit griechischen Buchstaben bezeichnet und die Meßfehler von x als δ. Die latente Variable beeinflußt die manifesten Variablen, was durch die Pfeile symbolisiert wird. Dabei wird die Stärke der Beeinflussung der beobachtbaren Variablen durch die Pfadkoeffizienten λ dargestellt. Es kann außerdem angenommen werden, daß die latenten Variablen interkorreliert sind, was durch Doppelpfeile symbolisiert wird. Folgendes Pfaddiagramm soll diese Zusammenhänge für ein Beispiel mit zwei interkorrelierten latenten (ξ_1 und ξ_2) und vier beobachtbaren x-Variablen darstellen.

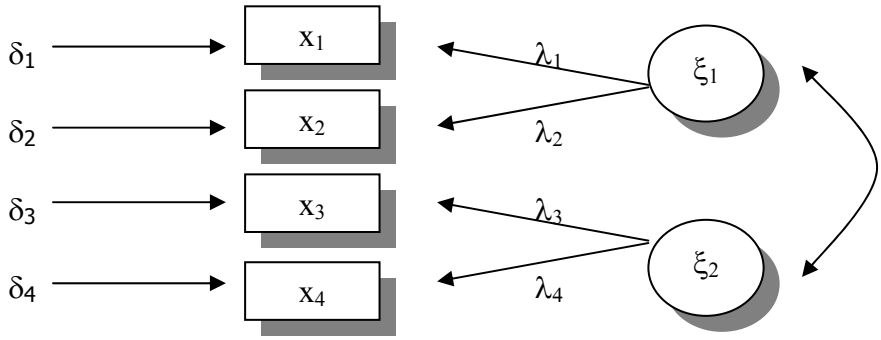


Abbildung 3.1: Beispiel eines Pfaddiagramms (vgl. Backhaus, Erichson, Plinke & Weiber, 1996)

Die Vorgehensweise zur Anwendung von LISREL beschreiben Pfeifer und Schmidt (1987). Danach wird unter Annahme multivariat verteilter manifester Variablen zuerst ein aus den Hypothesen bzw. dem theoretischen Modell abgeleitetes graphisches Modell erstellt. Anschließend werden aus diesem Pfaddiagramm Modellgleichungen abgeleitet, die zur Prüfung durch LISREL notwendig sind. Das obige Pfaddiagramm lautet in Gleichungsform:

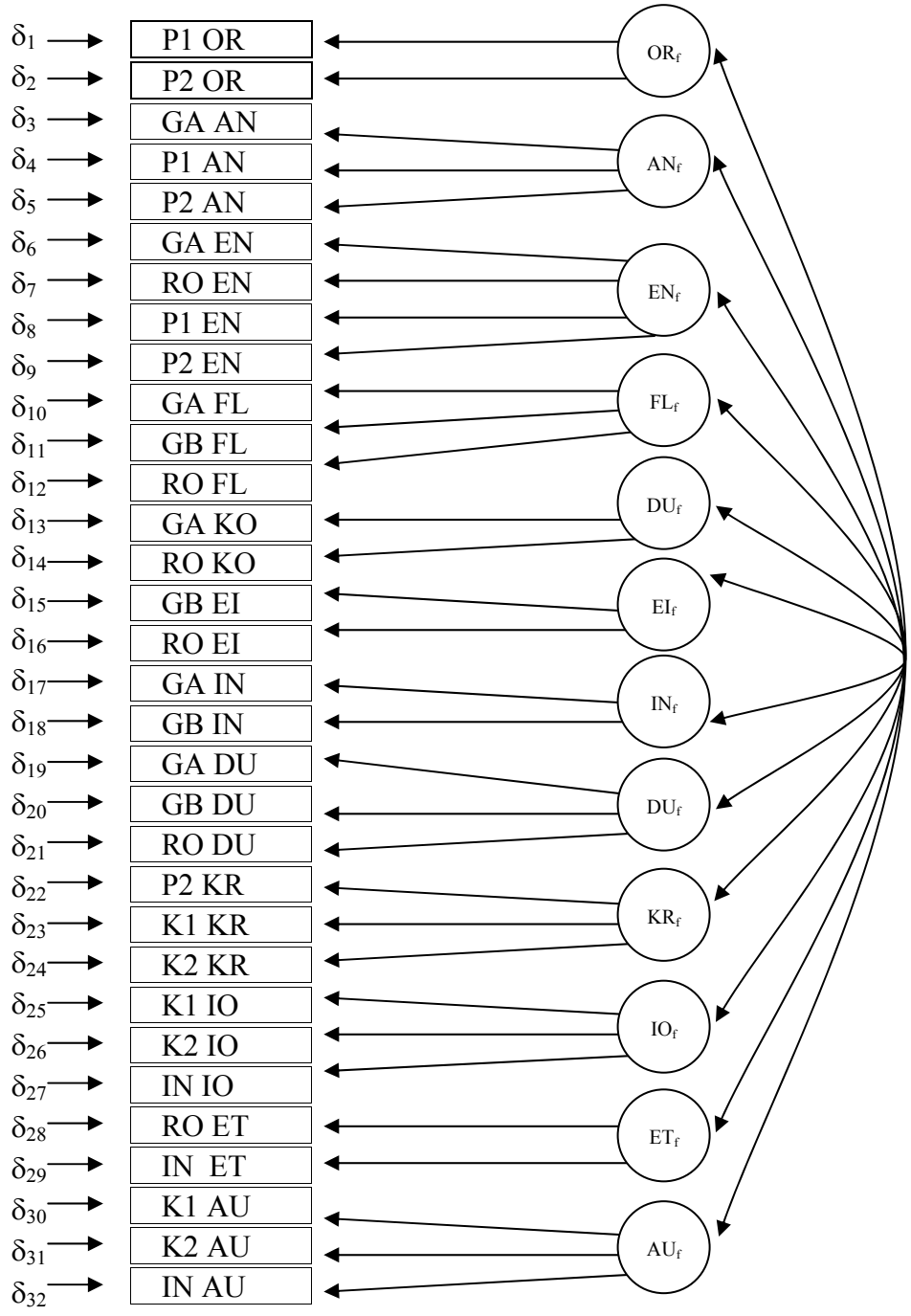
$$x_1 = \lambda_1 * \xi_1 + \delta_1$$
$$x_2 = \lambda_2 * \xi_2 + \delta_2$$
$$x_3 = \lambda_3 * \xi_3 + \delta_3$$
$$x_4 = \lambda_4 * \xi_4 + \delta_4$$

Auf das Modell des Assessment Centers übertragen, stellen die 32 Bewertungen der Einzeldimensionen (s. Tabelle 3.1: Ergebnismatrix) die manifesten Variablen dar. Die Übungen und/ oder die übergeordneten Personenmerkmale könnten in einem theoretischen Modell die latenten, nicht direkt erfaßten Variablen sein.

Diese Überlegungen sollen nun auf das untersuchte Assessment Center übertragen werden: Für die vorliegende MTMM-Matrix (s.o.) mit acht Übungen und zwölf Dimensionen bei insgesamt 32 Einzeldimensions-Bewertungen (x-Variablen) kann man sinnvollerweise drei faktorenanalytische Modelle überprüfen (vgl. Kleinmann, 1997). Die Konstruktion dieses Assessment Centers geht von Dimensionen und Übungen aus, die die Bewertungen bedingen. Auf eine Darstellung der Pfadkoeffizienten λ (s.o.) wurde bei den Modellen zugunsten der besseren Übersicht verzichtet. Dabei wird angenommen, daß die Übungsfaktoren miteinander korreliert sind. Dies geschieht aus der Überlegung heraus, daß bei der Konstruktion von Assessment Centern im allgemeinen davon ausgegangen wird, daß jede Übung einen Teil der Arbeit der späteren Zielposition reflektiert. Somit ist anzunehmen, daß sich die beobachteten Verhaltensstichproben überschneiden (s.a. Bycio et al., 1987). Diesem Zusammenhang wird durch eine Interkorrelation der Faktoren entsprochen, die im Pfaddiagramm (s. Abbildung 3.1) durch Doppelpfeile symbolisiert wird. Darüber hinaus wird aufgrund der im zweiten Kapitel vorgestellten Ergebnisse der Forschung angenommen, daß auch die verschiedenen Dimensionen miteinander korrelieren. Im folgenden sollen die drei zu testenden Modelle als Pfaddiagramm dargestellt werden. Dabei ist es zum besseren Verständnis der Bezeichnungen hilfreich zu wissen, daß der erste Teil des Namens der beobachtbaren Variablen die Übung wiedergibt und der zweite Teil das intendierte Personenmerkmal (Dimension). Die latenten Faktoren sind durch das tiefergestellte $_f$ symbolisiert.

Modell 1

Die Varianzen und Kovarianzen der MTMM-Matrix werden allein durch die zwölf übergeordneten Personenmerkmale der Kandidaten des Assessment Centers und Meßfehlern erklärt. Die 32 Einzeldimensions-Bewertungen werden demnach durch die aus dem Anforderungsprofil abgeleiteten Eigenschaften wie Analysefähigkeit, Kreativität usw. erklärt. Dieses Modell beschreibt die Höhe der Konstruktvalidität. Eine perfekte Passung bedeutet eine perfekte Konstruktvalidität. Die vermutete Beurteilungsstruktur ist durch folgendes faktorenanalytische Modell abbildbar.



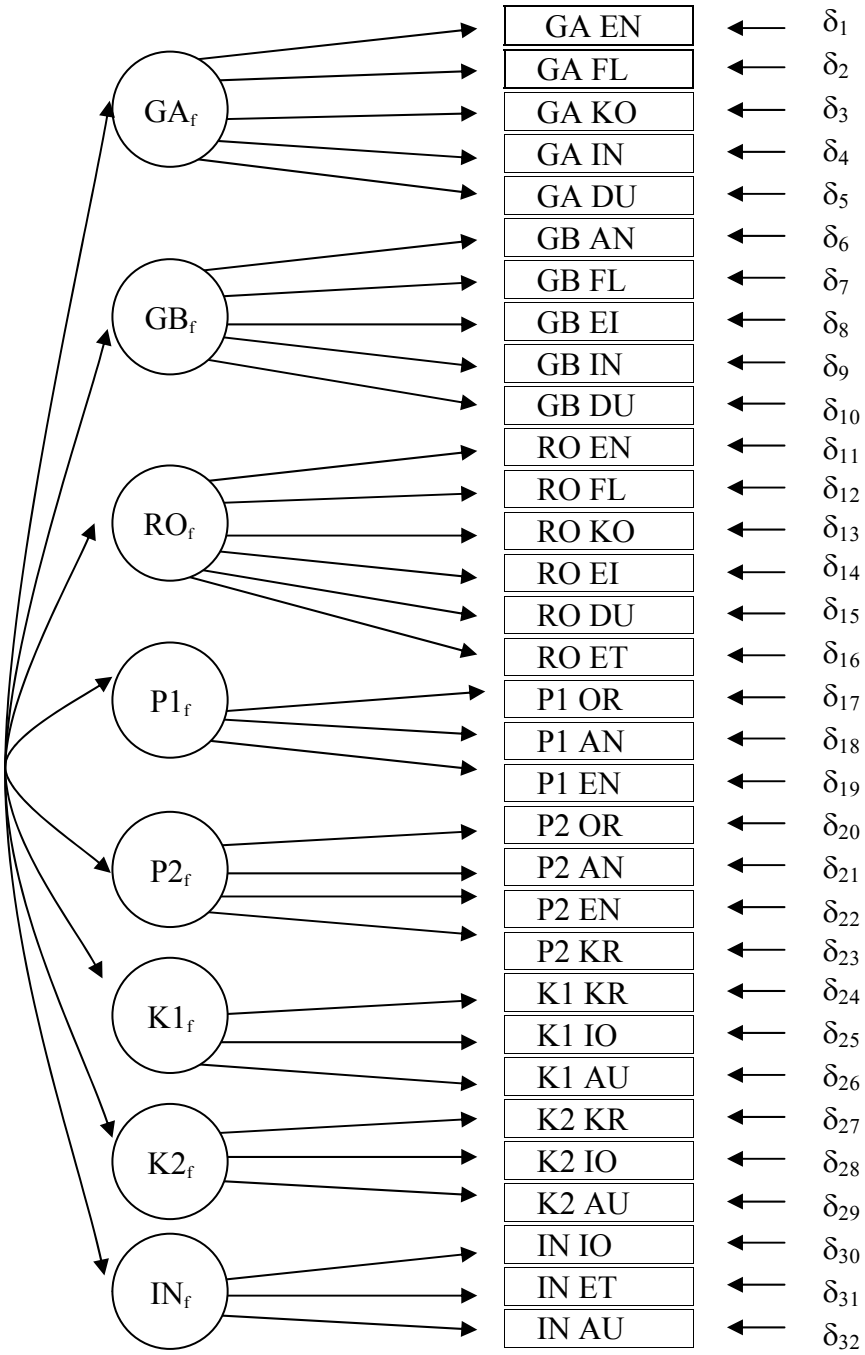
Anmerkung: GA: Gruppendiskussion „A"; GB: Gruppendiskussion „B"; RO: Rollenspiel; P1: Präsentations-Übung 1; P2: Präsentations-Übung 2; K1: Kreativitäts-Übung 1; K2: Kreativitäts-Übung 2; IN: Interview; OR: Organisation; AN: Analysefähigkeit; EN: Entscheidungsfähigkeit; FL: Flexibilität; KO: Kontaktfähigkeit; EI: Einfühlungsvermögen; IN: Integration; DU: Durchsetzung; KR: Kreativität; IO: Innovationsfähigkeit; ET: Ethische Grundhaltung; AU: Ausstrahlung; die latenten Variablen (Faktoren) sind durch Indizes $_f$ kenntlich gemacht.

Abbildung 3.2: Faktorenanalytisches Modell I: Zwölf Personenmerkmale (korreliert). Die manifesten Variablen sind nach den Personenmerkmalen sortiert.

In Abbildung 3.2 sind die latenten Variablen (Faktoren der Personenmerkmale) als Kreise, die manifesten Variablen als Rechtecke dargestellt. Es ist ersichtlich, daß die ersten beiden manifesten Variablen (Organisation in Präsentations-Übung 1 = P1OR und Organisation in Präsentations-Übung 2 = P2OR) von der ersten latenten Variablen „Organisation" (OR_f) und den Meßfehlern δ_1 bzw. δ_2 bestimmt werden. Die dritte (GAAN), vierte (P1AN) und fünfte (P2AN) manifeste Variable werden von dem Faktor Analysefähigkeit AN_f und den Meßfehlern bedingt. Analog dazu erklären die Meßfehler und die interkorrelierten latenten Variablen die weiteren beobachtbaren Variablen. Die latenten Variablen sind interkorreliert. Die Faktoren bilden in diesem Modell die vermuteten übergeordneten Personenmerkmale (Traits) ab, was im testtheoretischen Sinne für eine hohe Validität spräche.

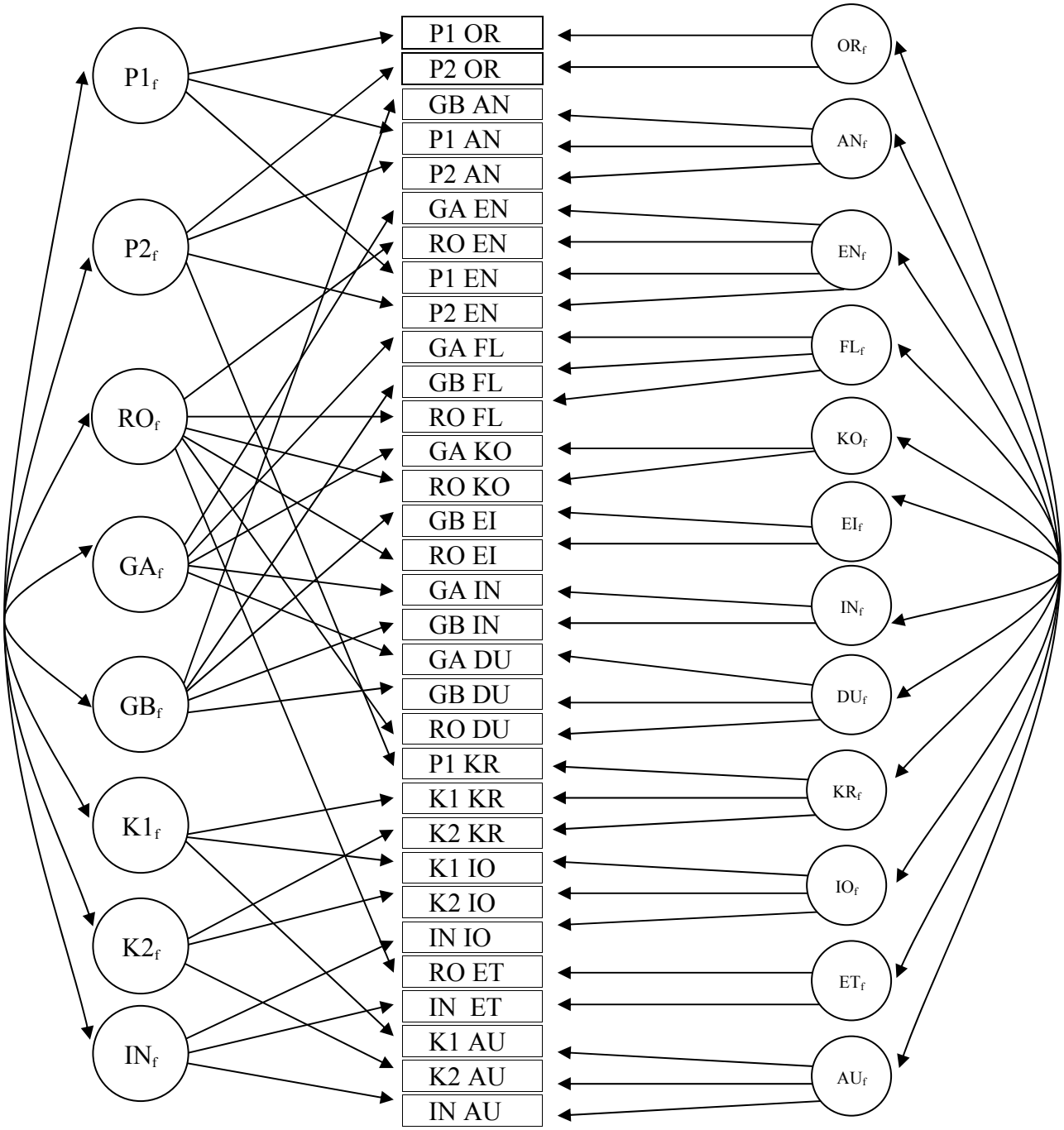
Das zweite mögliche Modell wird dadurch beschrieben, daß die Varianzen und Kovarianzen der manifesten Variablen allein durch Übungseinflüsse (Methodenfaktoren) und Meßfehler bestimmt werden. In Abbildung 3.3 wird das faktorenanalytische Modell 2 als Pfaddiagramm mit acht Übungsfaktoren dargestellt. Die manifesten Variablen sind nach Übungen sortiert. Die latenten Variablen sind - gemäß der Terminologie - wiederum als Kreise und die manifesten Variablen als Rechtecke dargestellt. Die ersten fünf beobachtbaren Variablen GAEN, GAFL, GAKO, GAIN und GADU werden von der latenten Variablen Übungsfaktor Gruppendiskussion „A" (GA_f) und den Meßfehlern δ_1 bis δ_5 bedingt. Der Übungsfaktor Gruppendiskussion „B" (GB_f) und der Meßfehler erklären bei diesem hypothetischen Modell die nächsten fünf Einzelbewertungen. In diesem Sinne werden alle manifesten Variablen bestimmt. Perfekte Passung dieses Modells bedeutet, daß die Bewertungen ausschließlich durch die Übungen bestimmt werden. Dann würde das Assessment Center geringe oder gar keine Konstruktvalidität vorweisen können.

Das dritte denkbare Modell zur Beschreibung der Beurteilungsstruktur in Assessment Centern ist in Abbildung 3.4 dargestellt. Die Varianzen und Kovarianzen der MTMM-Matrix werden durch Personenmerkmals- und Übungsfaktoren bestimmt. Modell 3 ist somit eine Kombination der ersten beiden Modelle. Die auf den ersten Blick verwirrende Pfeilanordnung auf der linken Seite ergibt sich, weil die beobachtbaren Variablen jeweils den Personenmerkmals- und den Übungsfaktoren zugeordnet werden müssen. Die Meßfehler wurden in dieser Darstellung weggelassen. Die manifesten Variablen sind zur besseren Übersicht nach den Dimensionen sortiert, was aber keine weitere Bedeutung hat.



Anmerkung: GA: Gruppendiskussion „A"; GB: Gruppendiskussion „B"; RO: Rollenspiel; P1: Präsentations-Übung 1; P2: Präsentations-Übung 2; K1: Kreativitäts-Übung 1; K2: Kreativitäts-Übung 2; IN: Interview; OR: Organisation; AN: Analysefähigkeit; EN: Entscheidungsfähigkeit; FL: Flexibilität; KO: Kontaktfähigkeit; EI: Einfühlungsvermögen; IN: Integration; DU: Durchsetzung; KR: Kreativität; IO: Innovationsfähigkeit; ET: Ethische Grundhaltung; AU: Ausstrahlung; die latenten Variablen (Faktoren) sind durch Indizes $_f$ kenntlich gemacht.

Abbildung 3.3: Faktorenanalytisches Modell 2: Acht Übungsfaktoren (korreliert). Die manifesten Variablen sind nach Übungen sortiert.



Anmerkung: GA: Gruppendiskussion „A"; GB: Gruppendiskussion „B"; RO: Rollenspiel; P1: Präsentations-Übung 1; P2: Präsentations-Übung 2; K1: Kreativitäts-Übung 1; K2: Kreativitäts-Übung 2; IN: Interview; OR: Organisation; AN: Analysefähigkeit; EN: Entscheidungsfähigkeit; FL: Flexibilität; KO: Kontaktfähigkeit; EI: Einfühlungsvermögen; IN: Integration; DU: Durchsetzung; KR: Kreativität; IO: Innovationsfähigkeit; ET: Ethische Grundhaltung; AU: Ausstrahlung; die latenten Variablen (Faktoren) sind durch Indizes $_f$ kenntlich gemacht.

Abbildung 3.4: Faktorenanalytisches Modell 3: Zwölf Personenmerkmale (korreliert) und acht Übungsfaktoren (korreliert). Die manifesten Variablen sind nach den Personenmerkmalen sortiert.

Das in Abbildung 3.4 dargestellte Pfaddiagramm verdeutlicht folgende Zusammenhänge: Nach diesem hypothetischen Faktormodell werden die 32 Einzelbewertungen (x-Variablen) sowohl von den Eigenschaftsfaktoren (rechte Seite) als auch den Übungsfaktoren (linke Seite) und den Meßfehlern bestimmt. Dabei wurde, wie in den beiden ersten Modellen angenommen, daß die latenten Übungsvariablen und die latenten Eigenschaftsvariablen interkorrelieren. Hier ist es außerdem wichtig, herauszufinden, ob konvergente und diskriminante Validität vorhanden sind. Dies geschieht anhand der Analyse des varianzerklärenden Beitrags jedes Faktors (vgl. Kleinmann, 1997).

Aus den drei Pfaddiagrammen sind in der weiteren Vorgehensweise von LISREL die Modellgleichungen abzuleiten (vgl. Pfeifer und Schmidt, 1987). Dies soll im folgenden exemplarisch für Modell 1 geschehen; Modell 1 lautet in Matrixschreibweise:

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \\ x_9 \\ x_{10} \\ x_{11} \\ x_{12} \\ x_{13} \\ x_{14} \\ x_{15} \\ x_{16} \\ x_{17} \\ x_{18} \\ x_{19} \\ x_{20} \\ x_{21} \\ x_{22} \\ x_{23} \\ x_{24} \\ x_{25} \\ x_{26} \\ x_{27} \\ x_{28} \\ x_{29} \\ x_{30} \\ x_{31} \\ x_{32} \end{pmatrix} = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \lambda_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda_3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda_4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda_5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda_6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda_7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda_8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda_9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda_{10} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda_{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda_{12} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda_{13} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda_{14} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \lambda_{15} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \lambda_{16} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{17} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{18} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{19} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{20} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{21} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{22} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{23} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{24} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{25} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{26} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{27} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{28} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{29} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{30} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{31} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{32} \end{pmatrix} * \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \\ \xi_4 \\ \xi_5 \\ \xi_6 \\ \xi_7 \\ \xi_8 \\ \xi_9 \\ \xi_{10} \\ \xi_{11} \\ \xi_{12} \end{pmatrix} + \begin{pmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \delta_5 \\ \delta_6 \\ \delta_7 \\ \delta_8 \\ \delta_9 \\ \delta_{10} \\ \delta_{11} \\ \delta_{12} \\ \delta_{13} \\ \delta_{14} \\ \delta_{15} \\ \delta_{16} \\ \delta_{17} \\ \delta_{18} \\ \delta_{19} \\ \delta_{20} \\ \delta_{21} \\ \delta_{22} \\ \delta_{23} \\ \delta_{24} \\ \delta_{25} \\ \delta_{26} \\ \delta_{27} \\ \delta_{28} \\ \delta_{29} \\ \delta_{30} \\ \delta_{31} \\ \delta_{32} \end{pmatrix}$$

Abbildung 3.5: Modell 1 in Matrixschreibweise

Die manifesten x-Variablen sind in der gleichen Reihenfolge wie im Pfaddiagramm (s. Abbildung 3.2) angeordnet. Das bedeutet, daß hier x_1 die manifeste Variable Organisation in der Präsentations-Übung 1 (in Abbildung 3.2 P1 OR), x_2 die Variable Organisation in der Präsentations-Übung 2 (in Abbildung 3.2 P2 OR), x_3 die Variable Analysefähigkeit in der Gruppendiskussion A (GA AN) und schließlich x_{32} die beobachtbare Variable Ausstrahlung im Interview (IN AU) darstellt. Die ersten beiden

manifesten Variablen (x_1 und x_2) werden demnach durch das Produkt der Ausprägung der Ladungsmatrix (λ_1 und λ_2) und dem Faktor ξ_1 und dem Meßfehler (δ_1 und δ_2) bestimmt. Der Wert Null auf der Ladungsmatrix bedeutet, daß die latente Variable keinen Einfluß auf die entsprechende x-Variable hat. Die Variablen x_3 , x_4 und x_5 sind im weiteren von der zweiten latenten Variable und den Meßfehlern bedingt. Diese Zusammenhänge sind analog auf die anderen manifesten Variablen zu übertragen.

Ziel ist es nun, wie oben bereits ausgeführt, zu prüfen, ob die empirisch gewonnenen Kovarianzen der MTMM-Matrix mit den spezifizierten LISREL-Modellen gut reproduzierbar sind. In der vorliegenden Untersuchung wird der Fragestellung nachgegangen, was in Assessment Centern gemessen wird, Dimensionen oder Übungen. Zur Beantwortung gibt LISREL verschiedene statistische Maße aus, die im folgenden kurz erläutert werden sollen:

In dieser Arbeit soll wie in vergleichbaren Untersuchungen von Fennekels (1987) und Kleinmann (1997) vorgegangen werden, die die Anpassung der hypothetischen Modelle mit der empirischen Datenstruktur mit Hilfe der χ^2-Prüfgröße getestet haben. Die χ^2-Prüfgröße wird unter der Voraussetzung angewendet, daß die beobachtbaren Variablen multivariat normalverteilt sind. Demnach läßt sich auf der Datengrundlage einer Kovarianzmatrix bei hinreichend großen N (=Anzahl der Versuchspersonen) eine Teststatistik berechnen, die unter der Nullhypothese, daß ein spezifiziertes faktorenanalytisches Modell eine empirische Kovarianzmatrix reproduziert, χ^2 -verteilt ist (vgl. Kleinmann, 1997). Dabei ergeben sich die Freiheitsgrade aus der Differenz der Zahl der Elemente in der Kovarianzmatrix und den zu schätzenden Parametern. Kleinmann (1997) betont, daß die χ^2 -verteilte Größe für den Fall der Übereinstimmung zwischen empirischer und reproduzierter Kovarianzmatrix Null wird. Somit lassen sich faktorenanalytische Modelle statistisch prüfen. Es sei hier jedoch noch auf einen wichtigen Unterschied zum üblichen Vorgehen beim Hypothesentesten hingewiesen: Es ist von einer guten Modellanpassung zu sprechen, wenn die χ^2-Prüfgröße insignifikant ist. Dementsprechend ist es das Ziel, die Nullhypothese zu bestätigen. Die Entwickler von LISREL Jöreskog und Sörbom (1989) unterstreichen jedoch, daß die χ^2-Prüfgröße v.a. bei großen Stichproben leicht zur Verwerfung von Modellen führt, obwohl eine zufriedenstellende Passung vorliegt. Demnach liegt die Bedeutung dieses Tests auch in der Möglichkeit herauszufinden, welches Modell die Daten am besten widerspiegelt. Laut Fennekels (1987) kann man noch von einer guten Anpassung sprechen, wenn das Verhältnis zwischen χ^2-Prüfgröße und Freiheitsgrade 3 zu 1 beträgt.

Neben der χ^2-Prüfgröße werden von LISREL noch weitere Kennwerte berechnet:

- „Goodness-of-Fit Index" (GFI) von Jöreskog und Sörbom (1989) – Maß für die Varianz und Kovarianz, die durch das hypothetische Modell erklärt wird. Es geht bei steigender Anpassungsgüte Passung gegen 1 und bei schlechter Anpassung gegen Null.
- „Adjusted Goodness-of-Fit Index" (AGFI) von Jöreskog und Sörbom (1989) – ebenfalls ein Maß dafür, wie gut ein Modell die Varianz und Kovarianz der Daten erklärt unter Berücksichtigung der Freiheitsgrade. Perfekte Anpassungsgüte ist bei 1 gegeben; keine bei Werte nahe Null.
- „Root Mean Square Residual" – Differenz zwischen der Varianz in der Stichprobe und der geschätzten Kovarianzmatrix, die um so näher bei Null liegt, je besser die Daten durch das Modell erklärt werden.

3.4.4 Vorgehensweise der Auswertung

Die Beantwortung der Fragestellung und die Prüfung der Hypothesen vollzieht sich in mehreren Schritten:

Zuerst soll die Reliabilität untersucht werden. Dabei möchte ich die von Lammers (1992) vorgeschlagene Vorgehensweise der Bestimmung der Reliabilität des Assessment Centers verfolgen, der nur die Interrater-Reliabilität für sinnvoll hielt. Für diese Vorgehensweise spricht auch der Aufbau dieses Assessment Centers (s. Abschnitt 3.1), in dem jeder Bewerber in der Regel von zwei Beobachtern bewertet wird. Dieses Setting erlaubt somit die Bestimmung der Beobachter-Übereinstimmung für alle 32 Einzeldimensionen. Die Übereinstimmung der Bewertungen verschiedener Beobachter soll mit Hilfe des Produkt-Moment-Korrelationskoeffizienten untersucht werden. Dieses Vorgehen geschieht, um die Ergebnisse mit denen anderer Studien zu vergleichen, die auch diesen Korrelationskoeffizienten verwendeten (u.a. Scholz, 1994; Fennekels, 1987). Außerdem gilt laut Bortz, Lienert und Boehnke (1990), daß die Verletzungen der Voraussetzung des Produkt-Moment-Korrelationskoeffizienten „in der Regel keinen nennenswerten Einfluß auf die Validität des parametrischen Signifikanztests" (S. 44) haben. Die Signifikanztests gelten demnach als äußerst robust.

Zur Beantwortung der Frage, was in diesem Assessment Center gemessen wird, sollen die drei oben vorgestellten Analysemethoden eingesetzt werden. Die Analyse der MTMM-Matrix nach Campbell und Fiske (1959) und die Hauptkomponentenanalyse sind die bisher gebräuchlichsten Verfahren zur Bestimmung der Konstruktvalidität von Assessment Centern. Daher sollen sie auch hier Verwendung finden. Um die Kritik an diesen Auswertungsverfahren von u.a. Fennekels (1987) und Kleinmann (1997) zu berücksichtigen, wird außerdem die konfirmatorische Faktorenanalyse mittels LISREL angewendet (zur Diskussion um angemessene Auswertungsmethoden siehe Abschnitt 2.4.2).

4 Ergebnisse

Gegenstand dieses Kapitels ist die Darstellung der Ergebnisse zur Interrater-Reliabilität und Konstruktvalidität. Die Hypothesenprüfung findet sich am Ende dieses Kapitels.

Zuerst sollen die Voraussetzung der in letzten Kapitel vorgestellten Analysemethoden und der hier angewendeten Korrelationskoeffizienten überprüft werden. Die Anwendung der faktorenanalytischen Verfahren und die Interpretation der Korrelationen setzt voraus, daß die Daten multivariat normalverteilt und intervallskaliert sind. Laut Bortz und Döring (1995) kann für Rating-Skalen in den Sozialwissenschaften Intervallskalenniveau angenommen werden. Somit gilt diese Voraussetzung als erfüllt. Die Voraussetzung der Normalverteilung wurde mit Hilfe des Kolmogorov-Smirnov Anpassungstests überprüft. Es zeigte sich dabei, daß die Annahme für keine Einzeldimension aufrechtzuerhalten ist. Es kann jedoch aufgrund der Überlegungen von Wittenberg (1991) vermutet werden, daß in der Grundgesamtheit die Daten normalverteilt sind. Wittenberg (1991) schlägt als „Daumenpeilung" zur Prüfung der Normalverteilung vor, die Schiefe und den Exzeß zu überprüfen. Danach sind deutliche Abweichungen von Null (Schiefe und Exzeß-Werte > □1,96|; vgl. Wittenberg, 1991, S.72) ein Indiz dafür, daß keine Normal-verteilung vorliegt. Die in der Tabelle 4.1 dargestellten Ergebnisse zur Schiefe und Exzeß liegen alle deutlich unter diesem Wert und unterstützen daher die Vermutung, daß eine Normalverteilung in der Grundgesamtheit vorliegt.

Letztlich soll das Datenmaterial mit Hilfe der oben vorgestellten Verfahren analysiert werden, um die Ergebnisse mit anderen Studien vergleichen zu können, die bei Verletzung der Voraussetzungen genauso vorgegangen sind (vgl. Bycio et al., 1987). Die Auswirkungen dieses Sachverhalts auf die Interpretation wird jeweils im Zusammenhang mit den verschiedenen Analysemethoden diskutiert.

Zu Beginn der methodischen Auswertung wurden die Daten deskriptiv erfaßt. Dabei soll folgende Tabelle eine Übersicht über Mittelwerte, Standardabweichungen, Schiefe und Exzeß der Bewertungen der Dimensionen geben. Die Tabelle ist nach Übungen sortiert.

Tabelle 4.1: Deskriptive Statistiken der Assessment Center Dimensionen

Einzeldimension in Übung	Mittelwert	Standard-abweichung	Schiefe	Exzeß	N
Gruppendiskussion „A"					
Entscheidungsfähigkeit	3.90	1.03	-0.07	0.48	316
Flexibilität	3.74	0.98	-0.44	0.06	315
Kontaktfähigkeit	3.89	1.07	-0.34	-0.32	315
Integration	3.83	1.02	-0.47	-0.05	315
Durchsetzung	3.67	1.19	-0.31	-0.56	315
Gruppendiskussion „B"					
Analysefähigkeit	3.50	1.13	-0.05	-0.31	310
Flexibilität	3.49	1.10	-0.39	-0.28	311
Einfühlungsvermögen	3.57	1.07	-0.33	-0.07	310
Integration	3.63	1.11	-0.55	-0.23	310
Durchsetzung	3.37	1.17	-0.18	-0.53	310
Rollenspiel					
Entscheidungsfähigkeit	3.78	1.10	-0.30	-0.12	312
Flexibilität	3.58	1.18	-0.22	-0.45	312
Kontaktfähigkeit	3.70	1.15	-0.28	-0.34	313
Einfühlungsvermögen	3.57	1.32	-0.16	-0.71	313
Durchsetzung	3.63	1.19	-0.11	-0.43	312
Ethische Grundhaltung	3.85	1.12	-0.55	0.12	313
Präsentations – Übung 1					
Organisation	3.53	1.32	-0.19	-0.78	314
Analysefähigkeit	3.52	1.30	-0.14	-0.73	313
Entscheidungsfähigkeit	3.83	1.19	-0.53	-0.18	313
Präsentations – Übung 2					
Organisation	3.38	1.06	-0.04	-0.38	312
Analysefähigkeit	3.24	1.18	0.16	-0.51	312
Entscheidungsfähigkeit	3.49	1.07	-0.18	-0.25	312
Kreativität	3.23	1.19	0.13	-0.60	312
Kreativitäts – Übung 1					
Kreativität	3.32	1.24	0.06	-0.55	317
Innovationsfähigkeit	3.32	1.16	-0.07	-0.47	317
Ausstrahlung	3.57	1.14	-0.37	-0.26	317
Kreativitäts – Übung 2					
Kreativität	3.31	1.32	0.01	-0.84	314
Innovationsfähigkeit	3.35	1.26	-0.06	-0.77	314
Ausstrahlung	3.63	1.17	-0.37	-0.29	313
Interview					
Innovationsfähigkeit	3.96	1.08	-0.64	0.25	317
Ethische Grundhaltung	4.36	0.92	-0.79	1.08	316
Ausstrahlung	4.26	1.10	-0.44	-0.16	317

Anmerkung: Abweichungen von (N = 317) sind auf fehlende Einträge zurückzuführen. Die Mittelwerte ergeben sich aus der 6-stufigen Skala: 6 = übertrifft die Anforderungen bei weitem; 5 = übertrifft die Anforderungen; 4 = erfüllt die Anforderungen voll; 3 = erfüllt die Anforderungen mit Abstrichen; 2 = erfüllt die Anforderungen nur zum Teil; 1 = erfüllt die Anforderungen nicht.

Die Mittelwerte der Assessment Center Bewertungen weisen eine Spannweite von 3.23 (Kreativität in der Präsentations-Übung 1) bis 4.36 (Ethische Grundhaltung im Interview) auf. Der Mittelwert aller Bewertungen ist 3.62. Die Standardabweichungen liegen zwischen 0.92 (Ethische Grundhaltung im Interview) und 1.32 (Einfühlungsvermögen im Rollenspiel, Organisation in der Präsentations-Übung 1 und Kreativität in der Kreativitäts-Übung 2). Es wurden Werte der Schiefe im Bereich von – 0.79 bis 0.16 ermittelt. Deutlich ist, daß die Verteilungen der meisten Assessment Center Dimensionen rechtssteil sind (Schiefe < 0). Nur die Bewertungen der Kreativität in den beiden Kreativitäts-Übungen und die Bewertungen der Analysefähigkeit und Kreativität in der Präsentations-Übung 2 haben positive Schiefewerte.
Der Exzeß der Bewertungen liegt zwischen – 0.84 (Kreativität in Kreativitäts-Übung 2) und 1.08 (Ethische Grundhaltung im Interview), wobei die meisten Werte negativ sind, was auf eine breitgipflige Verteilung deutet.

4.1. Ergebnisse zur Interrater-Reliabilität

Die Interrater-Reliabilität der Bewertungen, die in Form des Produkt-Moment-Korrelationskoeffizienten berechnet wurde, soll in der folgenden Tabelle für alle Einzeldimensionen aller Übungen wiedergegeben werden. Die Koeffizienten spiegeln den Zusammenhang zwischen den Bewertungen zweier Beobachter wider, die denselben Kandidaten in einer Übungen eingeschätzt haben. Dabei sind auch die gemittelten Interrater-Reliabilitäten der Übungen dargestellt, die mit Hilfe Fisher´s z-Transformation ermittelt wurden.

Da jeder Kandidat des Assessment Centers in jeder Übung von zwei Beobachtern bewertet wurde, läßt sich die Übereinstimmung der zwei Bewertungen für alle 32 Einzelbeurteilungen ermitteln. Die Korrelationen sind in obiger Tabelle dargestellt. Die Interrater-Reliabilitäten der Bewertungen haben eine Spannweite von r = 0.33 (Flexibilität in Gruppendiskussion „A") bis r = 0.76 (Analysefähigkeit in Präsentations-Übung 1), wobei die meisten Korrelationskoeffizienten über r = 0.50 liegen. Alle ermittelten Koeffizienten sind signifikant auf dem 1% Niveau bei einseitiger Testung.
Auffällig ist die große Spannweite zwischen den Koeffizienten der Beobachter-Übereinstimmung in den 32 Einzeldimensionen. Es bietet sich ein heterogenes Bild der Höhe der Koeffizienten, obwohl alle Korrelationen hochsignifikant sind. Dieses Bild zeigt sich auch in den gemittelten Übungswerten zur Beobachter-Übereinstimmung: Die zusammengefaßten Übungs-Reliabilitäten liegen zwischen r = 0.43 (Gruppendiskussion „A") und r = 0.71 (Präsentations-Übung 1). Die Beobachter sind scheinbar in verschiedenen Übungen unterschiedlich gut in der Lage, Kandidaten hinsichtlich der Personenmerkmale zu beurteilen.

Tabelle 4.2: Interrater-Reliabilität der Bewertungen

Einzeldimension in Übung	Interrater-Reliabilität	***N***
Gruppendiskussion „A"		
Entscheidungsfähigkeit	0.49	308
Flexibilität	0.33	307
Kontaktfähigkeit	0.43	308
Integration	0.34	308
Durchsetzung	0.55	309
insgesamt	0.43	
Gruppendiskussion „B"		
Analysefähigkeit	0.53	278
Flexibilität	0.48	276
Einfühlungsvermögen	0.45	277
Integration	0.50	276
Durchsetzung	0.51	277
insgesamt	0.49	
Rollenspiel		
Entscheidungsfähigkeit	0.52	276
Flexibilität	0.52	275
Kontaktfähigkeit	0.59	277
Einfühlungsvermögen	0.61	278
Durchsetzung	0.54	276
Ethische Grundhaltung	0.55	278
insgesamt	0.56	
Präsentations – Übung 1		
Organisation	0.73	305
Analysefähigkeit	0.76	304
Entscheidungsfähigkeit	0.64	303
Insgesamt	0.71	
Präsentations – Übung 2		
Organisation	0.57	300
Analysefähigkeit	0.63	300
Entscheidungsfähigkeit	0.55	300
Kreativität	0.57	301
Insgesamt	0.58	
Kreativitäts – Übung 1		
Kreativität	0.61	302
Innovationsfähigkeit	0.57	303
Persönliche Ausstrahlung	0.59	305
Insgesamt	0.59	
Kreativitäts – Übung 2		
Kreativität	0.57	300
Innovationsfähigkeit	0.61	300
Persönliche Ausstrahlung	0.56	301
Insgesamt	0.56	
Interview		
Innovationsfähigkeit	0.73	305
Ethische Grundhaltung	0.58	302
Persönliche Ausstrahlung	0.69	305
Insgesamt	0.67	

Anmerkung: Alle Korrelationen sind signifikant auf dem 1 % Niveau bei einseitiger Testung. Die Schwankungen von N sind auf fehlende Werte zurückzuführen. Alle zusammengefaßten (Übungs-) Korrelationen sind mittels Fisher´s z-Transformation berechnet.

Die Einschätzung, daß die Beobachter in den verschiedenen Übungen unterschiedlich gut übereinstimmen, zeigt die folgende Tabelle, in der die signifikanten Unterschiede zwischen den Korrelationen dargestellt werden.

Tabelle 4.3: Signifikante Unterschiede zwischen den Interrater-Reliabilitäten der Übungen. Die Tabelle ist nach Höhe der Beobachter-Übereinstimmung sortiert.

	P1	IN	K1	P2	K2	RO	GA	GB
Präsentations-Übung 1								
Interview								
Kreativitäts-Übung 1	*							
Präsentations-Übung 2	*							
Kreativitäts-Übung 2	*	*						
Rollenspiel	*	*						
Gruppendiskussion „A"	*	*	*	*				
Gruppendiskussion „B"	*	*	*	*	*	*		

Anmerkung: Grundlage des Vergleichs waren gemittelte Korrelationen. P1: Präsentations-Übung 1; IN: Interview; K1: Kreativitäts-Übung 1; P2: Präsentations-Übung 2; K2: Kreativitäts-Übung 2; RO: Rollenspiel; GA: Gruppendiskussion „A"; GB: Gruppendiskussion „B".

Es wird deutlich, daß die Beobachter-Übereinstimmung in den beiden Gruppendiskussionen schlechter ist als in allen anderen Übungen. Die drei Interaktionsübungen (Rollenspiel, Gruppendiskussion A + B) weisen die niedrigsten Korrelationen auf; die Interrater-Reliabilität in der Gruppendiskussion ist sogar signifikant kleiner als in allen Einzelübungen.
Die Werte zur Beobachter-Übereinstimmung sollen im folgenden hinsichtlich der zwölf übergeordneten Personenmerkmale untersucht werden. Dabei wurden die Korrelationen über die Fisher z-Transformation berechnet.

Tabelle 4.4: Interrater-Reliabilitäten der Assessment Center Dimensionen

Dimension	Interrater-Reliabilität
Organisation	0.66
Analysefähigkeit	0.65
Entscheidungsfähigkeit	0.55
Flexibilität	0.44
Kontaktfähigkeit	0.51
Einfühlungsvermögen	0.54
Integration	0.42
Durchsetzung	0.53
Kreativität	0.58
Innovationsfähigkeit	0.64
Ethische Grundhaltung	0.57
Ausstrahlung	0.62

Anmerkung: Alle Korrelationen sind signifikant auf dem 1 % Niveau bei einseitiger Testung. Alle Korrelationen sind mittels Fisher´s z-Transformation berechnet.

Die Beobachter-Übereinstimmung liegt für die Assessment Center Dimensionen im Bereich zwischen r = 0.42 für die Dimension „Integration" und r = 0.66 für „Organisation". Die Interrater-Reliabilität der verschiedenen übergeordneten Dimensionen unterscheiden sich somit z.T. deutlich untereinander. Insgesamt ist die Beobachter-Übereinstimmung der Dimensionsbewertungen ausreichend, zehn der zwölf ermittelten Korrelationskoeffizienten sind größer als r =0.50.

Bei genauerer Betrachtung wird deutlich, daß bei Eigenschaften, die in Übungen ausschließlich im Rollenspiel und den beiden Gruppendiskussionen beobachtet wurden, insgesamt niedrigere Interrater-Reliabilität zu finden sind. Die Korrelationen in den Dimensionen Flexibilität, Kontaktfähigkeit, Einfühlungsvermögen, Integration und Durchsetzung liegen zwischen 0.42 und 0.54 und sind damit alle kleiner als die Interrater-Reliabilität der restlichen Dimensionen, die alle (auch) in den anderen fünf Übungen bewertet wurden (kleinster Wert: 0.55). Diesen Zusammenhang kann man als Indiz dafür deuten, daß die Beobachter-Übereinstimmung viel mehr von der Übung abhängt als vom Konstrukt der Assessment Center Dimension.

4.2. Ergebnisse zur Konstruktvalidität

In diesem Abschnitt werden nun die Ergebnisse zur Konstruktvalidität des untersuchten Assessment Centers wiedergegeben. Dabei wird zuerst auf die Analyse der MTMM-Matrix eingegangen; die Ergebnisse der Hauptkomponentenanalyse und der konfirmatorischen Faktorenanalyse komplettieren den Abschnitt.

4.2.1 Analyse der MTMM-Matrix

Die folgende Tabelle zeigt für die beobachteten Eigenschaften und Übungen die Korrelationen der Multitrait-Multimethod-Matrix. Dabei werden die 32 Einzelbeurteilungen, die jeder Kandidat insgesamt erhält (vgl. Abschnitt 3.1), miteinander korreliert. Die ermittelten Koeffizienten sind Produkt-Moment-Korrelationen. Die Heterotrait-Monomethod Korrelationen (vgl. Abschnitt 3.4) sind schattiert, die Monotrait-Heteromethod Korrelationen fett und die Heterotrait-Heteromethod Koeffizienten normal abgebildet.

Tabelle 4.5: MTMM-Korrelationsmatrix für alle Dimensionen aller Übungen; die Monotrait-Heteromethod Korrelationen sind **fett** gedruckt, die Heterotrait-Monomethod Korrelationen sind schattiert und die Heterotrait-Heteromethod Korrelationen normal gedruckt.

Kriterium in Übung	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Gruppendiskussion „A"																
1 Entscheidungsfähigkeit																
2 Flexibilität	.67															
3 Kontaktfähigkeit	.71	.73														
4 Integration	.64	.80	.77													
5 Durchsetzung	.83	.64	.73	.67												
Gruppendiskussion „P"																
6 Analysefähigkeit	.38	.29	.25	.26	.34											
7 Flexibilität	.38	**.30**	.31	.27	.30	.81										
8 Einfühlungsvermögen	.32	.30	.29	.28	.25	.76	.85									
9 Integration	.32	.29	.31	**.29**	.28	.76	.83	.82								
10 Durchsetzung	.42	.29	.32	.31	**.40**	.81	.80	.77	.79							
Rollenspiel																
11 Entscheidungsfähigkeit	**.30**	.20	.23	.24	.34	.29	.29	.28	.26	.33						
12 Flexibilität	.26	**.27**	.27	.26	.29	.22	**.27**	.28	.25	.25	.66					
13 Kontaktfähigkeit	.26	.28	**.30**	.32	.31	.23	.27	.32	.28	.28	.65	.84				
14 Einfühlungsvermögen	.19	.29	.25	.28	.24	.16	.20	**.23**	.21	.17	.50	.80	.86			
15 Durchsetzung	.24	.18	.23	.20	**.33**	.27	.28	.26	.27	**.35**	.83	.72	.74	.63		
16 Ethische Grundhaltung	.16	.23	.21	.25	.21	.17	.21	.27	.25	.21	.53	.75	.83	.83	.64	
Präsentations – Übung 1																
17 Organisation	.18	.20	.18	.22	.20	.36	.34	.31	.29	.31	.15	.19	.22	.20	.19	.18
18 Analysefähigkeit	.16	.18	.15	.18	.15	**.30**	.28	.28	.23	.26	.13	.19	.23	.19	.18	.14
19 Entscheidungsfähigkeit	**.15**	.20	.19	.22	.20	.33	.30	.27	.25	.27	**.19**	.24	.27	.22	.25	.21
Präsentations – Übung 2																
20 Organisation	.33	.24	.19	.19	.25	.35	.33	.33	.25	.32	.29	.25	.23	.15	.23	.07
21 Analysefähigkeit	.29	.24	.20	.20	.23	**.37**	.34	.33	.27	.35	.32	.29	.26	.18	.26	.11
22 Entscheidungsfähigkeit	**.35**	.23	.23	.19	.28	.35	.35	.32	.27	.37	**.36**	.29	.25	.18	.32	.07
23 Kreativität	.31	.27	.27	.23	.27	.29	.33	.35	.22	.31	.32	.28	.28	.20	.26	.11
Kreativitäts – Übung 1																
24 Kreativität	.33	.25	.33	.23	.30	.32	.38	.36	.36	.35	.25	.22	.23	.19	.27	.23
25 Innovationsfähigkeit	.36	.24	.32	.24	.31	.36	.40	.39	.39	.40	.29	.24	.24	.19	.28	.23
26 Ausstrahlung	.35	.29	.37	.32	.31	.32	.37	.37	.38	.37	.31	.27	.29	.20	.30	.25
Kreativitäts – Übung 2																
27 Kreativität	.35	.30	.36	.29	.36	.35	.44	.38	.40	.37	.28	.29	.31	.21	.29	.26
28 Innovationsfähigkeit	.35	.28	.34	.27	.34	.33	.40	.35	.35	.32	.23	.23	.26	.16	.25	.20
29 Ausstrahlung	.37	.34	.42	.36	.37	.33	.41	.37	.37	.35	.32	.33	.39	.27	.34	.31
Interview																
30 Innovationsfähigkeit	.34	.30	37	.31	.36	.34	.38	.35	.37	.43	.34	.42	.42	.35	.36	.36
31 Ethische Grundhaltung	.22	.24	.30	.28	.25	.11	.19	.20	23	.22	.27	.32	.38	.36	.30	.36
32 Ausstrahlung	.35	.32	.39	.33	.34	.27	.32	.29	.31	.34	.33	.38	.43	.37	.34	.34

Kriterium in Übung	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
Gruppendiskussion „A"																
1 Entscheidungsfähigkeit																
2 Flexibilität																
3 Kontaktfähigkeit																
4 Integration																
5 Durchsetzung																
Gruppendiskussion „P"																
6 Analysefähigkeit																
7 Flexibilität																
8 Einfühlungsvermögen																
9 Integration																
10 Durchsetzung																
Rollenspiel																
11 Entscheidungsfähigkeit																
12 Flexibilität																
13 Kontaktfähigkeit																
14 Einfühlungsvermögen																
15 Durchsetzung																
16 Ethische Grundhaltung																
Präsentations – Übung 1																
17 Organisation																
18 Analysefähigkeit	.92															
19 Entscheidungsfähigkeit	.88	.89														
Präsentations – Übung 2																
20 Organisation	**.32**	.34	.32													
21 Analysefähigkeit	.33	**.32**	.31	.85												
22 Entscheidungsfähigkeit	.24	.26	**.26**	.86	.82											
23 Kreativität	.24	.25	.26	.77	.76	.80										
Kreativitäts – Übung 1																
24 Kreativität	.25	.25	.25	.31	.30	.26	**.26**									
25 Innovationsfähigkeit	.27	.24	.26	.33	.31	.30	28	.91								
26 Ausstrahlung	.29	.29	.31	.33	.30	.31	.30	.77	.81							
Kreativitäts – Übung 2																
27 Kreativität	.31	.29	.32	.28	.27	.27	**.25**	**.50**	.52	.48						
28 Innovationsfähigkeit	.31	.30	.32	.23	.24	.23	.21	.48	**.49**	.46	.92					
29 Ausstrahlung	.31	.30	.34	.24	.25	.25	.27	.49	.51	**.56**	.81	.81				
Interview																
30 Innovationsfähigkeit	.25	.28	.28	.34	.33	.31	.31	.43	**.46**	.44	.46	**.43**	.49			
31 Ethische Grundhaltung	.11	.13	.16	.11	.11	.10	.12	.29	.30	.33	.35	.32	.40	.70		
32 Ausstrahlung	.26	.26	.29	.25	.23	.23	.26	.35	.38	**.44**	.38	.36	**.48**	.78	.78	

Anmerkung: Alle Werte sind Korrelationen von gemittelten Dimensionsbewertungen. N liegt im Bereich zwischen 302 und 316. Abweichung sind auf fehlende Bewertungen zurückzuführen. Korrelation von r > 0,13 sind signifikant auf dem 1% Niveau bei einseitiger Testung. Auf eine Darstellung von N und eine Markierung der signifikanten Werte wurde zugunsten einer besseren Übersicht verzichtet.

Die Analyse der Korrelationen der Matrix nach Campbell und Fiske (1959) erfolgt anhand von vier Kriterien (vgl. Abschnitt 3.4). Sind alle vier erfüllt, so handelt es sich um ein konstruktvalides Instrument.

1. Um von konvergenter Validität sprechen zu können, sollen die Monotrait-Heteromethod Korrelationen signifikant größer als Null sein. Alle Monotrait-Heteromethod Korrelationen weichen auf dem 1 % Niveau bei einseitiger Testung signifikant von Null ab. Die Koeffizienten liegen dabei zwischen r = 0.15 und 0.56. Die mittlere Korrelation beträgt r = 0.35 (ermittelt über Fisher´s z-Trans-formation). Somit ist das Kriterium zur konvergenten Validität erfüllt. Problematisch ist jedoch, daß fast alle Korrelationen der MTMM-Matrix signifikant sind. Das bedeutet, daß selbst die gemittelte Heterotrait-Heteromethod Korrelation signifikant von Null verschieden ist und somit das erste Kriterium erfüllen würde. Auffällig sind im weiteren die Unterschiede zwischen verschiedenen Dimensionen. Während die Monotrait-Heteromethod Korrelation für Innovationsfähigkeit und Ausstrahlung im Mittel bei r = 0.46 bzw. r = 0.49 liegt, beträgt sie für Entscheidungsfähigkeit nur r = 0.27.

2. Die Heterotrait-Heteromethod Korrelationen sollen signifikant kleiner sein als die Monotrait-Heteromethod Korrelationen. Die Heterotrait-Heteromethod Korrela-tionen (bei einer Rang von r = 0.07 bis r = 0.52) sind entgegen der Erwartung nicht signifikant kleiner als die Monotrait-Heteromethod Korrelationen. Die gemittelte Monotrait-Heteromethod Korrelation von r = 0.35 ist dabei zwar größer als die gemittelte (Fisher´s z-Transformation) Heterotrait-Heteromethod Korrelation von r = 0.29. Dieser Unterschied läßt sich allerdings statistisch nicht absichern, so daß das zweite Kriterium als nicht erfüllt gilt.

3. Die Monotrait-Heteromethod Korrelationen (gemittelte Korrelation von r = 0.35) sind entgegen der Erwartung nicht größer als die Heterotrait-Monomethod Korrelationen, die bei einer Range von r = 0.50 bis 0.92 eine gemittelte Korrela-tion von 0.79 ergeben. Damit sind die Heterotrait-Monomethod Korrelationen sogar signifkant größer bei zweiseitiger Testung auf dem 1 % Niveau und in der gesamten Matrix die deutlich größten. Das dritte Kriterium ist also nicht erfüllt.

4. Es soll sich für alle Heterotrait-Monomethod Korrelationen möglichst das gleiche Muster ergeben wie für die Heterotrait-Heteromethod Korrelationen. Dabei sollten die Rangreihen der Trait-Interkorrelationen in allen Teilmatrizen möglichst identisch sein. Die Korrelationen der MTMM-Matrix ähneln sich entgegen der Erwartung in ihrem Muster eher innerhalb der Übungen (auch bei Heterotrait-Heteromethod Korrelationen) als innerhalb der Assessment Center Dimensionen. Dies wird deutlich, wenn man die Matrix spaltenweise betrachtet. So zeigen die Korrelationen der ersten Spalte ähnlichere Werte für jede *Übung* als für die *Dimension* Entscheidungsfähigkeit (Range von r = 0.15 bis r = 0.35). Ein gleiches Muster der Trait-Interkorrelationen ist jedoch weder für die Heterotrait-Monomethod noch für die anderen Korrelationen zu erkennen, so daß auch das vierte Kriterium als nicht erfüllt gilt.

Gemäß den Kriterien von Campbell und Fiske (1959) ist das untersuchte Assessment Center somit nicht konstruktvalide. Die Analyse deutet nicht daraufhin, daß Personenmerkmale in berufsrelevanten Übungen gemessen werden. Im Gegenteil, die Ergebnisse machen deutlich, daß die *Übungen* die Bewertungen stark beeinflussen oder sogar bestimmen.

Es scheinen sich noch einige interessante Tendenzen abzuzeichnen: Bei genauerer Betrachtung fällt auf, daß die drei letzten Übungen (Kreativitäts-Übungen 1 & 2, Interview) die höchsten konvergenten Validitätskoeffizienten aufweisen. Der Unterschied zwischen diesen drei Übungen (gemitteltes r = 0.48) und den restlichen (r = 0.30) läßt sich sogar statistisch bei zweiseitiger Testung auf dem 5 % Niveau absichern. Erwähnenswert ist auch, daß die Heterotrait-Heteromethod Korrelationen (mittleres r = 0.29) durchgängig kleiner sind als die Heterotrait-Monomethod Korrelationen (mittleres r = 0.79). Dieser Unterschied läßt sich bei zweiseitiger Testung auf dem 1% Niveau statistisch absichern. Die Korrelationen innerhalb der Übungen (Heterotrait-Monomethod) sind hochsignifikant größer als alle anderen Korrelationen.

4.2.2. Hauptkomponentenanalyse

Hier sollen nun die Ergebnisse der explorativen Faktorenanalyse vorgestellt werden.

Wie bereits oben aufgeführt, wurde dabei auf Basis der MTMM-Korrelationen eine Hauptkomponentenanalyse mit quadrierten multiplen Korrelationen als Kommunalitätenschätzung und anschließender Varimax-Rotation durchgeführt. Als Faktoren-Extraktionsbedingung wurde ein Eigenwert von 1 festgelegt. Die folgende Abbildung zeigt den Eigenwertverlauf der 32 Hauptkomponenten aus der Hauptkomponentenanalyse der MTMM-Matrix. Zur Veranschaulichung wurde eine gestrichelte Linie für □ = 1 eingefügt.

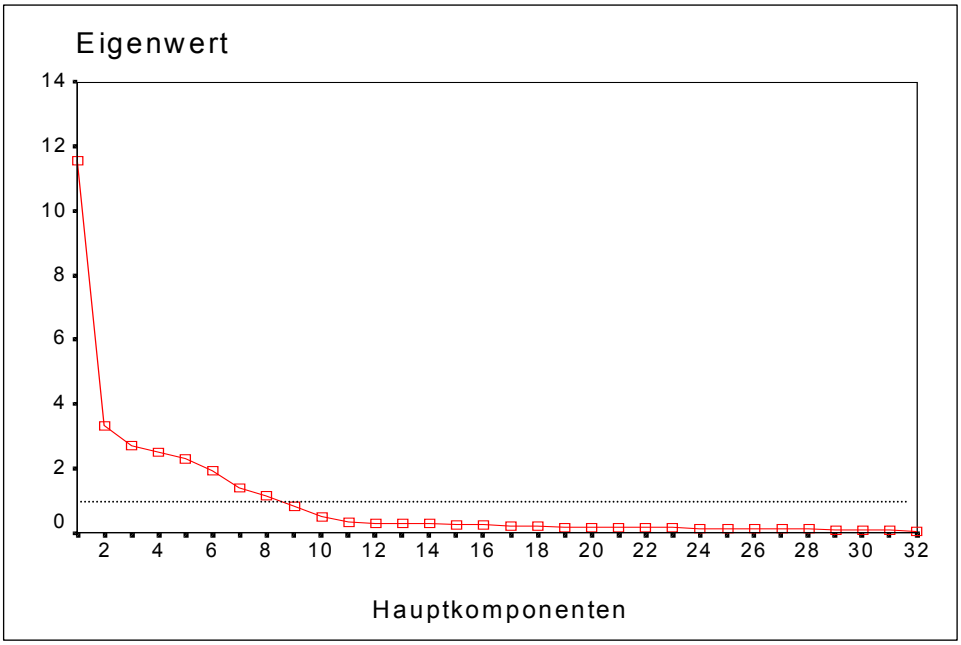


Abbildung 4.1: Eigenwertverlauf der 32 Hauptkomponenten aus der Hauptkomponentenanalyse der MTMM-Matrix

Acht der 32 Hauptkomponenten weisen einen Eigenwert größer als 1 auf und erklären insgesamt 84 % der Varianz der Variablen. Die folgende Übersicht soll die einzelnen Prozentanteile darstellen.

Tabelle 4.6: Erklärter Prozentanteil der Hauptkomponenten an der Gesamtvarianz

	Hauptkomponente							
	1	2	3	4	5	6	7	8
Eigenwert	11.58	3.32	2.72	2.50	2.29	1.95	1.40	1.14
Erklärter Anteil an der Gesamtvarianz	36.2%	10.4%	8.5%	7.8%	7.2%	6.1%	4.4%	3.6%

Es wird deutlich, daß die erste Hauptkomponente bei einem Eigenwert von 11.58 mehr als ein Drittel (36,2%) der Gesamtvarianz erklärt. Die zweite Hauptkomponente trägt bei einem Eigenwert von 3.32 mehr als zehn Prozent (10.4%) zur Varianzaufklärung bei, während schließlich die achte Hauptkomponente bei einem Eigenwert von 1.14 noch 3,6% der Gesamtvarianz erklärt. Die Kommunalitäten der 32 Variablen (hier: Einzeldimensionen), die angeben, in welchem Ausmaß die Variable durch die acht Faktoren aufgeklärt bzw. erfaßt wird, liegen zwischen 64 % (Entscheidungsfähigkeit im Rollenspiel) und 94% (Analysefähigkeit in Präsentations-Übung 1).

Für die Lösung mit acht Hauptkomponenten sind in der folgenden Tabelle die varimax-rotierten Ladungen der Variablen auf den Hauptkomponenten dargestellt. Dabei sind Faktorladungen größer als .30 schattiert abgebildet.

Tabelle 4.7: Ergebnisse der Hauptkomponentenanalyse (varimax-rotierte Komponentenmatrix). Die Variablen sind nach Übungen geordnet.

	Hauptkomponente							
	1	2	3	4	5	6	7	8
Gruppendiskussion „A"								
Entscheidungsfähigkeit	.089	.212	.804	.193	.006	.124	.113	.057
Flexibilität	.128	.132	.848	.080	.080	.046	.046	.098
Kontaktfähigkeit	.124	.130	.836	.060	.047	.133	.145	.159
Integration	.159	.117	.850	.039	.104	.034	.073	.114
Durchsetzung	.178	.158	.811	.119	.030	.100	.139	.059
Gruppendiskussion „B"								
Analysefähigkeit	.127	.843	.158	.164	.167	.080	.069	.022
Flexibilität	.119	.865	.156	.152	.105	.113	.168	.075
Einfühlungsvermögen	.166	.846	.123	.141	.101	.122	.090	.075
Integration	.122	.869	.143	.044	.085	.126	.109	.115
Durchsetzung	.124	.846	.204	.175	.076	.108	.075	.124
Rollenspiel								
Entscheidungsfähigkeit	.706	.189	.127	.238	-.028	.145	.101	.030
Flexibilität	.862	.094	.123	.155	.055	.046	.090	.143
Kontaktfähigkeit	.888	.120	.159	.106	.097	.024	.113	.162
Einfühlungsvermögen	.865	.047	.134	.023	.092	.019	.017	.166
Durchsetzung	.812	.162	.101	.150	.038	.139	.105	.049
Ethische Grundhaltung	.855	.089	.094	-.084	.071	.082	.061	.157
Präsentations - Übung 1								
Organisation	.073	.186	.089	.138	.917	.101	.133	.023
Analysefähigkeit	.067	.127	.058	.164	.928	.103	.110	.071
Entscheidungsfähigkeit	.118	.133	.083	.143	.906	.095	.135	.070
Präsentations – Übung 2								
Organisation	.065	.137	.093	.895	.161	.127	.055	.048
Analysefähigkeit	.110	.165	.081	.874	.158	.091	.063	.045
Entscheidungsfähigkeit	.139	.163	.111	.902	.078	.093	.056	.023
Kreativität	.136	.119	.152	.850	.075	.079	.038	.083
Kreativitäts – Übung 1								
Kreativität	.122	.162	.109	.122	.089	.880	.202	.116
Innovationsfähigkeit	.117	.198	.111	.147	.089	.879	.209	.128
Ausstrahlung	.136	.139	.184	.142	.157	.801	.189	.166
Kreativitäts – Übung 2								
Kreativität	.157	.190	.167	.095	.138	.223	.860	.146
Innovationsfähigkeit	.098	.158	.167	.064	.164	.205	.883	.145
Ausstrahlung	.216	.158	.218	.070	.167	.253	.753	.248
Interview								
Innovationsfähigkeit	.233	.204	.161	.192	.079	.196	.197	.755
Ethische Grundhaltung	.222	.053	.126	-.037	-.014	.101	.156	.862
Ausstrahlung	.220	.138	.196	.093	.131	.135	.127	.838

Anmerkung: Faktorladungen größer als .30 sind markiert.

Die Hauptkomponentenanalyse in Tabelle 4.7 bildet klar Übungsfaktoren ab. Die Faktorladungen der Einzeldimensionen einer Übung laden allesamt auf dem gleichen Faktor. Dabei sind Ladungen von .706 (Entscheidungsfähigkeit im Rollenspiel) bis .928 (Analysefähigkeit in der Präsentations-Übung 1) ermittelt worden. Da die Faktorladungen der Korrelation zwischen einer Variablen und einem Faktor entsprechen, läßt sich ablesen, daß der größte Teil der Varianz der Einzeldimensionen durch den jeweiligen Übungsfaktor erklärt wird. Dabei liegt die Spannweite der erklärten Varianz zwischen 50% (Entscheidungsfähigkeit im Rollenspiel) und 86% (Analysefähigkeit in der Präsentations-Übung 1).

Deutlich wird auch, daß Variablen, die nicht zur extrahierten Hauptkomponente gehören, nur sehr niedrig laden. Die Faktorladungen liegen bis auf die markierten Werte allesamt unter 0.30. Die Hauptkomponentenanalyse ermittelt keine Übungsfaktoren und gibt keine Hinweise darauf, daß den Bewertungen Assessment Center Dimensionen zugrunde liegen.

4.2.3 Konfirmatorische Faktorenanalyse

Die Ergebnisse der konfirmatorischen Faktorenanalyse, basierend auf der MTMM-Matrix, werden im folgenden vorgestellt. Dabei sollen drei Modelle getestet werden (s. Abschnitt 3.4), die aus den Konstruktionsprinzipien von Assessment Centern abgeleitet wurden. Die entsprechenden Rechnungen zur Überprüfung der Modelle wurden mit dem Programm LISREL 8 (Jöreskog & Sörbom, 1993) durchgeführt. Es sei bemerkt, daß bei allen LISREL-Auswertungen die jeweiligen Kovarianzen anstelle der Korrelationen der MTMM-Matrix analysiert wurden, weil lediglich beim Vorliegen einer empirischen Kovarianzmatrix die Teststatistik annähernd χ^2–verteilt ist.

Trotz der fehlenden Voraussetzung multivariat normalverteilter Variablen (s.o.) wurden die unter 3.4 vorgestellten Kennwerte der Anpassungsgüte bestimmt. Dies geschah, um die Ergebnisse mit anderen Untersuchungen zu vergleichen. Die Autoren dieser Studien sind bei Verletzung der Annahmen genauso vorgegangen (vgl. Bycio et al., 1987). Der χ^2 - Wert ist dadurch jedoch nicht – wie intendiert - als hypothesenprüfender Kennwert zu verstehen. Die Irrtumswahrscheinlichkeit p hat aufgrund der Verletzung der Voraussetzungen nur den Charakter einer explorativen Größe. In der folgenden Tabelle sind die Kennwerte für die Anpassungsgüte der drei Modelle wiedergegeben.

Tabelle 4.8: Kennwerte für die Anpassungsgüte verschiedener faktorenanalytischer Modelle

Modell	GFI	AGFI	RMR	χ^2	df	p
Modell 1 Zwölf Dimensionsfaktoren (korreliert)	0.46	0.29	0.17	5552.2	398	0.0
Modell 2 Acht Übungsfaktoren (korreliert)	0.80	0.75	0.05	1144.4	436	0.0
Modell 3 Zwölf Dimensionssfaktoren (korreliert) und acht Übungsfaktoren (korreliert)	0.94	0.86	0.02	365.7	242	0.0

Anmerkung: p: Irrtumswahrscheinlichkeit; GFI: Goodness-of-Fit Index; AGFI: Adjusted Goodness-of-Fit Index; RMR: Root Mean Square Residual.

Die Ergebnisse in Tabelle 4.8 zeigen, daß keines der geprüften Modelle die Daten statistisch abgesichert reproduziert. Die Irrtumswahrscheinlichkeiten liegen allesamt deutlich unter der Grenze (p = 0.0 für alle drei Modelle). Das bedeutet, die Modelle sind mit dem Datensatz nicht hinreichend vereinbar. Es sei hier noch einmal auf den wichtigen Unterschied zum üblichen Vorgehen beim Hypothesentesten hingewiesen: Bei diesem Prüfverfahren ist von einer guten Modellanpassung zu sprechen, wenn die χ^2-Prüfgröße insignifikant ist, das heißt, möglichst groß ist. Wie unter 3.4 bereits ausgeführt, unterstreichen die LISREL-Entwickler Jöreskog und Sörbom (1989), daß die χ^2-Prüfgröße v.a. bei großen Stichproben leicht zur Verwerfung von Modellen führt, obwohl eine zufriedenstellende Passung vorliegt. Demnach liegt die Bedeutung der LISREL Kennwerte auch in der Möglichkeit herauszufinden, welches Modell die Daten am besten widerspiegelt.
Wird die Tabelle 4.8 auf dem Hintergrund dieser Überlegungen betrachtet, zeigt sich, daß die Modelle in ihrer Anpassungsgüte doch stark differieren. Einen Anhaltspunkt gibt der χ^2-Wert. Dabei ist das Verhältnis zwischen Anzahl der Freiheitsgrade und der χ^2-Prüfgröße von Bedeutung. In der Tabelle ist erkennbar, daß der χ^2-Wert für Modell 1 mit 5552.2 bei 398 Freiheitsgraden am größten ist. Das Verhältnis zwischen χ^2-Wert und Freiheitsgraden für Modell 1 liegt somit bei 14 zu 1. Für das zweite Modell wurde ein χ^2-Wert von 1144.4 mit 436 Freiheitsgraden ermittelt; das Verhältnis (χ^2-Wert zu Freiheitsgraden) ist 2.6 zu 1. Beim dritten Modell liegt das Verhältnis bei 1.5 zu 1 bei einem χ^2-Wert von 365.7 und 242 Freiheitsgraden.
Einen weiteren Hinweise, welches Modell die Daten am besten reproduziert, gibt der Goodness-of-Fit Index (GFI). Laut Pfeifer und Schmidt (1987) ist der GFI auch bei Verletzung der Normalverteilungsvoraussetzung sehr robust. Wie in Tabelle 4.8 ersichtlich, variiert der GFI zwischen 0.46 bei Modell 1 und 0.94 bei Modell 3. Der Adjusted Goodnes-of-Fit Index liegt zwischen 0.29 (Modell 1) und 0.86 (Modell 3). Die Werte belegen, daß Modell 1 mit zwölf Dimensionsfaktoren die Daten am schlechtesten repräsentiert. Der Index liegt für das dritte Modell (zwölf Dimensions- und acht Übungsfaktoren) am nächsten am Wert 1, der perfekten Passung. Auch das Modell 2 mit acht Übungsfaktoren gibt die Daten offensichtlich besser wieder als

Modell 1. Die Werte für das Root Mean Square Residual zeigen die gleiche Tendenz. Modell 3 liegt mit RMR = 0.02 dem perfekten Wert von 0 am nächsten. Beim zweiten Modell beträgt das Root Mean Square Residual 0.05 und beim ersten 0.17.

Diese Ergebnisse deuten daraufhin, daß das Modell 3, das sowohl Übungseinflüsse als auch Dimensionseinflüsse annimmt, die Daten am besten erklärt. Die LISREL-Analyse des Modells führt jedoch zu einigen theoretisch nicht möglichen Parametern. Das heißt, daß LISREL Korrelationen produziert, die über dem möglichen Wert von 1 liegen. Die von LISREL ermittelten Werte der Anpassungsgüte (s. Tabelle 4.8) sind daher nicht exakte Werte, sondern nur bedingte Schätzungen (s. Pfeifer & Schmidt, 1987). Dies kann laut Jöreskog und Sörbom (1989) bedeuten, daß die Daten sich nur sehr schlecht oder gar nicht durch das Modell erklären lassen. Die Parameter des dritten Modells sollen daher nicht weiter interpretiert werden (s. Pfeifer & Schmidt, 1987); es werden im weiteren nur LISREL Spezifikationen der Modelle 1 und 2 genauer betrachtet.

Die folgende Tabelle 4.9 gibt die Kovarianzen wieder, die von LISREL mittels der Maximum-Likelihood-Methode (s.o.) geschätzt werden. Die Werte zeigen die gleiche Tendenz wie die Kennwerte der Anpassungsgüte (s. Tabelle 4.8). Die Kovarianzen der manifesten Variablen mit den vermuteten Übungsfaktoren in Modell 2 sind deutlich größer als die Kovarianzen der manifesten Variablen mit den vermuteten Dimensionsfaktoren (Personenmerkmalsfaktoren) in Modell 1. Dabei liegen die Werte für Modell 2 zwischen .67 (Flexibilität in Gruppendiskussion „A" und Ethische Grundhaltung im Interview) und 1.17 (Organisation und Analysefähigkeit in Präsentations-Übung 1). Für das erste Modell schätzt LISREL Kovarianzen im Bereich von .29 (Kreativität in Präsentations-Übung 2) bis .90 (Kontaktfähigkeit im Rollenspiel). Zu erwähnen ist auch, daß die Kovarianzen für jede manifeste (x-) Variable beim zweiten Modell größer ist. Auch die Meßfehler unterstreichen die Einschätzung, daß das zweite Modell die Daten deutlich besser repräsentiert. Während die Werte der Kovarianzen zwischen Meßfehler und manifester Variable für Modell 1 zwischen .29 (Kontaktfähigkeit im Rollenspiel) und 1.02 (Kreativität in Präsentations-Übung 2) liegen, sind die Kovarianzen im zweiten Modell deutlich niedriger. Die geschätzten Kovarianzen variieren hier zwischen .06 (Innovations-fähigkeit in Kreativitäts-Übung 1) und .51 (Entscheidungsfähigkeit im Rollenspiel).
Auffällig ist, daß die Meßfehler im ersten Modell meistens höher mit den manifesten Variablen kovariieren als mit den eigentlich vermuteten Dimensionsfaktoren. Das bedeutet, daß die Daten sich eher durch die Meßfehler erklären lassen als durch die Dimensionen. Dieser Umstand zeigt deutlich, daß das Modell 1 mit zwölf Übungsfaktoren die Beobachter-Bewertungen nicht adäquat repräsentiert.

Tabelle 4.9: Von LISREL geschätzte Kovarianzen (Maximum-Likelihood-Methode)

x- Variable	Modell 1 (zwölf Dimensionsfaktoren)		Modell 2 (acht Übungsfaktoren)	
	Dimensionen	Meßfehler	Übungen	Meßfehler
Gruppendiskussion „A"				
Entscheidungsfähigkeit	.34	.71	.76	.25
Flexibilität	.30	.58	.67	.21
Kontaktfähigkeit	.32	.74	.80	.21
Integration	.33	.65	.71	.25
Durchsetzung	.58	.77	.89	.31
Gruppendiskussion „B"				
Analysefähigkeit	.40	.85	.87	.25
Flexibilität	.60	.58	.90	.13
Einfühlungsvermögen	.37	.73	.84	.16
Integration	.70	.48	.88	.19
Durchsetzung	.63	.68	.91	.25
Rollenspiel				
Entscheidungsfähigkeit	.38	.82	.68	.51
Flexibilität	.57	.77	.93	.23
Kontaktfähigkeit	.90	.29	1.01	.10
Einfühlungsvermögen	.69	.95	1.07	.29
Durchsetzung	.66	.68	.83	.43
Ethische Grundhaltung	.86	.32	.89	.27
Präsentations - Übung 1				
Organisation	.77	.93	1.17	.14
Analysefähigkeit	.84	.79	1.17	.11
Entscheidungsfähigkeit	.70	.68	1.00	.18
Präsentations – Übung 2				
Organisation	.48	.65	.87	.12
Analysefähigkeit	.60	.75	.95	.20
Entscheidungsfähigkeit	.48	.67	.87	.13
Kreativität	.29	1.02	.89	.32
Kreativitäts – Übung 1				
Kreativität	.78	.66	1.05	.16
Innovationsfähigkeit	.75	.50	1.01	.06
Ausstrahlung	.73	.48	.84	.31
Kreativitäts – Übung 2				
Kreativität	.89	.60	1.13	.10
Innovationsfähigkeit	.86	.54	1.08	.12
Ausstrahlung	.83	.38	.88	.29
Interview				
Innovationsfähigkeit	.46	.77	.85	.26
Ethische Grundhaltung	.34	.54	.67	.21
Ausstrahlung	.56	.68	.91	.16

In der folgenden Tabelle 4.10 wird der Zusammenhang zwischen manifesten Variablen und den vermuteten Faktoren weiter untersucht. Die Varianzbeiträge der Dimensionsfaktoren (Modell 1) und der Übungsfaktoren (Modell 2) zu den 32 manifesten Variablen sollen dabei als Prozentanteile dargestellt werden.

Dies geschieht, um einschätzen zu können, wie gut die vermuteten Faktoren die manifesten Variablen erklären. Dazu wurden die quadrierten Korrelationskoeffizienten ermittelt (vgl. Bühl & Zöfel, 1996).

Tabelle 4.10: Varianzbeiträge (LISREL – Schätzungen) der Dimensionsfaktoren (Modell 1) und der Übungsfaktoren (Modell 2) zu den 32 manifesten Variablen (x-Variablen)

x-Variable	erklärte Varianz durch Dimensionsfaktoren in Modell 1	erklärte Varianz durch Übungsfaktoren in Modell 2
Gruppendiskussion „A"		
Entscheidungsfähigkeit	14 %	70 %
Flexibilität	14 %	68 %
Kontaktfähigkeit	12 %	76 %
Integration	15 %	67 %
Durchsetzung	31 %	72 %
Gruppendiskussion „B"		
Analysefähigkeit	16 %	75 %
Flexibilität	39 %	86 %
Einfühlungsvermögen	16 %	81 %
Integration	51 %	80 %
Durchsetzung	37 %	77 %
Rollenspiel		
Entscheidungsfähigkeit	15 %	47 %
Flexibilität	30 %	79 %
Kontaktfähigkeit	74 %	91 %
Einfühlungsvermögen	34 %	80 %
Durchsetzung	39 %	62 %
Ethische Grundhaltung	70 %	75 %
Präsentations - Übung 1		
Organisation	39 %	91 %
Analysefähigkeit	47 %	93 %
Entscheidungsfähigkeit	42 %	85 %
Präsentations – Übung 2		
Organisation	27 %	86 %
Analysefähigkeit	32 %	82 %
Entscheidungsfähigkeit	25 %	85 %
Kreativität	7 %	71 %
Kreativitäts – Übung 1		
Kreativität	48 %	87 %
Innovationsfähigkeit	53 %	95 %
Ausstrahlung	53 %	70 %
Kreativitäts – Übung 2		
Kreativität	57 %	92 %
Innovationsfähigkeit	58 %	91 %
Ausstrahlung	64 %	73 %
Interview		
Innovationsfähigkeit	22 %	74 %
Ethische Grundhaltung	18 %	68 %
Ausstrahlung	32 %	84 %

Die in Tabelle 4.10 dargestellten Varianzbeiträge unterscheiden sich bei den zwei betrachteten Modellen deutlich. Die Dimensionsfaktoren erklären im ersten Modell zwischen 7 % (Kreativität in Präsentations-Übung 2) und 74 % (Kontaktfähigkeit im Rollenspiel) der manifesten Variablen. Trotz des relativ hohen Werts von 74 % weisen die meisten quadrierten Korrelationskoeffizienten niedrige Varianzbeiträge (die meisten Werte liegen deutlich unter 50 %) auf. Auch läßt sich keine „dimensionsspezifische" Tendenz erkennen. Das heißt, es scheint keinen Personenmerkmalsfaktor zu geben, der besonders gut die zugehörigen manifesten Variablen erklärt. Oder anders ausgedrückt, es scheint kein Personenmerkmal zu geben, das die Beobachter adäquat erkennen und bewerten. Beispielsweise ist der

Varianzbeitrag von 74 % für Kontaktfähigkeit im Rollenspiel nicht als Hinweis zu werten, daß das Personenmerkmal Kontaktfähigkeit gut von den Beobachtern erkannt wurde, da die erklärte Varianz für Kontaktfähigkeit in der Gruppendiskussion „A" nur 14 % beträgt.

Die erklärten Varianzen im zweiten Modell sind höher und liegen zwischen 47 % (Entscheidungsfähigkeit im Rollenspiel) und 95 % (Innovationsfähigkeit in der Kreativitäts-Übung 1). Auffällig ist, daß die Werte des zweiten Modells für jede manifeste Variable größer sind als die des ersten Modells. Die 32 manifesten Variablen werden also durch die Übungsfaktoren viel besser erklärt als durch Dimensionsfaktoren. Insgesamt zeigt sich aber auch, daß die erklärten Varianzen in Abhängigkeit von der Übung und der Dimension zu betrachten sind. In beiden Modellen streuen die Werte sehr stark sowohl innerhalb der Übungen als auch zwischen den Übungen.

4.3 Hypothesenprüfung

Im folgenden sollen - analog zur Gliederung der Ergebnisse - zuerst die zweite Hypothese zur Interrater-Reliabilität und dann die Hypothesen 1 und 1a zur Konstruktvalidität überprüft werden.

4.3.1 Hypothesen zur Interrater-Reliabilität

Ausgehend von der Forschung zur Interrater-Reliabilität im Assessment Center (s. Abschnitt 2.4.1) wurde angenommen, daß Beobachter, die ausreichend trainiert wurden, Kandidaten mit genügend hoher Übereinstimmung bewerteten (Hypothese 2). Dabei wurden nicht so hohe Werte erwartet wie in den Studien, bei denen den Bewertungen ein Informationsaustausch vorherging (vgl. Schmitt, 1977 und Jones, 1981). Anhand der Beschreibung des untersuchten Assessment Centers kann angenommen werden, daß die Beobachter ausreichend trainiert wurden, so daß diese Voraussetzung als gegeben gewertet wird. Die oben dargestellten Ergebnisse zur Interrater-Reliabilität machen deutlich, daß die Beobachter mit guter Übereinstimmung die Kandidaten des untersuchten Assessment Centers bewertet haben. Alle Korrelationen sind höchstsignifikant und zumeist deutlich über $r = 0.50$, so daß die erste Hypothese, wonach Assessment Center Beobachter nach ausreichenden Training mit genügend hoher Übereinstimmung bewerten, als bestätigt gelten kann.

4.3.2 Hypothesenprüfung zur Konstruktvalidität

Abgeleitet aus den Konstruktionsprinzipien des untersuchten Assessment Centers, wonach Personenmerkmale in berufsrelevanten Übungen bewertet werden, wurde folgendes zur Konstruktvalidität angenommen:
Beobachter, die ausreichend trainiert wurden, können im Assessment Center Personenmerkmale in verschiedenen Übungen bewerten (Hypothese 1).

Die Ergebnisse der Analyse der MTMM-Matrix bestätigen scheinbar diese Hypothese. Beobachter können demnach Personenmerkmale konvergent valide (s.o.) bewerten. Dieser Zusammenhang konnte auch statistisch abgesichert werden. Die weitere Analyse der MTMM-Matrix läßt jedoch vermuten, daß dieses Ergebnis nicht hypothesenkonform zu interpretieren ist, da fast alle Korrelationen der MTMM-Matrix signifikant von Null verschieden sind. Die konvergente Validität könnte sich nur auf Grund der Tatsache ergeben, daß *alle* Bewertungen stark miteinander korrelieren. Die gefundene Konvergenzen würden dann nicht auf übereinstimmenden Bewertungen von Personenmerkmalen basieren. Die Ergebnisse der konfirmatorischen Faktorenanalyse unterstützen diese Einschätzung. Es sind keine statistisch abgesicherten Ergebnisse dazu ermittelt worden, daß Beobachter wirklich Personenmerkmale in verschiedenen Übungen bewerten. Zusammenfassend läßt sich sagen, daß die vorliegenden Ergebnisse nicht eindeutig zu interpretieren sind. Die erste Hypothese konnte nicht abschließend bestätigt werden.

Die erste Hypothese wurde in bezug auf eine faktorenanalytische Auswertung weiter spezifiziert. Für ein Assessment Center mit ausreichend trainierten Beobachtern wurde angenommen, daß ein signifikanter Teil der Verhaltensvarianz durch Dimensionsfaktoren erklärt wird (Hypothese 1a). Die Hauptkomponentenanalyse auf Basis der MTMM-Matrix zeigt deutlich, daß Übungsfaktoren die Bewertungen bedingen. Die Ergebnisse geben keinen Hinweis darauf, daß auch Dimensionsfaktoren die Bewertungen der Beobachter erklären. Entgegen der Erwartung werden keine Dimensionsfaktoren extrahiert. Die Ergebnisse der konfirmatorischen Faktorenanalyse unterstreichen ebenfalls, daß Dimensionsfaktoren keinen signifikanten Einfluß auf die Bewertungen haben. Insgesamt wird deutlich, daß Übungsfaktoren die Daten deutlich besser erklären als Dimensionsfaktoren. Somit konnte die Hypothese 1a nicht bestätigt werden.

5 Diskussion

Wie sind die Ergebnisse zur Interrater-Reliabilität in die bisherige Forschung einzuordnen?

Die Ergebnisse zur Interrater-Reliabilität liegen im erwarteten Bereich. Das bedeutet, daß die hier ermittelten Korrelationskoeffizienten (Spannweite von r = 0.33 bis r = 0.76) vergleichbar mit den Ergebnissen anderen Untersuchungen sind (vgl. Borman, 1982; Jones, 1981; Lammers, 1992; Scholz, 1994). Die Höhe der Korrelationen weist auf eine befriedigende, wenngleich z.T. niedrige Beobachter-Übereinstimmung hin. Somit wird durch die vorliegenden Ergebnisse die Einschätzung von Scholz (1994) sowie Thornton und Byham (1982) unterstützt, daß Assessment Center reliabel sind. Ungeklärt ist jedoch die große Spannweite der Korrelationen, die auch in den vergleichbaren Studien gefunden wurde. Die Unterschiede zwischen den verschiedenen Koeffizienten lassen weitere Vermutungen bzw. Interpretationen zu.

In Übungen, in denen nur drei oder vier Einzeldimensionen zu beobachten waren, bewerteten die Beobachter zuverlässiger als in Übungen, in denen fünf oder sechs Einzeldimensionen beurteilt werden mußten. Ein möglicher Grund hierfür könnte darin liegen, daß eine höhere Zahl an zu beobachtenden Dimensionen die Beobachter überfordert. Dieser Zusammenhang konnte bereits in der Studie von Gaugler und Thornton (1989) ermittelt werden.

Die unterschiedlich hohe Interrater-Reliabilität in den verschiedenen Übungen läßt auch die Annahme zu, daß die *Übungstypen* die Beobachter-Übereinstimmung bedingen. Das könnte bedeuten, daß in Einzelübungen, wie der Präsentations-Übung oder dem Interview, die Beobachter eher in der Lage sind, reliabel zu bewerten, während in Interaktionsübungen (Rollenspiel und Gruppendiskussionen) die Beobachter weniger übereinstimmend beurteilen. Wie in Tabelle 4.3 ersichtlich, unterstützen die vorliegenden Befunde diese Vermutung. Das Rollenspiel und die beiden Gruppendiskussionen weisen die niedrigsten Interrater-Reliabilitätskoeffizienten auf. Ähnliche Ergebnisse wurden auch in anderen Studien ermittelt (u.a. Scholz, 1994). Die Bewertungen einer Gruppendiskussion differieren laut Scholz (1994) stärker, weil der Ablauf der Übung weniger kontrollier- und standardisierbar ist.
Zusammenfassend läßt sich sagen, daß die Beziehung zwischen Übungstyp und Interrater-Reliabilität nicht geklärt ist; weitere Forschung hierzu wäre wünschenswert.

Wie sind die Ergebnisse zur Konstruktvalidität in die bisherige Forschung einzuordnen?

Die hier ermittelten Ergebnisse zur MTMM-Matrix stimmen mit Befunden anderer Studien (u.a. Russell, 1987; Bycio et al., 1987) überein. Assessment Center können offensichtlich den von Campbell und Fiske (1959) postulierten Kriterien der Konstruktvalidität nicht standhalten. Nur das erste Kriterium wird in dieser Untersuchung, wie schon bei Kleinmann (1997), tendenziell bestätigt. Die Analyse der MTMM-Matrix deutet scheinbar hypothesenkonform darauf, daß die Beobachter

Personenmerkmale in verschiedenen Übungen bewerten. Der Zusammenhang konnte auch statistisch abgesichert werden. Es ist jedoch nicht klar, ob das Ergebnis als Bestätigung der Hypothese zu verstehen ist, da fast alle Korrelationen der MTMM-Matrix signifikant von Null verschieden sind. Die konvergente Validität könnte sich nur auf Grund der Tatsache ergeben, daß alle Bewertungen stark miteinander korrelieren (zur genaueren Darstellung dieses Zusammenhangs siehe oben). Darüber hinaus wurden weitere interessante Ergebnisse ermittelt, auf die im folgenden eingegangen wird.

In den letzten drei Übungen wurden die signifkant höchsten konvergenten Validitätskoeffizienten bestimmt. Warum? Eine mögliche Einflußgröße können die Kreativitäts-Übungen sein. Da in beiden Aufgaben die gleichen Einzeldimensionen abgefragt werden, würde die Ähnlichkeit der beiden Kreativitäts-Übungen die hohen konvergenten Koeffizienten bedingen.
Eine weitere Erklärung für das Zustandekommen der Ergebnisse könnte darin liegen, daß die Übungen alle am Ende des Assessment Centers durchgeführt werden. Die Beobachter hätten demnach während der 2-tägigen Veranstaltung „gelernt", konvergentere Bewertungen zu geben (vgl. Fennekels, 1987). Die weiteren Ergebnisse der MTMM-Matrix unterstützen diese Vermutung jedoch nicht. Es ergeben sich nicht stetig steigende konvergente Validitätskoeffizienten je nach Zeitpunkt der Übung, wie nach dieser Erklärung zu vermuten wäre. Außerdem zeigt sich, daß die Heterotrait-Monomethod Korrelationen der beiden Präsentations- und Kreativitäts-Übungen und dem Interview höher sind als bei den Gruppendiskussionen und dem Rollenspiel. Laut Scholz (1994) kann sich ein solches Ergebnis ergeben, weil bei Interaktionsübungen ein größeres Verhaltensspektrum beobachtet werden kann und somit auch eher verschiedene Personenmerkmale beobachtbar sind.

Insgesamt macht die Analyse der Korrelationen MTMM-Matrix deutlich: Es werden nicht die intendierten Personenmerkmale gemessen. Vielmehr zeigen die Ergebnisse den großen Einfluß der Übungen auf die Bewertungen. Die Korrelationen innerhalb der Übungen (Heterotrait-Monomethod) sind hochsignifikant größer als alle anderen Korrelationen. Das heißt, Beobachter beurteilen die Kandidaten nicht nach bestimmten Personenmerkmalen, wie Kreativität, Entscheidungsfähigkeit usw., sondern nach ihrer Performance in einer Übung.

Auch die Hauptkomponentenanalyse repliziert die Befunde anderer Studien (u.a. Hoenle, 1995; Sackett & Dreher, 1982; Neubauer, 1989). Demnach werden mittels der Hauptkomponentenanalysen Übungsfaktoren und nicht Dimensionsfaktoren extrahiert. Die Ergebnisse lassen sich als klarer Beleg dafür interpretieren, daß die Beobachter nach Übungen bewerten. Laut Bortz (1993) ist die genaue Interpretation der Hauptkomponentenanalyse jedoch schwierig; die Ergebnisse sind als „hypothesengenerierend" zu verstehen und erlauben keine direkte Überprüfung von Annahmen (s.a. 3.4). Demnach müßte in weiteren Untersuchungen geklärt werden, ob die Übungen wirklich die Bewertungen bestimmen. Die folgende Interpretation soll daher als Überlegung verstanden werden, die es zu überprüfen gilt.

Die Hauptkomponenten, die den größten Anteil der Gesamtvarianz erklären, sind die Interaktions-Übungen (s. Tabelle 4.6). Insgesamt erklären diese drei Aufgaben über die Hälfte der Varianz (55,1%). Das Interview (extrahierte Hauptkomponente 8) trägt hingegen nur 3,6 % bei. Dieser Befund läßt sich so interpretieren, daß Gruppendiskussionen und Rollenspiele ein größeres Verhaltensspektrum abdecken und somit „mehr" Verhalten von Kandidaten beobachtbar machen (vgl. Scholz, 1994). Innerhalb dieser Übungen können die Beobachter demnach die Kandidaten besser und umfassender einschätzen. Das Ziel von Assessment Centern, Kandidaten zu beurteilen, wäre somit eher durch den Einsatz von Interaktions-Übungen als durch andere Übungstypen zu erreichen. Die Überprüfung dieser Zusammenhänge durch weitere Forschungsstudien steht noch aus.

Die konfirmatorische Faktorenanalyse konnte den vermuteten Einfluß von Personenmerkmalen (Dimensionen) auf die Bewertung der Kandidaten nicht ermitteln. Anders als in den Laborstudien von Kleinmann (1997) und Kudisch et al. (1997) wurden hier keine Dimensionsfaktoren mit Hilfe von LISREL generiert. Im Gegenteil, die vorliegenden Ergebnisse zeigen deutlich, daß Übungsfaktoren die Daten besser repräsentieren. Die Arbeit unterstützt somit die Befunde von Bycio et al. (1987) und Fennekels (1987), die Assessment Center aus der Praxis mit Hilfe von LISREL untersuchten. Sie fanden heraus, daß die Übungen und nicht die erhofften Personenmerkmale die Bewertungen bedingen. Dem Einwand von Kleinmann (1997), daß diese Ergebnisse durch die hohe Zahl von Dimensionen beeinflußt wurden, wird durch die vorliegende Arbeit widersprochen. Auch in Übungen, wo nur drei Dimensionen zu bewerten waren, ergaben sich keine Personenmerkmals - Faktoren.

Die Bedeutung der LISREL Kennwerte liegt in der Möglichkeit, herauszufinden, welches Modell die Daten am besten widerspiegelt. Laut Fennekels (1987) kann man noch von einer guten Anpassung sprechen, wenn das Verhältnis zwischen □□-Prüfgröße und Freiheitsgraden 3 zu 1 beträgt. Betrachtet man die Ergebnisse der konfirmatorischen Faktorenanalyse vor diesem Hintergrund, zeigt sich, daß das Modell 2 mit Übungsfaktoren die Daten gut reproduziert. Bei einem □□-Wert von 1144.4 mit 436 Freiheitsgraden beträgt das Verhältnis 2.6 zu 1. Für Modell 1 ergab sich hingegen bei einem □□-Wert von 5552.2 mit 398 Freiheitsgraden nur ein Verhältnis von 14 zu 1. Die LISREL Statistiken des dritten Modells können - wie unter Abschnitt 4.2.3 ausgeführt - nicht interpretiert werden. Hier sollen daher nur die ersten beiden Modelle diskutiert werden. Die höhere Anpassungsgüte des Übungsfaktor – Modells zeigt sich auch bei den Werten des Goodness-of-Fit Index (GFI). Während für das zweite Modell ein GFI von 0.80 ermittelt wurde, liegt der Kennwert für das erste Modell bei 0.46. Nach der konfirmatorischen Faktorenanalyse

lassen sich somit die Assessment Center Daten als Übungsbewertungen interpretieren. Dabei ist jedoch zu berücksichtigen, daß die Übungen Bewertungen nicht ausschließlich erklären.

Insgesamt zeigen die Ergebnisse der Analyse der MTMM-Matrix, der Hauptkomponentenanalyse und der konfirmatorischen Faktorenanalyse sehr deutlich, daß die Beobachter Übungen und nicht Personenmerkmale bewerten. Anders als in der Studie von Kleinmann (1997) produzierten die Analysemethoden keine unterschiedlichen Ergebnisse. In keinem der drei Verfahren konnte der Einfluß von Dimensionen auf die Bewertungen zweifelsfrei bestätigt werden. Die ursprüngliche Fragestellung, was in Assessment Centern gemessen wird, Übungen oder Dimensionen, kann somit als beantwortet gelten. Die Analyse der Konstruktvalidität des untersuchten Assessment Centers erbrachte letztlich die gleichen Ergebnisse wie die Studien von Sackett und Dreher (1982), Turnage und Muchinsky (1982), Neubauer (1989) sowie Bycio et al. (1987) und Fennekels (1987), wonach das Verfahren nicht konstruktvalide im testtheoretischen Sinne ist. Die Hypothese von Neubauer (1989), „So ergeben z.B. Faktorenanalysen über die Einzelurteile STETS [Hervorhebung im Original; Anm. d. Verf.] Übungsfaktoren und keine Merkmalsfaktoren" (S.203) scheint durch diese Ergebnisse bestätigt zu sein. Das untersuchte Assessment Center mißt demnach nicht die vermuteten Personenmerkmale in berufsrelevanten Übungen, sondern, ob ein Kandidat eine Übung gut oder schlecht absolviert hat.

Wie lassen sich diese Befunde erklären?

Ein Grund für die geringe Konstruktvalidität von Assessment Centern kann nach Lammers (1992) mangelnde Interrater-Reliabilität sein (s. Abschnitt 2.4). Bei Anwendung einer Beobachter-Rotation würden dann die Korrelationen innerhalb einer Übung (Heterotrait-Monomethod) die Bewertungen *eines* Beobachter vergleichen, während die konvergenten Validitätskoeffizienten (Monotrait-Heteromethod) durch Bewertungen *verschiedener* Beobachter zustande kämen. Eine geringe Beobachter-Übereinstimmung könnte somit die niedrige konvergente Validität bedingen. Denkbar ist, daß dieser Effekt die vorliegenden Ergebnisse beeinflußt hat. Die Befunde zur Interrater-Reliabilität deuten jedoch auf eine befriedigende Höhe hin. Der von Lammers (1992) postulierte Zusammenhang scheint somit hier nicht entscheidend gewesen zu sein.

Darüber hinaus ist ein Einfluß der besonderen Ausrichtung des untersuchten Assessment Centers möglich. Das Verfahren wurde als Auswahlinstrument für Paare konzipiert (s.o.). Halo-Effekte könnten somit die Beurteilungen verzerrt haben. Es gilt außerdem zu überlegen, ob die verschiedenen Übungstypen die Bewertungen bestimmen. Die Ergebnisse der drei Analyseverfahren zeigen, daß die Übungen z.T. erheblich differieren. Die Beobachter bewerten die Kandidaten - wie in der Untersuchung von Templer (1995) - in den verschiedenen Übungen unterschiedlich. Inwieweit die beschriebenen Effekte die Ergebnisse der Interrater-Reliabilität und Konstruktvalidität beeinflußt haben, läßt sich jedoch hier nicht feststellen.

Fazit

Die vorliegenden Befunde zeigen, daß der Aufgabentyp einen entscheidenden Einfluß auf die Bewertungen der Beobachter hat (vgl. Kleinmann, 1997). Ob Interaktions-Übungen einen größeren und vielleicht wichtigeren Anteil zur Qualität des Assessment Center Verfahrens beitragen als andere Übungstypen, wie z.B. Einzelübungen, gilt es in der weiteren Forschung zu überprüfen. Hier sei die Empfehlung von Templer (1995) aufgegriffen, im Assessment Center möglichst viele verschiedene Aufgabentypen einzusetzen, wobei eine besondere Gewichtung auf Interaktions-Übungen gelegt werden sollte.

Bei Betrachtung der Ergebnisse der Assessment Center Forschung fällt auf, daß die Studien, in denen ein Einfluß der Dimensionen auf die Bewertungen gefunden wurde, Laborstudien waren (Guldin & Schuler, 1997; Kleinmann, 1997; Kudisch et al., 1997). Dieser Umstand läßt die Interpretation zu, daß die standardisierten Bedingungen einer Laboruntersuchung nicht in Assessment Centern der Praxis umgesetzt werden können. Die gefundenen Ergebnisse könnten also durch Einflußgrößen bedingt sein, die mit der praktischen Umsetzung von Assessment Centern zusammenhängen. Sowohl die Laborstudien als auch die Analysen von Assessment Centern der Praxis ergaben jedoch, daß Übungen die Bewertungen entscheidend bestimmen. So muß letztlich vielleicht doch der Einschätzung von Neubauer (1989) zugestimmt werden:

„Die Übungen entscheiden über die Gesamturteile; Merkmale sind innerhalb der Übungen hilfreiche Beobachtungshinweise" (S.204).

6 Literatur

Arbeitskreis Assessment Center (Hrsg.) (1989). *Das Assessment Center in der betrieblichen Praxis. Erfahrungen und Perspektiven*. Hamburg: Windmühle.

Arbeitskreis Assessment Center (Hrsg.) (1995). *Assessment Center auf dem Prüfstand. Bewährungskontrolle und Qualitätssicherung am Beispiel eines Unternehmens-ACs.* Hamburg: Windmühle.

Arbeitskreis Assessment Center (Hrsg.) (1996). *Assessment Center als Instrument der Personalentwicklung. Schlüsselkompetenzen, Qualitätsstandards, Prozeßoptimierung.* Hamburg: Windmühle.

Backhaus, K., Erichson, B., Plinke, W. & Weiber, R. (1996). *Multivariate Analysemethoden. Eine anwendungsorientierte Einführung*. (8. Aufl.). Berlin: Springer-Verlag.

Borman, W. (1982). Validity of Behavioral Assessment for Predicting Military Recruiter Performance. *Journal of Applied Psychology, 67*, 3-9.

Bortz, J. (1993). *Statistik für Sozialwissenschaftler.* (4. vollst. überarbeitete Aufl.). Berlin: Springer-Verlag.

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

Bortz, J., Lienert, G.A. & Boehnke, K. (1990). *Verteilungsfreie Methoden in der Biostatistik.* Berlin: Springer-Verlag.

Bray, D.W. (1964). The Management Progress Study. *American Psychologist*, *19*, 419-420.

Bray, D.W. & Grant, D.L. (1966). The assessment center in the measurements of potential for business management. *Psychological Monographs*, *80* (17), 1-27.

Bühl, A. & Zöfel, P. (1996). *Professionelle Datenanalyse mit SPSS für Windows.* Bonn: Addison-Wesley.

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

Byham, W.C. (1970). Assessment centers for spotting future managers. *Harvard Business Review, 48*, 150-167.

Byham, W.C. (1977). Application of the assessment center method. In J.L. Moses & W.C. Byham (Eds.), *Applying the Assessment Center Method* (pp. 31-43). New York: Pergamon Press.

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

Domsch, M. & Jochum, I. (1989). Zur Geschichte des Assessment Centers – Ursprünge und Werdegänge. In C. Lattmann (Hrsg.), *Das Assessment-Center-Verfahren der Eignungsbeurteilung. Sein Aufbau, seine Anwendung und sein Aussagegehalt* (S. 1-18). Heidelberg: Physica-Verlag.

Fennekels, G. (1987). *Validität des Assessment-Centers bei Führungskräfteauswahl und –entwicklung.* Unveröffentlichte Dissertation, Rheinische Friedrich-Wilhelms-Universität Bonn.

Finkle, R.B (1976). Managerial asessment centers. In M.D. Dunette (Ed.), *Handbook of industrial and organizational psychology* (pp. 861-888). Chicago: Rand McNally.

Fisseni, H.-J. & Fennekels, G. (1995). *Das Assessment Center. Eine Einführung für Praktiker.* Göttingen: Verlag für Angewandte Psychologie.

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

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

Guldin, A. und Schuler, H. (1997). Konsistenz und Spezifität von AC-Beurteilungskriterien: Ein neuer Ansatz zur Konstruktvalidierung des Assessment Center-Verfahrens. *Diagnostica 43*, 230-254.

Harburger, W. (1992). Soziale Validität im individuellen Erleben von Assessment-Center- Probanden. *Zeitschrift für Arbeits- und Organisationspsychologie*, *36*, 147-151.

Hinrichs, J.R. & Haanperä, S. (1976). Reliability of measurement in situational exercises: an assessment of the assessment center method. *Personnel Psychology*, *29*, 31-40.

Hoenle, S. (1995). Ergebnisse zur Konstruktvalidität. In Arbeitskreis Assessment Center (Hrsg.), *Assessment Center auf dem Prüfstand. Bewährungskontrolle und Qualitätsicherung am Beispiel eines Unternehmens-ACs* (S.39-52). Hamburg: Windmühle

Holling, H. & Leippold, W. (1991). Zur sozialen Validität von Assessment Centern. In H. Schuler & U. Funke (Hrsg.), *Eignungsdiagnostik in Forschung und Praxis* (S. 305-312). Stuttgart: Verlag für Angewandte Psychologie.

Horn, U. (1996). *Integrative Entwicklungsbegleitung statt Assessment Center: Ein neuer Ansatz zur Auswahl von Führungsnachwuchskräften.* Hamburg: S + W Steuer- und Wirtschaftsverlag.

Huck, J.R. (1977) The research base. In J.L. Moses & W.C. Byham (Eds.), *Applying the Assessment Center Method* (pp. 261-291). New York: Pergamon Press.

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

Jeserich, W. (1989). Überblick und Einleitung. In Arbeitskreis Assessment Center (Hrsg.), *Das Assessment Center in der betrieblichen Praxis. Erfahrungen und Perspektiven* (S. 7-11). Hamburg: Windmühle.

Jeserich, W. (1995). Assessment Center (AC). In W. Sarges (Hrsg.), *Management-Diagnostik* (S. 717-728). (2. überarbeitete Auflage). Göttingen: Hogrefe.

Jochmann, W. (1999). Vorwort. In W. Jochmann (Hrsg.), *Innovationen im Assessment-Center* (S. V-VII). Stuttgart: Schäffler-Poeschel Verlag.

Jöreskog, K.G. & Sörbom, D. (1989*). LISREL 7 - A guide to the program and applications.* (2nd Ed.). Chicago: SPSS Publications.

Jöreskog, K.G. & Sörbom, D. (1993). *LISREL 8 – Structural equation modeling with the SIMPLIS Command Language*. Chicago: Scientific Software.

Jones, A. (1981). Inter-rater reliability in the assessment of group exercises at a UK assessment centre. *Journal of Occupational Psychology*, *54*, 79-86.

Jung, P. & Leiter, R. (1989). Definition und Zielsetzung der Assessment Center-Methode. In Arbeitskreis Assessment Center (Hrsg.), *Das Assessment Center in der betrieblichen Praxis. Erfahrungen und Perspektiven* (S. 30-32). Hamburg: Windmühle.

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. (1997). *Assessment Center. Stand der Forschung – Konsequenzen für die Praxis.* Göttingen: Verlag für Angewandte Psychologie.

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

Klimoski, R.J. & Strickland, W.J. (1977). Assessment centers – valid or merely prescient. *Personnel Psychology*, *30*, 353-363.

Kompa, A. (1989). *Assessment Center: Bestandsaufnahme und Kritik.* München: Rainer Hampp Verlag.

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. [CD-ROM]. *Journal of Social Behavior & Personality*, *12,* 129-144. Abstract from: EPClient: PsycLIT Item 10074-006.

Kuptsch, C. (1994). *Der Chamäleon-Effekt: über den Einfluß interindividueller Verhaltensvariabilität auf die konvergente Validität eines Personalauswahl- und Personalentwicklungs- Verfahrens*. Unveröffentlichte Dissertation Universität Kiel.

Lammers, F. (1992). *Zur Problematik des Beobachterverhaltens im Assessment-Center*. Unveröffentlichte Dissertation, Universität Osnabrück.

Lattmann, C. (Hrsg.) (1989). *Das Assessment-Center-Verfahren der Eignungsbeurteilung. Sein Aufbau, seine Anwendung und sein Aussagegehalt.* Heidelberg: Physica-Verlag.

Leiter, R. (1996). Vorwort. In Arbeitskreis Assessment Center (Hrsg.), *Assessment Center als Instrument der Personalentwicklung. Schlüsselkompetenzen, Qualitätsstandards, Prozeßoptimierung* (S. 9-10). Hamburg: Windmühle.

Lienert, G.A. (1969). *Testaufbau und Testanalyse* (3. Auflage). Weinheim: Beltz-Verlag.

Lorenzo, R.L. (1984). Effekts of assessorship on managers´ proficiency in acquiring, evaluating, and communicating information about people. *Personnel Psychology*, *37*, 617-634.

MacKinnon, D.W. (1977). From selecting spies to selecting managers – the OSS assessment program. In J.L. Moses & W.C. Byham (Eds.), *Applying the Assessment Center Method* (pp. 13-30). New York: Pergamon Press.

Maukisch, H. (1986). Erfolgskontrollen von Assessment Center-Systemen: Der Stand der Forschung. *Psychologie und Praxis. Zeitschrift für Arbeits- und Organisationspsychologie*, *30*, 86-91.

Maukisch, H. (1989). Informationswert und Ökonomie der diagnostischen Prinzipien von Assessment Center Systemen zur Erfassung von Management Potential. In C. Lattmann (Hrsg.), *Das Assessment-Center-Verfahren der Eignungsbeurteilung. Sein Aufbau, seine Anwendung und sein Aussagegehalt* (S.251-289). Heidelberg: Physica-Verlag.

Moses, J.L. & Byham, W.C. (Eds.) (1977). *Applying the Assessment Center Method*. New York: Pergamon Press.

Neubauer, R. (1980). Die Assessment-Center-Technik: Ein verhaltensorientierter Ansatz zur Führungskräfteauswahl. In R. Neubauer und L. v. Rosenstiel (Hrsg.), *Handbuch der Angewandten Psychologie, Band 1: Arbeit und Organisation* (S. 122-158). München: Moderne Industrie.

Neubauer, R. (1989). Implizite Eignungstheorien im Assessment Center (AC). In C. Lattmann (Hrsg.), *Das Assessment-Center-Verfahren der Eignungsbeurteilung: Sein Aufbau, seine Anwendung und sein Aussagegehalt* (S. 191-221). Heidelberg: Physica-Verlag.

Neuberger, O. (1989) Assessment Center – Ein Handel mit Illusionen? In C. Lattmann (Hrsg.), *Das Assessment-Center-Verfahren der Eignungsbeurteilung. Sein Aufbau, seine Anwendung und sein Aussagegehalt* (S. 291-307). Heidelberg: Physica-Verlag.

Obermann, C. (1992). *Assessment Center. Entwicklung - Durchführung - Trends.* Wiesbaden: Gabler.

Pfeifer, A. & Schmidt, P. (1987). *LISREL - Die Analyse komplexer Strukturgleichungsmodelle.* Stuttgart: Fischer Verlag.

Reilly, R.R., Henry, S. & Smither, J.W. (1990). An examination of the effects of using behavior checklists on the construct validity of assessment center dimensions. *Personnel Psychology, 43*, 71-84.

Revenstorf, D. (1980). *Faktorenanalyse*. Mainz: Kohlhammer.

Richards, S.A. & Jaffee, C.L. (1972). Blacks supervising whites: A study of interracial difficulties in working together in a simulated organization. *Journal of Applied Psycholgy*, *56*, 234-240.

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.

Russell, C.J. (1987). Person characteristic versus role congruency explanations for assessment center ratings. *Academy of Management Journal*, *30*, 817-826.

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

Sarges, W. (1996). Einleitung des Herausgebers: Die Assessment Center-Methode – Herkunft, Kritik und Weiterentwicklungen. In W. Sarges (Hrsg.), *Weiterentwicklungen der Assessment Center-Methode*. Göttingen: Verlag für Angewandte Psychologie.

Schmitt, N. (1977). Interrater agreement in dimensionality and combination of assessment center judgments. *Journal of Applied Psychology*, *62*, 171-176.

Scholz, G. (1994). *Das Assessment Center: Konstruktvalidität und Dynamisierung.* Stuttgart: Verlag für Angewandte Psychologie.

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

Schröder, P. (1997). *Das Erleben und die Bewertung des Assessment Center Verfahrens eines Großunternehmens aus der Sicht der Teilnehmer. Ein empirischer Beitrag zur sozialen Validität.* Unveröffentlichte Diplomarbeit, Universität Hamburg.

Schuler, H. (1987). Assessment Center als Auswahl- und Entwicklungsinstrument: Einleitung und Überblick. In H. Schuler & W. Stehle (Hrsg.), *Assessment Center als Methode der Personalentwicklung* (S. 1-35). Stuttgart: Verlag für Angewandte Psychologie.

Schuler, H. (1989). Die Validität des Assessment Centers. In C. Lattmann (Hrsg.), *Das Assessment-Center-Verfahren der Eignungsbeurteilung. Sein Aufbau, seine Anwendung und sein Aussagegehalt* (S. 223-250). Heidelberg: Physica-Verlag.

Schuler, H. (1990). Personenauswahl aus der Sicht der Bewerber: Zum Erleben eignungsdiagnostischer Situationen. *Zeitschrift für Arbeits- und Organisationspsychologie*, *34*, 184-191.

Schuler, H. & Moser, K. (1995). Geschichte der Managementdiagnostik. In W. Sarges (Hrsg.), *Management-Diagnostik* (S. 32-42). (2. überarbeitete Auflage). Göttingen: Hogrefe.

Schuler, H. & Stehle, W. (1983). Neuere Entwicklungen des Assessment Center Ansatzes – beurteilt unter dem Aspekt der sozialen Validität. P*sychologie und Praxis. Zeitschrift für Arbeits- und Organisationspsychologie*, *27*, 33-44.

Schuler, H. & Stehle, W. (Hrsg.) (1987). *Assessment Center als Methode der Personalentwicklung.* Stuttgart: Verlag für Angewandte Psychologie.

Schwertfeger, B. (1999). Die Guten ins Töpfchen. In Assessment Centern werden die Manager von morgen ausgelesen. In: *Die Zeit,* Nr.13, S.9.

Sichler, R. (1989). Das Erleben und die Verarbeitung eines Assessment-Center-Verfahrens: Ein empirischer Beitrag zur „Sozialen Validität" eignungsdiagnostischer Situationen. *Zeitschrift für Arbeits-Organisationspsychologie*, *33*, 139-145.

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.

Simoneit, M. (1933). *Wehrpsychologie. Ein Abriß ihrer Probleme und praktischen Folgerungen.* Berlin: Verlag Bernard & Graefe.

Templer, K.-J. (1995). Zusammenhänge zwischen Aufgabentypen beim Assessment Center. *Zeitschrift für Arbeits- und Organisationspsychologie*, *39,* 179-181.

Thornton, G.C. & Byham, W.C. (1982). *Assessment Centers and Managerial Performance.* New York: Academic Press.

Turnage, J.J. & Muchinsky, P.M. (1982). Transsituational variability in human performance within assessment centers. *Organizational Behavior and Human Performance*, *30*, 174-200.

Wittenberg, R. (1991). *Grundlagen computerunterstützter Datenanalyse*. Stuttgart: Gustav Fischer Verlag.

Wolf, B., Barrel, G. & Hoenle, S. (1995). Einleitung. In Arbeitskreis Assessment Center (Hrsg.), *Assessment Center auf dem Prüfstand. Bewährungskontrolle und Qualitätsicherung am Beispiel eines Unternehmens-ACs* (S. 9-12). Hamburg: Windmühle.

Wolf, B., Barrel, G. & Hoenle, S. (1996). Ein Unternehmens-AC auf dem Prüfstand – Validierung in der Praxis. In Arbeitskreis Assessment Center (Hrsg.), *Assessment Center als Instrument der Personalentwicklung. Schlüsselkompetenzen, Qualitätsstandards, Prozeßoptimierung* (S. 221-241). Hamburg: Windmühle.

7 Anhang

- Aufbau eines Beurteilungsbogens am Beispiel der Präsentations-Übung 1-

Beurteilungsbogen

Präsentations-Übung 1

Beobachter	Teilnehmer(in)
_____	_____

	übertrifft die Anforderungen bei weitem	übertrifft die Anforderungen	erfüllt die Anforderungen voll	erfüllt die Anforderungen mit Abstrichen	erfüllt die Anforderungen nur zum Teil	erfüllt die Anforderungen nicht
	6	5	4	3	2	1
Organisation	☐	☐	☐	☐	☐	☐
Analysefähigkeit	☐	☐	☐	☐	☐	☐
Entscheidungsfähigkeit	☐	☐	☐	☐	☐	☐

Stärken

_____
_____
_____
_____
_____
_____

Schwächen

_____
_____
_____
_____
_____
_____

Erklärung

Ich versichere hiermit, daß ich die vorliegende Arbeit mit dem Thema:

„Konstruktvalidität und Assessment Center. Ein empirischer Beitrag."

selbständig verfaßt und keine anderen Hilfsmittel als die angegebenen verwendet habe. Die Stellen, die anderen Werken dem Wortlaut oder dem Sinn nach entnommen sind, habe ich in jedem einzelnen Fall durch Angabe der Quelle, auch der benutzten Sekundärliteratur, als Entlehnung gekennzeichnet.

Hamburg, den 09.12.1999

Kristof Kupka