Die Anwendung von Multitrait-Multimethod Modellen in der Assessment-Center-Diagnostik der Bundeswehr

Rolf Steyer
Ivailo Partchev, Sandra Menz und Katrin Seiß

Bericht Nr. 2 zum Forschungsvorhaben InSanI – 1190 – V 9000

Stand: 21. Februar 2001

Anschrift der Autoren
Universität Jena
Institut für Psychologie
Lehrstuhl für Methodenlehre und Evaluationsforschung
Am Steiger 3, Haus 1
07743 Jena

Tel. 03641/945231
Fax 03641/945232
Email: rolf.steyer@uni-jena.de
Homepage: http://www.uni-jena.de/svw/metheval

Die Anwendung von Multitrait-Multimethod Modellen in der Assessment-Center-Diagnostik der Bundeswehr

Rolf Steyer, Ivailo Partchev, Sandra Menz und Katrin Seiß

Wie beim computerunterstützten adaptiven Testen ist die Bundeswehr auch mit der Einführung der Assessment-Center-Diagnostik (zur Einführung s. z. B. Höft & Funke, in Druck) modernen Entwicklungen im Bereich der Eignungsdiagnostik gefolgt. Der zweifellos hohen inhaltlichen Validität, die sicherlich einer der Gründe für die hohe Akzeptanz des Verfahrens ist, steht aber durchaus eine beträchtliche Unsicherheit über die Kriteriumsvalidität gegenüber, die letztlich auf Zweifeln an der Reliabilität der erhobenen Urteile beruht. Wie zuverlässig sind die Urteile der Rater? Wie gut können die zu beurteilenden Eigenschaften der Probanden gemessen werden? Welchen Einfluss haben die unterschiedlichen Übungen (Präsentation, Postkorb, Gruppendiskussion etc.) auf die Urteile der Rater? Dieses sind nur einige der Fragen, die ein derart komplexes Verfahren der Eignungsdiagnostik aufwirft. Weitere Fragen ergeben sich, wenn man auch berücksichtigt, dass die *Beobachtungen* der Probanden *in Gruppen* gemacht werden, womit auch sozialpsychologische Effekte ins Spiel kommen. Verhält sich ein Proband Müller in einer Gruppe A von Probanden genauso wie in einer Gruppe B? Hätte Proband Müller in Gruppe B eine sehr viel aktivere Rolle („Führung“) einnehmen können als in Gruppe A, in der Proband Meier dies getan hat?

Man kann wohl sagen, dass man mit der Assessment-Center-Diagnostik bewusst die Komplexität des Alltags anstrebt, in der Hoffnung, damit validere Urteile und Prognosen über das Verhalten der Probanden im späteren beruflichen Alltag treffen zu können. Dies ist das gerade Gegenteil dessen, was man in der herkömmlichen Diagnostik anstrebt: Verhaltensunterschiede zwischen Probanden sollen nach Möglichkeit nur auf ihre Unterschiede in der betrachteten Eigenschaft zurückgeführt werden können; alle anderen Varianzquellen sind nach Möglichkeit auszuschalten. Ziel ist in der traditionellen Diagnostik also keinerlei Rater-Effekte zu haben (Objektivität), keine Situationseffekte (Konsistenz), homogene Items, die nur die betreffende Eigenschaft messen, etc.

Insgesamt gesehen dürfte es zum gegenwärtigen Zeitpunkt kaum möglich sein, ein abschließendes Urteil darüber zu fällen, ob die Assessment-Center-Diagnostik der herkömmlichen Diagnostik in den wichtigsten Validitätskriterien überlegen ist, zumal da zu bedenken ist, dass das Validitätskriterium „Vorgesetztenurteil“ die Assessment-Center-Diagnostik einseitig bevorzugt, handelt es sich doch um den gleichen Urteilsmodus. Damit wird eine wichtige Varianzquelle konstantgehalten, die zu einer artifiziellen Erhöhung der Korrelation zwischen Urteilen im Assessment Center und dem Vorgesetztenurteil führt.

Ziel der vorliegenden Arbeit ist es, Möglichkeiten der Analyse von in Assessment Centern erhobenen Daten aufzuzeigen, die Aufschluss über die Wichtigkeit verschiedener Varianzquellen in den diagnostischen Urteilen der Assessoren geben. Dazu geben wir zuerst einen Überblick über die einschlägige Literatur, stellen Modelle zur Analyse von Multitrait-Multimethod-Daten dar, und illustrieren dann die vorgestellten Modelle anhand eines Datensatzes aus den in der Offizierbewerberprüfzentrale (OPZ) der Bundeswehr durchgeführten Assessment Centern.

1. Zum Stand der Literatur

Jeserich (1981, S.33) definiert das Assessment-Center-Verfahren als „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.“ In dieser Definition werden die Grundprinzipien des Verfahrens deutlich: Ein Teilnehmer wird von *mehreren* Beobachtern in *mehreren* Situationen hinsichtlich *mehrerer* Merkmale beobachtet und beurteilt.

Die Beurteilung der Güte und Angemessenheit verschiedener statistischer Verfahren zur Analyse von Daten, die in Assessment-Center-Verfahren gewonnen werden, setzt zunächst eine kritische Auseinandersetzung mit der Frage voraus, inwieweit mit dem Design von Assessment Centern den Grundprinzipien des Verfahrens Rechnung getragen wird. Hierbei ist festzustellen, dass die Anzahl der erhobenen Merkmale in den bislang zu diesem Thema durchgeführten Studien erheblich variiert. Während in einigen Studien lediglich drei Merkmale erhoben wurden (z.B. Kleinmann, Andres, Fedtke, Godbersen & Köller, 1994; Kleinmann & Köller, 1997; Kuptsch, Kleinmann & Köller, 1998), kam es in anderen Untersuchungen zur Anwendung von bis zu 16 verschiedenen Beurteilungsdimensionen (z.B. Sackett & Dreher, 1982). Auch die Anzahl der in Assessment-Center-Verfahren eingesetzten Übungen variiert in doch beträchtlichem Ausmaß. Neben Studien, in denen die interessierenden Merkmale in nur drei verschiedenen Situationen erhoben wurden (z.B. Kleinmann et al., 1994; Kleinmann et al., 1997; Kuptsch et al., 1998), existieren auch Verfahren mit bis zu acht unterschiedlichen Übungen (z.B. Fleenor, 1996). Das Grundprinzip der Beobachtung eines Teilnehmers durch mehrere Assessoren wurde bisher jedoch nur teilweise umgesetzt. Während beispielsweise in dem von Kleinmann et al. (1994) untersuchten Verfahren ein Teilnehmer von einer pro Übung zufällig zusammengesetzten Dreiergruppe von Beobachtern beurteilt wird, verwenden andere Autoren ein Beobachtungsdesign, bei dem jeder Beobachter die ihm zugewiesenen Teilnehmer allein beurteilt (z.B. Turnage & Muchinsky, 1982; Bycio, Alvares & Hahn, 1987).

Das optimale Design von Assessment Centern trägt jedoch nicht nur den Grundprinzipien des Verfahrens Rechnung, sondern ermöglicht auch die Kontrolle eventuell verfälschender Einflussfaktoren auf das Verhalten der Teilnehmer, wodurch eine exakte und zuverlässige Abschätzung von Personparametern möglich wird. Dieser Gesichtspunkt hinsichtlich des Designs von Assessment-Center-Verfahren findet in Studien zu diesem Thema unterschiedlich starke Berücksichtigung. Diese Einflussfaktoren an sich, sowie Ansätze zu ihrer Kontrolle werden im Folgenden näher erläutert: (a) die Assessoren (Beobachter und Rollenspieler) des Assessment Centers, (b) weitere Teilnehmer und schließlich (c) unerwünschte Effekte der verschiedenen Übungen des Verfahrens, welche, neben den erwünschten übungsspezifischen Effekten, Einfluss auf das Verhalten und die Beurteilung der Teilnehmer nehmen können.

1.1 Assessoren als Fehlerquellen im Assessment Center

Beobachter im Assessment Center können in verschiedener Hinsicht Probleme aufwerfen. Eine recht umfassende Darstellung dazu findet man z. B. bei Lammers (1992). So können die als *Beobachter* agierenden Assessoren zu einer Quelle potentieller Konfundierung von übungsspezifischen Effekten und beobachtungsspezifischen Effekten werden, wenn Korrelationen zwischen einer bestimmten Übung und einem bestimmten Beobachter beispielsweise dadurch entstehen, dass ein bestimmter Beobachter bzw. eine Gruppe von Beobachtern in nur einer Übung tätig wird, wie es beispielsweise in einer Studie von Bycio et al. (1987) der Fall war. Gerade bei dieser Zuordnung von Beobachtern zu Übungen haben bestimmte von der Literatur belegte und nachfolgend kurz angeführte Einflüsse der Beobachter verfälschenden Effekte. So beschreiben Andres und Kleinmann (1993) klassische Beobachterfehler, wie z.B. Halo-Effekte, Primacy-Effekte und Ähnlichkeitsphänomene. Guzzo, Wagner, Maguire, Herr

und Hawley (1986) weisen darauf hin, dass in Assessment Centern Beurteilungen bestimmter Aspekte des Gruppenprozesses (z.B. die Art und Weise, wie Gruppenmitglieder mit Konflikten umgehen) vor dem Hintergrund von Wissen über die Ergebnisse der Gruppenarbeit getroffen werden (z.B. ob eine Assessment-Center-Aufgabe erfolgreich absolviert wurde). Folgt man den sozialpsychologischen Befunden der Autoren, so ist anzunehmen, dass diese Beurteilungen eben aufgrund des Wissens über den Erfolg oder Nichterfolg der Gruppe verfälscht sein könnten. Neben diesen Effekten, die theoretisch die Beurteilung aller Teilnehmer verzerren können, verweisen Andres und Kleinmann (1993) weiterhin auf Sympathieeffekte, die hinsichtlich der Beurteilung einzelner Teilnehmer wirksam werden können. Einen weiteren Effekt beschreibt Schriesheim (1981, zit. nach James, Demaree & Wolf, 1984): Bei der Beurteilung von Vorgesetzen durch Assessoren, die diesen im realen Arbeitsleben untergeordnet sind, zeigen sich sog. Mildeeffekte. Aber auch konkrete Eigenschaften von Beobachtern können u. U. Einfluss auf Beobachtung und Beurteilung der Teilnehmer haben. Schmitt (1993) verweist auf die Möglichkeit eines Einflusses der Rasse der Beobachter auf die Ratings, indem er nachweisen konnte, dass in dem von ihm untersuchten Verfahren einige Ratings von weißen Beobachtern signifikant höher als die von schwarzen Beobachtern waren und dass schwarze Beobachter Teilnehmer ihrer Rasse höher als weiße Teilnehmer beurteilten. Lowry (1993) untersuchte ebenfalls den Einfluss verschiedener Beobachtercharakteristika hinsichtlich ihrer Effekte auf die Ratings der Teilnehmer und belegte, dass die Beobachtermerkmale Alter und Rang/Position signifikante Varianzquellen darstellen. Insgesamt jedoch ist die Befundlage hinsichtlich des Einflusses bestimmter Merkmale der Beobachter auf die Ratings der Teilnehmer recht uneinheitlich. So konnte beispielsweise Carrier (1990, zit. nach Lowry, 1993) keine Effekte des Ranges bzw. der Position der Beobachter in der Organisation feststellen. Weiterhin muss in Bezug auf die soeben angeführten Befunde berücksichtigt werden, dass deren Generalisierbarkeit über die jeweils untersuchte Stichprobe hinaus kritisch zu betrachten und damit keineswegs abgesichert ist. [In der Studie von Lowry (1993) konnten die signifikanten Merkmale lediglich 2% der Varianz aufklären. Bei Schmitt (1993) wurden sehr geringe Unterschiede aufgrund der großen Stichprobe signifikant.]

Schließlich wird verschiedentlich und wiederholt in der Literatur darauf hingewiesen, dass die begrenzten Informationsverarbeitungskapazitäten der Beobachter beim Design von Assessment Centern Berücksichtigung finden sollten. Nach Schneider und Schmitt (1992) stellt eine kognitive Überlastung der Beobachter eine potentielle Fehlerquelle dar. In diesem Sinne argumentieren auch Gaugler und Thornton (1989), indem sie nachweisen konnten, dass die Anzahl der von den Beobachtern zu beurteilenden Dimensionen die Genauigkeit der Klassifikation und Beurteilung von Verhalten beeinflusst: Beobachter, die das Verhalten der Teilnehmer anhand einer geringen Anzahl von Dimensionen zu beurteilen hatten, klassifizierten dieses Verhalten hinsichtlich dieser Dimensionen genauer und beurteilten es mit grösserer Genauigkeit als Beobachter, die mit einer größeren Anzahl von Dimensionen arbeiteten.

Doch neben den Assessoren, die als Beobachter agieren, können nach Schneider und Schmitt (1992) auch die als *Rollenspieler* am Assessment Center beteiligten Assessoren einen verfälschenden Einfluss auf das Verhalten der Teilnehmer und damit auf die mittels des Verfahrens gewonnenen Daten haben. Alle oben für die Beobachter aufgeführten Fehler- bzw. Konfundierungsquellen können auch bei diesen Assessoren ihren Ausgang nehmen und müssen beim Design des jeweiligen Verfahrens Berücksichtigung finden.

In der Literatur existieren verschiedene Ansätze zur Kontrolle bzw. Ausschaltung vieler der bisher angeführten Effekte von Assessoren auf die Beobachtung und Beurteilung der Teilnehmer. Andres und Kleinmann (1993) haben ein Rotationssystem für die Beobachtungssituation in Assessment Centern entwickelt, bei dem zum einen keine feste Kopplung von einem bestimmten Beobachter und einer bestimmten Übung entsteht (alle Beobachter

durchlaufen im Verlauf des Verfahrens alle Übungen) und zum anderen jeder Beobachter jeden Teilnehmer genau einmal beobachtet. In ähnlicher Weise wurden in dem von Schneider und Schmitt (1992) untersuchten Verfahren die als Rollenspieler agierenden Assessoren den verschiedenen Übungen derart zugewiesen, dass die Teilnehmer mit jeweils unterschiedlichen Assessoren interagierten. Auch Kudisch, Ladd und Dobbins (1997) berücksichtigen das Problem der Konfundierung von Beobachter- und Übungseffekten, indem ein Beobachtungsdesign eingesetzt wird, bei dem verschiedene Beobachter in verschiedenen Übungen beurteilen. Eine weitere Maßnahme zur Reduktion von Beobachtungsfehlern (und auch zur Standardisierung des Verhaltens der Assessoren, die als Interaktionspartner der Teilnehmer agieren) stellt ein systematisches Training der Assessoren dar. Verschiedene Autoren messen dem Beobachtertraining bzw. seiner Gründlichkeit unterschiedliche Bedeutung zu. Laut Gaugler und Thornton (1989) besteht kein Beleg für die Annahme, dass ein bestimmtes Minimum an Zeit notwendig wäre, um Assessoren hinsichtlich der grundlegenden Beobachtungs- und Beurteilungsprozesse zu trainieren. Verschiedene Autoren (u.a. Schneider & Schmitt, 1992) weisen jedoch darauf hin, dass ein gründliches Beobachtertraining zur Verbesserung der Konstruktvalidität des Assessment Centers beitrage und aufgrund dessen von Bedeutung sei. Sagie und Magnezy (1997) stellten fest, dass die von Psychologen vorgenommenen Ratings durch eine Fünf-Faktoren-Struktur beschreibbar waren, wohingegen Manager nur zwei verschiedene Dimensionen nutzten, anhand derer sie beobachtete Verhaltensweisen klassifizierten. Die Autoren schlussfolgern aus diesem Befund, dass Managern in einem kurzen Beobachtertraining nicht dieselben analytischen Kompetenzen vermittelt werden können, wie Psychologen sie aufgrund langjähriger Erfahrung erworben haben, und liefern damit ein weiteres Argument für die Notwendigkeit eines umfassenden Trainings der Assessoren. Ein weiterer Faktor, der beim Design eines Assessment-Center-Verfahrens berücksichtigt werden sollte, ist schließlich die begrenzte Informationsverarbeitungskapazität der Beobachter, weswegen die Anzahl zu beurteilender Dimensionen gering gehalten werden sollte. Kleinmann et al. (1997) schlagen vor, dass Beobachter nicht mehr als drei Dimensionen pro Teilnehmer und Übung beobachten und beurteilen sollten. Neben dieser Reduktion der zu beobachtenden Dimensionen zeichnet sich ein weiterer Trend zur Handhabung dieses Aspektes ab: der Einsatz von Verhaltenschecklisten zur Reduktion der kognitiven Anforderungen an die Beobachter (vgl. Donahue, Truxillo, Cornwell und Gerrity, 1997). Reilly et al. (1990) und Donahue et al. (1997) meinen, dass eine derartige Checkliste die an die Beobachter gestellten kognitiven Anforderungen tatsächlich reduziert.

Neben den bisher beschriebenen möglichen Richtungen einer Einflussnahme (einseitig gerichtete Einflüsse der Assessoren auf die Teilnehmer und wechselseitige Interaktionen zwischen Assessoren und Teilnehmern) besteht auch die Möglichkeit, dass *Interaktionen zwischen den Assessoren* zu Fehlerquellen hinsichtlich Beobachtung und Beurteilung der Teilnehmer werden können. Dem Prinzip des Assessment-Center-Verfahrens, dass die Teilnehmer jeweils von mehreren Assessoren beobachtet werden, wird nicht nur bei der Beobachtung der Teilnehmer, sondern auch bei deren anschließender Beurteilung in sog. Beobachterkonferenzen Rechnung getragen. Dementsprechend ist es allgemein üblich, dass nicht nur die Beobachtung eines Teilnehmers, sondern auch dessen Beurteilung von einer Gruppe mehrerer Beobachter vorgenommen wird. Nach Byham (1980) rechtfertigt sich dieser Gruppenansatz aufgrund mehrerer mit ihm verbundener Vorteile. Zum einen entsteht für die beteiligten Beobachter die Möglichkeit, gesammelte Verhaltensinformationen zu vergleichen und zu kontrastieren. Weiterhin wird der Einfluss der oben beschriebenen Beobachtungsfehler reduziert, indem die letztlich getroffene Entscheidung hinsichtlich der Beurteilung des Teilnehmers eine von allen Beobachtern getragene Gruppenentscheidung ist. Schließlich wird hierdurch auch die Wahrscheinlichkeit richtiger Entscheidungen aufgrund der Kombination verschiedener Meinungen maximiert. Mehrere Autoren weisen jedoch darauf hin, dass mit

dieser Vorgehensweise zwar eine einseitige verfälschende Einflussnahme der Beobachter auf die Teilnehmer kontrolliert werden kann, sich andererseits jedoch neue Probleme aus diesem Ansatz ergeben. Andres und Kleinmann (1993, S.19) argumentieren, dass „Machtspiele und Beobachtungsbündnisse“ zwischen Beobachtern kontrolliert werden müssen, da „Sympathie bzw. Spannungen zwischen den Beobachtern [...] sich schnell auf die Bewertung der einzelnen Teilnehmer auswirken.“ Um diese negativen Interaktionseffekte bei der Beurteilung der Teilnehmer zu vermeiden, schlagen die Autoren die bereits angesprochene Rotation zwischen Beobachtern vor, so dass keine „ungleichmäßige Variation der Zusammensetzung der Beobachtergruppe [...] diese unerwünschten Effekte“ verstärken kann (Andres & Kleinmann, 1993, S. 20). Einen anderen Ansatz zur Kontrolle des negativen Einflusses derartiger Interaktionseffekte (der gleichzeitig die an die Beobachter gestellten kognitiven Anforderungen reduziert) diskutiert Lowry (1993). Er schlägt eine „non-evaluative intergrating discussion“ (Lowry, 1993, S. 498) vor. Bei einer derart durchgeführten Beobachterkonferenz tragen die Beobachter nur die von ihnen notierten Verhaltensinformationen, nicht jedoch deren Bewertung vor; die jeweils zuhörenden Beobachter dürfen ebenfalls keine Evaluation dieser Informationen vornehmen. Diese Vorgehensweise erleichtert es den Assessoren, sich an Verhaltensinformationen zu erinnern, die sie während des Beobachtungsvorganges möglicherweise nicht notiert haben, und dennoch zu einem unabhängigen und nicht von Interaktionseffekten verfälschten Urteil zu gelangen.

1.2 Andere Teilnehmer als Fehlerquellen im Assessment Center

Sozialpsychologische Effekte der Beeinflussung des Verhaltens eines Einzelnen aufgrund der Gegenwart einer Gruppe kommen auch im Assessment Center, und hier vor allem in Gruppenübungen, zum tragen. Aber nicht nur das Verhalten des einzelnen Teilnehmers, sondern auch dessen Beurteilung durch die Beobachter findet nicht unabhängig vom Verhalten anderer Assessees, der Interaktion mit diesen und deren Beurteilung statt. Butler (1989, zit. nach Schneider & Schmitt, 1992) stellte fest, dass die Beurteilung eines Teilnehmers in einer Gruppendiskussion günstiger ausfällt, wenn dieser Teilnehmer zusammen mit zwei unterdurchschnittlichen Assessees agiert, während die Bewertung seines Verhaltens weniger positiv ausfällt, wenn er die Übung in Interaktion mit zwei überdurchschnittlichen Teilnehmern absolviert. Schmitt und Hill (1977, zit. nach Schmitt, 1993) konnten nachweisen, dass die Beurteilung weiblicher Kandidaten hinsichtlich der Dimension Führungsverhalten ungünstiger ausfällt, wenn die Frauen in einer Gruppe interagieren, die sich hauptsächlich aus männlichen Teilnehmern zusammensetzt und lieferten damit einen Beleg für den Einfluss der Gruppenzusammensetzung auf die Beurteilung des Einzelnen im Assessment Center.

Zur Kontrolle dieser gruppenspezifischen Interaktionseffekte setzen Schneider und Schmitt (1992) ein System der Rotation der Teilnehmer zwischen den Übungen ein, das zur Folge hat, dass die Teilnehmer in verschiedenen Gruppenübungen mit jeweils unterschiedlichen weiteren Teilnehmern interagieren. Auch Andres und Kleinmann (1993, S. 19) berücksichtigen in ihrem Rotationssystem die Tatsache, dass es sich „je nach Stärken-/Schwächenprofil der Teilnehmer [...] ungünstig auswirken [kann], wenn zwei Teilnehmer zu häufig oder zu selten in der gleichen Untergruppe zusammen sind.“

1.3 Unerwünschte übungsspezifische Effekte als Fehlerquellen im Assessment Center

Neben den erwünschten Effekten verschiedener Übungen (nach Neidig & Neidig, 1984, u.a. die Möglichkeit zur Verhaltensbeobachtung in verschiedenen, tätigkeitsbezogenen Übungen

sowie die Konfrontation der Teilnehmer mit unterschiedlichen Anforderungen) können auch unerwünschte übungsspezifische Effekte zu Fehlerquellen im Assessment-Center-Verfahren werden. Ein erster Faktor, der laut Schneider und Schmitt (1992) zu Verfälschungen führen kann, ist eine stets gleiche Abfolge der Übungen des Assessment Centers. Um diesen Effekt der Reihenfolge der Übungen kontrollieren zu können, schlagen die Autoren vor, den Ablaufplan des Assessment Centers systematisch so zu variieren, dass nicht alle Teilnehmer die Übungen in immer gleicher Reihenfolge durchlaufen. Ein zweiter verfälschender Faktor ergibt sich aus der Tatsache, dass unterschiedliche Übungen u. U. in verschiedenem Ausmaß die Möglichkeit bieten, die jeweils interessierenden Verhaltensweisen zu beobachten. Deshalb sei darauf zu achten, dass die Übungen und Dimensionen eines Assessment-Center-Verfahrens wohlüberlegt ausgewählt werden, um sicherzustellen, dass sich für die Assessoren die Möglichkeit eröffnet, die dimensionsbezogenen Verhaltensweisen mit relativ gleicher Häufigkeit über verschiedene Übungen hinweg beobachten zu können.[1]

2. Verfahren zur Analyse von Assessment-Center-Daten

In der Literatur finden sich verschiedene statistische Vorgehensweisen zur Analyse der in Assessment-Centern gewonnenen Daten. Ein vielfach angewandtes Verfahren ist die Analyse von sog. Multitrait-Multimethod Korrelationsmatrizen (MTMM-Matrizen) nach Campbell und Fiske (1959). Dabei stellen die in den Assessment-Centern eingesetzten Übungen die Methoden dar; die erfassten Merkmale sind Indikatoren für die Traits. Jedes Merkmal wird mit verschiedenen Übungen bzw. Methoden beurteilt. Unter Anwendung dieses Verfahrens lassen sich konvergente und diskriminante Validität des Assessment-Center-Verfahres abschätzen.[2] In der Literatur wird vielfältige Kritik an diesem Verfahren geübt. Nach Kleinmann et al. (1994) besteht neben anderen Problemen der Hauptkritikpunkt am Verfahren von Campbell und Fiske darin, dass die Auswertung der MTMM-Matrizen auf der Grundlage der Korrelationen zwischen messfehlerbehafteten, manifesten Variablen geschehe, wohingegen die nachfolgende Interpretation Schlussfolgerungen über latente Variablen (Trait- und Methodenfaktoren) enthalte.

In Anbetracht der Schwachstellen dieses Verfahrens werden gegenwärtig ergänzend faktorenanalytische Methoden (konfirmatorische Faktorenanalyse, CFA) für die Analyse von MTMM-Daten verwendet. Der Vorteil der CFA besteht in der Möglichkeit, die Daten modellgeleitet zu analysieren. Dabei ist es möglich, Faktorenladungen und/oder auch Korrelationen zwischen Faktoren (z. B. auf 0) zu fixieren und damit gezielt faktorenanalytische Modelle zu überprüfen. Computerprogramme wie LISREL (Jöreskog & Sörbom, 1999) und EQS (Bentler, 2001) werden zur Schätzung der entsprechenden theoretischen Parameter eingesetzt. Im folgenden werden wir ein solches konfirmatorisches MTMM-Modell vorstellen und an den Daten erläutern, wie sie in der Assessment-Center-Diagnostik in der Offizierbewerberprüfzentrale (OPZ) der Bundeswehr erhoben werden. Dabei handelt es sich um das Modell von Eid (2000). Zum besseren Verständnis beschreiben wir zunächst die Datenstruktur und erläutern dann die Modelle, die wir später mit diesen Daten analysieren.

---

1 Gerade um diesen Aspekt dreht sich eine in der Literatur geführte Kontroverse zwischen Sackett und Dreher (1982, 1984) und Neidig und Neidig (1984). Letztere vertreten im Gegensatz zu Sackett und Dreher, die die oben beschriebene Position befürworten, die Ansicht, dass es nicht Anliegen eines Assessment-Center-Verfahrens sein könne, Teilnehmer mit Übungen zu konfrontieren, die jeweils die gleiche Möglichkeit bieten, dimensionsbezogene Verhaltensweisen zu zeigen. Vielmehr käme es darauf an, die Tätigkeitsbezogenheit der Übungen des Verfahrens sicherzustellen.

2 Der interessierte Leser sei an dieser Stelle auf die detaillierte Beschreibung in Kleinmann et al. (1997) verwiesen.

2.1 Datenerhebung

Insgesamt liegen uns die Daten einer Stichprobe von $n = 297$ Bewerbern vor, die das Assessment center der OPZ durchlaufen haben. Dabei wurden u. a. die folgenden Eigenschaften (*traits*) erhoben: *Führungsfähigkeit* (Leader), *soziale Kompetenz* (Social), die *sprachliche Ausdrucksfähigkeit* (Language) und die *Urteilsfähigkeit* (Judgment) und zwar anhand der folgenden *Übungen*: *Rundgespräch* (Discussion), *Planspiel* (Game), und *Interview* (Interview). Einen dritten Faktor bilden die *Beurteiler* oder *Rater*: ein *Stabsoffizier*(StOff), ein *Prüfoffizier*(PrOff) und ein *Psychologe*(Psycho). Für jeden der $n = 297$ Bewerber liegen somit für die anstehende Betrachtung von MTMM-Modellen $4 \times 3 \times 3 = 36$ Variablen (Urteile) vor.[3]

Aus Gründen der Einfachheit haben wir nun zwei verschiedene Wege der Vereinfachung dieses Datensatzes beschritten. Zum einen haben wir (a) die *Urteile über der Rater* gemittelt. Dabei haben wir folgende Behandlung fehlender Werte vorgenommen. War mindestens ein Urteil vorhanden, so wurde über die vorliegenden (ein, zwei oder drei) Urteile gemittelt. Lag kein einziges Urteil vor, so galt das betreffende aggregierte Urteil als fehlend. Auf diese Weise erhielten wir für jeden von $n = 296$ Bewerber $4 \times 3 = 12$ Variablen (= über die Rater gemittelte Urteile). In diesem Datensatz stellen die Übungen die Methoden dar. Tabelle 1 identifiziert die Labels der 12 manifesten Variablen, wie sie gerade beschrieben wurden. In Tabelle 3 sind die Korrelationen, Standardabweichungen und Mittelwerte angegeben, auf denen die nachfolgenden Analysen beruhen.

Hier Tabelle 1 einfügen

Hier Tabelle 2 einfügen

Zum anderen wurden (b) die Urteile *über die verschiedenen Übungen* gemittelt. Die oben bereits beschriebene Missing-data-Technik führte so zu einer Stichprobengröße von $n = 216$. Auch hier liegen $4 \times 3 = 12$ Variablen (= über die Übungen gemittelte Urteile) vor. In Tabelle 2 sind wieder die 12 manifesten Variablen und ihre Labels angegeben. In Tabelle 4 findet man ihre Korrelationen, Standardabweichungen und Mittelwerte.

Hier Tabelle 3 einfügen

Hier Tabelle 4 einfügen

Mit beiden auf diese Weise reduzierten Datensätzen können wir MTMM-Modelle analysieren. In beiden Datensätzen kommen die oben genannten traits *Führungsfähigkeit* (Leader), *soziale Kompetenz* (Social), die *sprachliche Ausdrucksfähigkeit* (Language) und die *Urteilsfähigkeit* (Judgment) vor. Während im ersten Datensatz die Übungen die Rolle der Methoden einnehmen, sind es im zweiten die Beurteiler.

---

[3] Andere Variablen seien zunächst ausgeklammert. Außerdem wird aus Gründen der Einfachheit vernachlässigt, dass es sich bei der Beurteilern nicht immer um dieselben Personen handelt.

2.2 Konfirmatorische faktoranalytische MTMM-Modelle

Konfirmatorische faktoranalytische MTTM werden bereits seit einigen Jahrzehnten verwendet (s. z. B. Marsh & Grayson, 1995; Widaman, 1985). Dabei liegen Daten zugrunde, bei denen jede von mehreren Eigenschaften (Traits) mit mehreren Methoden (hier: Übungen oder Rater) erhoben werden. Das Grundprinzip dabei ist, einerseits die Traits als latente Variablen (Faktoren) und andererseits die differentiellen Methodeneffekte als Werte von anderen, mit den ersteren unkorrelierten latenten Variablen (Faktoren) darstellen zu können.

Nachdem schon über Jahrzehnte beobachtet wurde, dass die Schätzalgorithmen bei solchen Modellen oft nicht konvergieren oder die Lösungen sehr instabil waren, wurden in den letzten Jahren die Gründe dafür gefunden und alternative Modelle entwickelt, die dieses Problem nicht mehr aufweisen. Der Grund für die beobachteten Probleme ist die *lokale Unidentifiziertheit* der MTMM-Modelle, d. h. es existiert keine eindeutige Lösung für bestimmte Parameterkonstellationen (daher lokal) (s. Grayson & Marsh, 1994; Steyer, 1995; Wothke 1996)

Eid (2000) hat gezeigt, dass man diese Probleme dadurch vermeiden kann, dass man nicht, wie bei den herkömmlichen MTMM-Modellen üblich, genauso viele Methodenfaktoren einführt wie Methoden verwendet werden, sondern genau einen Faktor weniger. Eine der Methoden spielt dann einfach die Rolle eine *Referenzmethode*. Ihre Besonderheiten sind daher in der betreffenden Traitvariablen enthalten, die dann allerdings ein *methodenspezifischer* Trait ist. Gleichzeitig sind auch die Methodenfaktoren in dem Sinne spezifisch, da sie in Abhebung von der als Referenzmethode gewählten Methode definiert sind.

2.3 Analyse der über die Rater gemittelten Beurteilungen

An dem in Abbildung 1 dargestellten Pfaddiagramm eines konfirmatorischen faktoranalytischen MTMM-Modells soll diese Art von Modellen näher erläutert werden. Abbildung 1 bezieht sich auf die *nichtstandardisierte Lösung*, der die *Kovarianzmatrix* der 12 manifesten Variablen zugrunde liegt, die wir in Tabelle 1 beschrieben haben, nämlich die über die Rater gemittelten Urteile bzgl. der vier Eigenschaften, die in jeder der drei Übungen beobachtet wurden.

Die vier Ovale auf der rechten Seite der Abbildung 1 repräsentieren die vier *methodenspezifischen* latenten Traits, wobei hier als Referenzmethode das Rundgespräch gewählt wurde. Bei den vier latenten Variablen handelt es sich also um die True-Score-Variablen *Führungsfähigkeit (Leader)*, *Soziale Kompetenz (Social), Sprachliche Ausdrucksfähigkeit (Language)* und *Urteilsfähigkeit (Judgment)*, jeweils bezogen auf *das Rundgespräch*. Diese Rundgespräch-spezifischen Traits erklären nun auch einen erheblichen Teil der Varianz der (über die Rater gemittelten) Urteile der in den anderen Übungen beurteilten Eigenschaften (s. die erste Zeile in Tabelle 6).[4] Des weiteren ist auf der rechten Seite der Abbildung 1 zu erkennen, wie stark diese Rundgespräch-spezifischen Traits miteinander korrelieren – hier allerdings dargestellt als Kovarianzen. In Tabelle 5 sind die zugehörigen Korrelationen angegeben. Die Korrelation zwischen allen Traits sind hier größer als .80. Die sprachliche Ausdrucksfähigkeit (Language) korreliert mit der Urteilsfähigkeit (Judgment) sogar zu .91. Dies

[4] Wie groß diese erklärte Varianz ist, ergibt sich aus dem Produkt der quadrierten Ladung (als Zahlen an den Pfeilen dargestellt) mit der Varianz der latenten Variablen (die jeweilige Zahl neben dem Oval). Der erklärte Varianz*anteil* errechnet sich dann aus dem Quotienten der erklärten Varianz an der Varianz der betrachteten manifesten Variablen. In der standardisierten Lösung, die wir später darstellen werden, ergeben sich einige Vereinfachungen dadurch, dass manifeste und latente Variablen auf die Varianz 1 normiert werden.

zeigt, dass eine Beurteilung differenzieller Fähigkeiten nur begrenzt stattfindet oder aber, dass diese Fähigkeiten in der betrachteten Gruppe von Bewerbern eben sehr hoch korreliert sind.

Hier Abbildung 1 einfügen

Hier Tabelle 5 einfügen

Auf der linken Seite der Abbildung 1 sind die beiden Methodenfaktoren, hier also der Planspiel- (Game) und der Interviewfaktor (Interview) dargestellt. Dabei handelt es sich um denjenigen Anteil der True-score-Variablen der betrachteten manifesten Variablen, der nicht durch den jeweiligen Rundgespräch-spezifischen Trait (mittels einer linearen Regression) erklärt werden kann. Der Methodenfaktor (d. h. der Planspielfaktor) der Variablen „Führungsfähigkeit im Planspiel“ (LeaGam) bspw. ist also als Residuum seiner True-score-Variablen bzgl. seiner Regression auf die True-score-Variable der Variablen „Führungsfähigkeit im Rundgespräch“ (LeaDis) definiert (s. Eid, 2000). Anders ausgedrückt enthalten die Methodenfaktoren also alle systematischen Bestandteile der betreffenden manifesten Variablen, die sich nicht durch eine lineare Regression auf die True-score-Variable der Referenzvariable erklären lassen. Die so als Residuen definierten Methodenfaktoren können grundsätzlich miteinander korrelieren. Im vorliegenden Beispiel wird diese Korrelation .056 geschätzt. Eine Korrelation der Methodenfaktoren mit den Traitfaktoren wird durch die Definition von Traits und Methodenfaktoren als Residuen ausgeschlossen (s. Eid, 2000).

Der Wert eines Bewerbers auf dem Planspielfaktor, multipliziert mit der betreffenden Ladung der betrachteten manifesten Variablen auf dem Planspielfaktor, drückt daher aus, inwieweit das messfehlerbereinigte Urteil über seine Fähigkeit im Planspiel von dem abweicht, was aus seiner im Rundgespräch eingeschätzten Fähigkeit bereits vorhersagbar ist. Varianzanalytisch ausgedrückt gibt der durch einen Methodenfaktor erklärte Varianzanteil der betreffenden manifesten Variablen an, wie stark die Interaktion Person × Methode ist. Die Varianzanteile der Methodenfaktoren geben daher an, wie *heterogen* die verwendeten Methoden sind, hier, welche Rolle es also spielt, *welche* Übung zur Beurteilung der Eigenschaft des Bewerbers herangezogen wird. Allerdings ist zu beachten, dass diese erklärten Varianzanteile zunächst nur auf die gewählte Referenzmethode bezogen sind. Bei der Referenzmethode kann daher auch nicht zwischen den beiden erklärten Varianzanteilen unterschieden werden. In Tabelle 6 sind nun die durch die Traits und die durch die Methodenfaktoren erklärten Varianzanteile angegeben.

Hier Tabelle 6 einfügen

Die im Rundgespräch ermittelten Fähigkeiten erklären den größten Teil der Varianz der Beurteilungen im Planspiel und im Interview. Im Schnitt erklären sie ca. 58% der Varianz. Die Methodenfaktoren „Planspiel“ und „Interview“ erklären im Schnitt noch einmal ca. 18% der Varianz. Die Reliabilitäten der über die Rater gemittelten Urteile liegen fast alle über .80. Lediglich zwei Urteilsvariablen bilden hier eine Ausnahme: Die Beurteilung der Führungsfähigkeit und der sozialen Kompetenz im Interview. Die Fehlervarianz der Urteile über die Führungsfähigkeit beträgt 62%, die über die soziale Kompetenz 31%.

Die Traits beim vorliegenden Modell sind *methodenspezifisch*, ein methodenübergreifender Trait kommt in diesen Modellen nicht vor. Allerdings lassen sich die Korrelationen *aller*

methodenspezifischen Traits berechnen, auch wenn diese nicht alle explizit im Modell vorkommen. Diese methodenspezifischen Traits lassen sich als die True-score-Variablen der manifesten Variablen interpretieren. Im hier vorliegenden Modell gibt es davon also 12 Stück. Diese Korrelationen geben natürlich auch darüber Auskunft, wie stark die drei Rundgespräch-spezifischen Traits miteinander korreliert sind, ebenso wie die drei Planspiel-spezifischen Traits und die Interview-spezifischen Traits miteinander korrelieren. Diese Korrelationen (s. Tab. 7) unterscheiden sich von den Korrelationen der manifesten Variablen (s. Tab. 1) dadurch, dass sie messfehlerbereinigt sind. Ein Vergleich mit den entsprechenden Korrelationen der manifesten Variablen zeigt, dass die Korrelationen der messfehlerbereinigten Variablen weit höher sind, als die der manifesten Variablen. Auch ist das Muster typischer MTMM-Korrelationen deutlicher zu erkennen: hohe Korrelationen innerhalb des gleichen Traits, hohe Korrelationen auch innerhalb der gleichen Methoden und niedrigere Korrelationen dann, wenn weder Traits noch Methoden identisch sind (s. Tab. 7).[5]

Hier Tabelle 7 einfügen

Neben den Korrelationen dieser 12 True-score-Variablen kann man natürlich auch die Korrelationen der Summen der jeweils drei (übungsspezifischen) True-score-Variablen betrachten, mit denen ein gemeinsamer Trait gebildet werden soll.[6] Die entsprechenden Korrelationen (s. Tab. 8) sind jedoch in diesem Beispiel nur unwesentlich anders als die Korrelationen der übungsspezifischen True-score-Variablen.

Hier Tabelle 8 einfügen

2.4 Analyse der über die Übungen gemittelten Beurteilungen

Die oben dargestellten Analysen sollen nun auch für den zweiten Datensatz (s. Tab. 2 und 4), die über die Übungen gemittelten Beurteilungen präsentiert werden. In Abbildung 2 findet man das betreffende Pfaddiagramm.

Hier Abbildung 2 einfügen

Als Referenzmethode wurde die Beurteilung durch den Stabsoffizier gewählt. Die Beurteilungen durch den Prüfoffizier und den Psychologen tauchen daher explizit als Methodenfaktoren auf. Die Traits auf der rechten Seite des Pfaddiagramms sind also Stabsoffizier-spezifisch, d. h. sie sind als True-score-Variablen der Beurteilungen der jeweiligen Fähigkeit *durch den Stabsoffizier* zu interpretieren. Auch hier sind die Korrelationen zwischen allen Traits größer als .80 (s. Tab. 9), und auch hier ist die größte Korrelation mit .94 zwischen *Sprachlicher Ausdrucksfähigkeit* (Language) und *Urteilsfähigkeit* (Judgment).

Hier Tabelle 9 einfügen

---

5 Die Tabelle 7 erhält man, in dem man die 12 True-score-Variablen explizit als latente Variablen in das Strukturgleichungsmodell aufnimmt. Aus Gründen der Einfachheit wurde auf ein entsprechendes Pfaddiagramm verzichtet. Die entsprechenden LISREL-Steuerkarten findet man in Anhang A.

6 Die entsprechenden LISREL-Steuerkarten sind die gleichen wie die unter Fußnote 5 verwendeten.

Auf der linken Seite der Abbildung 2 sind wiederum die beiden Methodenfaktoren aufgeführt, hier der Prüfoffizier- und der Psychologenfaktor. Dabei handelt es sich um denjenigen Anteil der True-score-Variablen der betrachteten manifesten Variablen, der nicht durch den jeweiligen Stabsoffizier-spezifischen Trait (mittels einer linearen Regression) erklärt werden kann. Der Methodenfaktor (d. h. der Prüfoffizierfaktor) der Variablen „Führungsfähigkeit beurteilt durch den Prüfoffizier“ (LeaOff) bspw. ist also als Residuum seiner True-score-Variablen bzgl. seiner Regression auf die True-score-Variable der Variablen „Führungsfähigkeit beurteilt durch den Stabsoffizier“ (LeaSta) definiert (s. Eid, 2000). Die beiden Methodenfaktoren korrelieren in diesem Beispiel deutlich stärker miteinander, als dies beim vorangegangenen Beispiel der Fall war, nämlich .44.

Auch für dieses Beispiel können wir wieder die durch die Traits bzw. die Methodenfaktoren erklärten Varianzanteile angeben (s. Tab. 10). Dabei ist wiederum zu beachten, dass die Traits und die Methodenfaktoren *methodenspezifisch* (hier: Stabsoffizier-spezifisch) definiert sind. Bei der hier gewählten Referenzmethode kann daher auch nicht zwischen den beiden erklärten Varianzanteilen unterschieden werden. In Tabelle 10 sind nun die erklärten Varianzanteile angegeben.

Hier Tabelle 10 einfügen

Die durch den Stabsoffizier eingeschätzten Fähigkeiten erklären den größten Teil der Varianz der Beurteilungen des Prüfoffiziers und des Psychologen. Im Schnitt erklären sie ca. 53% der Varianz. Die Methodenfaktoren „Prüfoffizier“ und „Psychologe“ erklären im Schnitt noch einmal ca. 27% der Varianz. Acht von zwölf Reliabilitäten der über die Übungen gemittelten Urteile liegen über .80. Vier liegen zwischen .71 (Soziale Kompetenz beurteilt durch den Stabsoffizier) und .78 (Führungsfähigkeit beurteilt durch den Psychologen). Bei dieser Form der Aggregation fällt also keine der Reliabilitäten aus dem Rahmen.

Auch wenn die Traits wiederum methodenspezifisch sind, lassen sich die Korrelationen *aller* methodenspezifischen Traits berechnen.[7] Auch hier sind die methodenspezifischen Traits als die True-score-Variablen der manifesten Variablen zu interpretieren. Im hier vorliegenden Modell gibt es davon also 12 Stück. Diese Korrelationen geben natürlich auch darüber Auskunft wie stark die drei Stabsoffizier-spezifischen Traits miteinander korreliert sind, ebenso wie die drei Prüfoffizier-spezifischen Traits und die Psychologen-spezifischen Traits miteinander korrelieren. Diese Korrelationen (s. Tab. 11) unterscheiden sich von den Korrelationen der manifesten Variablen (s. Tab. 4) dadurch, dass sie messfehlerbereinigt sind. Ein Vergleich mit den entsprechenden Korrelationen der manifesten Variablen zeigt, dass die Korrelationen der messfehlerbereinigten Variablen weit höher sind, als die der manifesten Variablen. Auch ist das Muster typischer MTMM-Korrelationen deutlicher zu erkennen: hohe Korrelationen innerhalb des gleichen Traits, hohe Korrelationen auch innerhalb der gleichen Methoden und niedrigere Korrelationen dann, wenn weder Traits noch Methoden identisch sind (s. Tab. 11).

Hier Tabelle 11 einfügen

Neben den Korrelationen dieser 12 True-score-Variablen kann man natürlich auch hier wieder die Korrelationen der Summen der jeweils drei (Rater-spezifischen) True-score-

---

[7] Diese Ergebnisse erhält man analog wie in Fußnote 5 angegeben. Die entsprechenden LISREL-Steuerkarten findet man wieder im Anhang.

Variablen betrachten, mit denen ein gemeinsamer Trait gebildet werden soll.[8] Die entsprechenden Korrelationen (s. Tab. 12) sind jedoch in diesem Beispiel ebenfalls nicht sehr anders als die die Korrelationen der übungsspezifischen True-score-Variablen, die wir bereits in Tabelle 9 betrachtet haben. Die höchste Korrelation (.94) ist die zwischen Sprachlicher Ausdrucksfähigkeit (Language) und Urteilsfähigkeit (Judgment).

Hier Tabelle 12 einfügen
---

3. Diskussion

Die Vorteile der oben dargestellten CFA-MTMM-Modelle sind darin zu sehen, dass zum einen die in ihnen vorkommenden latenten Variablen alle wohldefiniert sind, entweder als True-score-Variablen oder aber als Residuen bzgl. einer linearen Regression von bestimmten True-score-Variablen auf eine andere. Zum anderen sind diese Modelle global und lokal identifiziert. Die Lösungen sind stabil und in der Regel beschreiben die Modelle die Daten sehr gut. Beide dargestellten Modelle sind die von Eid (2000) eingeführten Standardmodelle. Es wurde in beiden Fällen keine einzige datengeleitete Modellveränderung vorgenommen.

Dafür muss man jedoch den Nachteil in Kauf nehmen, dass die Traits methodenspezifisch sind, d. h. es resultiert keine direkte Information über einen wie auch immer definierten Trait, der *allen Übungen* (Abb. 1) bzw. *allen Ratern* (Abb. 2) gemeinsam ist. Entsprechend sind auch die Methodenfaktoren in dem Sinne spezifisch, dass sie als Residuen bzgl. der linearen Regression einer True-score-Variablen auf die True-score-Variable der Referenzmethode definiert sind. Durch die Betrachtung der Korrelationen der Summen der spezifischen True-score-Variablen (s. die Tabellen 8 und 12) kann man diesen Nachteil jedoch zu einen Teil wieder wettmachen, sofern man bereit ist, den allgemeinen, d.h. nicht-methodenspezifischen Trait als die Summe der betreffenden methodenspezifischen True-score-Variablen zu definieren.

In der Einleitung haben wir verschiedene Fragen aufgeworfen, die auch in der Assessment-Center-Diagnostik von Bedeutung sind. Wir werden nun diese Fragen nacheinander basiert auf den oben dargestellten Analysen beantworten.

3.1 Wie zuverlässig sind die Beurteilungen durch die Rater?

Diese Frage lässt sich auf verschiedene Weise präzisieren. Fragen wir nach der Reliabilität der über die drei Rater gemittelten Urteile, so sind die Antworten in Tabelle 6 zu finden. Die für die hier analysierten Daten wichtigsten Ergebnisse lassen sich wie folgt zusammenfassen: Fast alle Reliabilitäten der über die Rater gemittelten Beurteilungen liegen zwischen .80 und .90. Lediglich die Reliabilität der Beurteilung der *Führungsfähigkeit im Interview* ist mit .38 extrem gering. Auffällig niedriger ist auch die Reliabilität der Beurteilung der *Sozialen Kompetenz im Interview* mit .69. Insbesondere im ersten Fall, also der *Führungsfähigkeit im Interview*, ist zu prüfen, woran dies liegt, ob Verbesserungen möglich sind oder ob diese Beurteilung ganz gestrichen werden sollte.

---

8 Die entsprechenden LISREL-Steuerkarten findet man wieder im Anhang.

Die Frage nach der Zuverlässigkeit der Beurteilungen durch die Rater kann man auch für die über die Übungen aggregierten (gemittelten) Urteile der Rater stellen. Hier zeigt sich, dass im Schnitt ca. 25% der Varianz der manifesten Variablen durch die Raterfaktoren (Methodenfaktoren) erklärt werden (s. Tab. 10). Das sind immerhin mehr als 30% der systematischen Varianz.

3.2 Wie gut können die Eigenschaften der Bewerber gemessen werden?

Oben ging es um die Reliabilität der Beurteilungen der vier Eigenschaften in den drei Übungen gemittelt über die drei Rater, oder aber gemittelt über die drei Übungen. Hier dagegen fragen wir nach der durch die Traits erklärten Varianzanteile der beiden Arten von gemittelten Beurteilungen. Die relevanten Ergebnisse hierzu findet man wiederum in den Tabellen 6 bzw. 10. Dabei zeigen sich nur geringfügige Unterschiede in den durchschnittlichen, durch die Traits erklärten Varianzanteile. Nennenswert ist lediglich der extrem geringe, durch die Führungsfähigkeit erklärte Varianzanteil von 28% im Interview bei der Aggregation über die Rater.

Bei der Aggregation über die Übungen gibt es keine so auffallend niedrige, durch einen Trait erklärten Varianzanteile. Der niedrigste durch einen Trait erklärte Varianzanteil liegt hier bei 46%. Dabei handelt es sich um die *Soziale Kompetenz eingeschätzt durch die Psychologen*. Heir scheinen die Urteile der Psychologen relativ weit von den als Referenzmethode gewählten Urteilen der Stabsoffiziere abzuweichen. Der Psychologenfaktor (Methodenfaktor) erklärt hier 29% der Varianz der über die Übungen gemittelten Urteile.

3.3 Welchen Einfluss haben die unterschiedlichen Übungen und die unterschiedlichen Rater?

Die Beurteilungen der Rater wurden in drei verschiedenen Übungen erhoben: dem Rundgespräch, dem Planspiel und dem Interview. In der ersten oben dargestellten Analyse wurden diese Übungen als „Methoden“ aufgefasst. Welchen Einfluss haben die unterschiedlichen Übungen nun auf die Beurteilungen durch die Rater? Auch hierzu finden wir die Antwort in Tabelle 6. Danach werden im Schnitt ca. 18% der Varianz der über die Rater gemittelten Urteile durch die Methoden *Planspiel* bzw. *Interview* erklärt. Die Beurteilungen sind also zu einem beträchtlichen Teil *spezifisch* für die jeweils betrachtete Übung.

Aus Tabelle 10 können wir entnehmen, dass auch die über die Übungen aggregierten Urteile *spezifisch* für den jeweiligen Rater sind. Im Schnitt werden 27% der Varianz der über die Übungen gemittelten Urteile durch die Methoden *Prüfoffizier* bzw. *Psychologe* erklärt.

3.4 Was sind die praktischen Konsequenzen aus unseren Befunden?

Zwei unserer Befunde legen praktische Konsequenzen nahe. Zum einen ist dies die auffallend geringe Reliabilität (.38) der gemittelten Urteile der Führungsfähigkeit im Interview (s. Tab. 6). Hier wäre nach den Gründen zu suchen und gegebenenfalls zu prüfen, ob etwa ein Interviewleitfaden zu reliableren Urteilen führt.

Ein zweiter auffälliger Befund sind die hohen Korrelationen zwischen den Traits, insbesondere die zwischen den beiden Traits *Sprachliche Ausdrucksfähigkeit* und *Urteilsfähigkeit* (.91 bzw. .94, s. die Tabellen 8 und 12). Auch hier wären Maßnahmen zu prüfen, die möglicherweise zu Beurteilungen führen, die eine größere Differenzierung zwischen den Traits

erlauben, wie z. B. eine Einführung oder Revision von Beurteilungsleitfäden oder ein Beurteilertraining. Möglicherweise wäre auch ein Verzicht auf den vierten Trait ein Gewinn, da dieser nur wenig neue Information bringt und möglicherweise eine reliable und valide Beurteilung der anderen Traits erschwert (kognitive Überforderung der Beurteiler).

3.5 Ausblick

Die hier dargestellten und angewandten Analysen der Daten, die in den Assessment-Center der OPZ erhoben wurden, geben eine recht gute Antwort auf die Frage, wie stark die Übungen einerseits und die Rater andererseits die über die Rater bzw. über die Übungen gemittelten Urteile beeinflussen. Entsprechende Analysen ließen sich prinzipiell auch über die nicht gemittelten Ratings anstellen. Dabei würde man die Ratings direkt als metrische Variablen behandeln. Dies wäre aus Anwendungssicht sicher vertretbar, aber grundsätzlich nicht optimal, da das vorliegende Datenniveau nicht berücksichtigt würde. Grundsätzlich sind entsprechende Analysen auch mit Modellen möglich, die eigens für ordinale Daten, wie die hier vorliegenden Ratings, entwickelt wurden. Zum einen bieten sich hier die mehrdimensionalen logistischen Modelle an, wie sie mit ConQuest (Wu, Adams, & Wilson, 1998) analysiert werden können. Zum anderen kann man aber auch die Probitmodelle verwenden, wie sie in PRELIS/LISREL (Jöreskog & Sörbom, 1999), EQS (Bentler, 2001) und Mplus (Muthén, 1998) realisiert sind. Für beide Modellklassen müsste man aber auch prüfen, ob die hier vorliegenden Stichprobengrößen – bei den nichtaggregierten Urteilen werden diese jeweils wegen fehlender Werte niedriger liegen – für solche anspruchsvolleren Modelle, in denen entsprechend wesentlich mehr Parameter (4 Schwellenparameter für jedes Item) geschätzt werden müssen, ausreichen. Es ist zu befürchten, dass diese Stichprobengröße selbst bei Verwendung von ausgefeilteren Techniken der Behandlung fehlender Werte für eine zuverlässige Analyse mit Modellen für ordinale Daten noch nicht ausreicht. Eine entsprechende Simulationsstudie könnte hier Aufschluss geben.

Wie bereits erwähnt haben die oben verwendeten MTMM-Modelle von Eid (2000) den Vorteil, dass sie auch global identifiziert sind und daher zu stabilen und zuverlässigen Schätzungen der Parameter führen. Außerdem sind die latenten Variablen mathematisch wohldefiniert. Dennoch haben sie den interpretatorischen Nachteil, dass alle Aussagen zu einem gewissen Teil nur auf die jeweils gewählte Referenzmethode bezogen gelten. Es gibt dabei zunächst weder den für alle Methoden (Übungen oder Rater) gemeinsamen Trait, noch den auf einen solchen gemeinsamen Trait bezogenen Methodenfaktor. Entsprechende Trait- und Methodenbegriffe zu entwickeln, die nicht die Probleme der bisher vorliegenden Begriffe und Modelle aufweisen, ist eine wichtige Aufgabe für die Methodenlehre. Gemeinsame Traits als Summen der Übungs- (bzw. Rater-)spezifischen True-score-Variablen zu definieren, ist ein erster Schritt in diese Richtung.

Darüber hinaus wäre es auch für die hier vorliegenden Datenstrukturen wünschenswert, MTMM-Modelle zu entwickelten, mit denen die hier vorliegende dreidimensionale Datenstruktur bearbeitet werden könnte, ohne den Weg über die Aggregation über eine Faktor (Rater bzw. Übungen) zu gehen.

Aus inhaltlicher Sicht wären nun natürlich Validierungsstudien wünschenswert. Wie gut lässt sich aus den hier ermittelten latenten Variablen (Traits) später festgestellte Eignung im Dienst prädizieren? Wie steht es um den Vergleich mit der prognostischen Validität von psychometrischen Testverfahren? Lassen sich Eignungskriterien entwickeln, die nicht auf der Beurteilung durch Vorgesetzte oder auch Untergebene beruhen? Diese Frage ist insofern von Bedeutung, da beide Kriterien die in Assessment Centern erhobenen Variablen begünstigen.

Literatur

Andres, J. & Kleinmann, M. (1993). Die Entwicklung eines Rotationssystems für die Beobachtungssituation im Assessment-Center. *Zeitschrift für Arbeits- und Organisationspsychologie, 37*, 19-25.

Bentler, P.M. (2001). *EQS 6: Structural equations program manual*. Encino, CA: Multivariate Software.

Bycio, P., Alvares, K. M. & Hahn, J. (1987). Situational Specificity in Asssessment Center Ratings: A Confirmatory Factor Analysis. *Journal of Applied Psychology, 72,* 463-474.

Byham, W. C. (1980). Starting an assessment center the correct way. *The Personnel Administrator, Feb. ,* 27-32.

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

Donahue, L. M., Truxillo, D. M., Cornwell, J. M. & Gerrity, M. J. (1997). Assessment Center Construct Validity and Behavioral Checklist: Some Additional Findings. *Journal of Social Behavior and Personality, 12,* 85-108.

Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika*, *65*, 241-261.

Fleenor, J. W. (1996). Constructs And Developmental Assessment Centers: Further Troubling Empirical Findings. *Journal of Business and Psychology, 10,* 319-335.

Gaugler, B. B. & Thornton, G. C. (1989). Number of Assessment Center Dimensions as a Determinant of Assessor Accuracy. *Journal of Applied Psychology, 74,* 611-618.

Grayson, D., & Marsh, H. W. (1994). Identification with deficient rank loading matrices in confirmatory factor analysis: Multitrait-multimethod models. *Psychometrika, 59,* 121-134

Guzzo, R. A., Wagner, D. B., Maguire, E., Herr, B. & Hawley, C. (1986). Implizit Theories and the Evaluation of Group Process and Performance. *Organizational Behavior and Human Decision Processes, 37,* 279-295.

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

James, L. R., Demaree, R. G. & Wolf, G. (1984). Estimating Within-Group Interrater Reliability With and Without Response Bias. *Journal of Applied Psychology, 69,* 85-98.

Jeserich, W. (1981). *Mitarbeiter auswählen und fördern. Handbuch der Weiterbildung für die Praxis in Wirtschaft und Verwaltung.* München: Hanser.

Jöreskog, K. G. & Sörbom, D. (1993). *LISREL VIII. User's reference guide*. Chicago: Scientific Software International.

Jöreskog, K. G. & Sörbom, D. (1999). *LISREL (Version 8.30) and PRELIS (Version 2.30) [computer software]*. Chicago: Scientific Software International.

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

Kleinmann. M. & Köller, O. (1997). Construct Validity of Assessment Centers: Appropriate Use of Confirmatory Factor Analysis and Suitable Construction Principles. *Journal of Social Behavior and Personality, 12,* 65-84.

Kudisch, J. D., Ladd, R. T. & Dobbins, G. H. (1997). New Evidence on the Construct Validity of Diagnostic Assessment Centers: The Findings May Not Be So Troubling After All. *Journal of Social Behavior and Personality, 12,* 129-144.

Kuptsch, C., Kleinmann, M. & Köller, O. (1998). The Chameleon Effect in Assessment Centers: The Influence of Cross-Situational Behavioral Consistency on the Convergent Validity of Assessment Centers. *Journal of Social Behavior and Personality, 13,* 103-116.

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

Lowry, P. E. (1993). The Assessment Center: An Examination of the Effects of Assessor Characteristics on Assessor Scores. *Public Personnel Management, 22*, 487-501.

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

Muthén, L. K. & Muthén, B. O. (1998) *Mplus: User's Guide*. Los Angeles, CA: Muthén & Muthén.

Neidig, R. D. & Neidig, P. J. (1984). Multiple Assessment Center Exercises and Job Relatedness. *Journal of Applied Psychology, 69,* 182-186.

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

Sackett, P. R. & Dreher, G. F. (1982). Constructs and Assessment Center Dimensions: Some Troubling Empirical Findings. *Journal of Applied Psychology, 67,* 401-410.

Sackett, P. R. & Dreher, G. F. (1984). Situation Specificity of Behavior and Assessment Center Validation Strategies: A Rejoinder to Neidig and Neidig. *Journal of Applied Psychology, 69,* 187-190.

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

Schmitt, N. (1993). Group Composition, Gender, and Race Effects on Assessment Center Ratings. In H. Schuler, J. L. Farr et al. (eds.), *Personnel Selection and Assessment: Individual and Organizational Perspectives* (315-332). Hillsdale, NJ: Erlbaum.

Schneider, J. R. & Schmitt, N. (1992). An Exercise Design Approach to Understanding Assessment Center Dimension and Exercise Constructs. *Journal of Applied Psychology, 77*, 32-41.

Steyer, R. (1995). Das MTMM-Modell ist nicht identifiziert. *Newsletter der Fachgruppe Methoden, 2*, 5-6.

Turnage, J. J. & Muchinsky, P. M. (1982). Transsituational Variability in Human Performance within Assessment Centers. *Organizational Behavior And Human Performance, 30,* 174-200.

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

Wothke, W. (1996). Models for multitrait-multimethod matrix analysis. In G. A. Marcoulides & R. E. Schumacker (eds.). *Advanced structural equation modeling: issues and techniques* (pp. 7-56). Mahwah, NJ: Erlbaum.

Wu, M. L., Adams, R. J., & Wilson (1998). *ACER ConQuest*. *Generalised item response modelling software*. Melbourne: The Australian Council for Educational Research Ltd.

Tabelle 1

Labels der 12 manifesten Variable (über die Rater gemittelten) Beurteilungen der vier Eigenschaften in den drei Übungen

LeaDis	Führungsfähigkeit im Rundgespräch
LeaGam	Führungsfähigkeit im Planspiel
LeaInt	Führungsfähigkeit im Interview
SocDis	Soziale Kompetenz im Rundgespräch
SocGam	Soziale Kompetenz im Planspiel
SocInt	Soziale Kompetenz im Interview
LanDis	Sprachliche Ausdrucksfähigkeit im Rundgespräch
LanGam	Sprachliche Ausdrucksfähigkeit im Planspiel
LanInt	Sprachliche Ausdrucksfähigkeit im Interview
JudDis	Urteilsfähigkeit im Rundgespräch
JudGam	Urteilsfähigkeit im Planspiel
JudInt	Urteilsfähigkeit im Interview

Tabelle 2

Labels der 12 manifesten Variable (über die Übungen gemittelten) Beurteilungen der vier Eigenschaften, die durch die drei Rater beurteilt wurden

LeaSta	Führungsfähigkeit eingeschätzt durch den Stabsoffizier
LeaOff	Führungsfähigkeit eingeschätzt durch den Prüfoffizier
LeaPsy	Führungsfähigkeit eingeschätzt durch den Psychologen
SocSta	Soziale Kompetenz eingeschätzt durch den Stabsoffizier
SocOff	Soziale Kompetenz eingeschätzt durch den Prüfoffizier
SocPsy	Soziale Kompetenz eingeschätzt durch den Psychologen
LanSta	Sprachliche Ausdrucksfähigkeit eingeschätzt durch den Stabsoffizier
LanOff	Sprachliche Ausdrucksfähigkeit eingeschätzt durch den Prüfoffizier
LanPsy	Sprachliche Ausdrucksfähigkeit eingeschätzt durch den Psychologen
JudSta	Urteilsfähigkeit eingeschätzt durch den Stabsoffizier
JudOff	Urteilsfähigkeit eingeschätzt durch den Prüfoffizier
JudPsy	Urteilsfähigkeit eingeschätzt durch den Psychologen

Tabelle 3

Korrelationsmatrix, Standardabweichungen und Mittelwerte der 12 über die Rater gemittelten Urteile

	LeaDis	LeaGam	LeaInt	SocDis	SocGam	SocInt	LanDis	LanGam	LanInt	JudDis	JudGam	JudInt
LeaDis	1.000											
LeaGam	0.744	1.000										
LeaInt	0.494	0.451	1.000									
SocDis	0.696	0.596	0.306	1.000								
SocGam	0.617	0.718	0.332	0.735	1.000							
SocInt	0.524	0.434	0.424	0.597	0.553	1.000						
LanDis	0.714	0.608	0.416	0.663	0.597	0.547	1.000					
LanGam	0.685	0.707	0.449	0.665	0.717	0.558	0.839	1.000				
LanInt	0.561	0.510	0.459	0.543	0.511	0.652	0.765	0.742	1.000			
JudDis	0.693	0.584	0.366	0.671	0.597	0.589	0.791	0.724	0.648	1.000		
JudGam	0.659	0.753	0.399	0.652	0.720	0.525	0.713	0.793	0.614	0.730	1.000	
JudInt	0.521	0.449	0.477	0.477	0.475	0.689	0.637	0.592	0.775	0.667	0.583	1.000
SD	0.757	0.880	1.220	0.596	0.649	0.743	0.760	0.713	0.827	0.811	0.735	1.015
M	4.025	4.038	4.065	3.800	3.786	3.677	3.952	4.046	3.988	4.032	4.043	4.170

Tabelle 4

Korrelationsmatrix, Standardabweichungen und Mittelwerte der 12 über die drei Übungen gemittelten Urteile

	LeaSta	LeaOff	LeaPsy	SocSta	SocOff	SocPsy	LanSta	LanOff	LanPsy	JudSta	JudOff	JudPsy
LeaSta	1.000											
LeaOff	0.670	1.000										
LeaPsy	0.644	0.662	1.000									
SocSta	0.638	0.434	0.501	1.000								
SocOff	0.576	0.696	0.568	0.575	1.000							
SocPsy	0.476	0.461	0.692	0.538	0.571	1.000						
LanSta	0.701	0.524	0.557	0.694	0.550	0.539	1.000					
LanOff	0.579	0.684	0.537	0.456	0.673	0.484	0.657	1.000				
LanPsy	0.496	0.507	0.612	0.475	0.544	0.673	0.676	0.665	1.000			
JudSta	0.714	0.510	0.521	0.682	0.574	0.538	0.810	0.634	0.589	1.000		
JudOff	0.609	0.697	0.560	0.524	0.686	0.561	0.647	0.755	0.617	0.685	1.000	
JudPsy	0.479	0.489	0.658	0.511	0.525	0.709	0.624	0.567	0.788	0.629	0.653	1.000
SD	0.758	0.883	0.818	0.628	0.643	0.640	0.777	0.789	0.701	0.861	0.864	0.840
M	4.064	4.116	4.032	3.838	3.863	3.684	4.012	4.116	3.943	4.193	4.159	4.039

Tabelle 5

Korrelationsmatrix der Rundgespräch-spezifischen Traits

	Leader	Social	Language	Judgment
Leader	1.00			
Social	0.82	1.00		
Language	0.81	0.81	1.00	
Judgment	0.83	0.85	0.91	1.00

Tabelle 6

Durch Trait bzw. Methodenfaktoren erklärte Varianzanteile und Reliabilitäten der manifesten Variablen für die in Tabelle 3 angegebenen Daten

Varianzanteil erklärt durch	LeaDis	LeaGam	LeaInt	SocDis	SocGam	SocInt	LanDis	LanGam	LanInt	JudDis	JudGam	JudInt
Traits	.88	.62	.28	.80	.64	.49	.89	.80	.64	.82	.67	.51
Methodenfaktoren	--	.22	.10	--	.17	.20	--	.08	.19	--	.17	.34
Reliabilität	.88	.84	.38	.80	.81	.69	.89	.88	.83	.82	.84	.85

Tabelle 7

Korrelationsmatrix der True-score-Variablen der 12 über die Rater gemittelten Urteile

	Lea Dis	LeaGam	LeaInt	SocDis	SocGam	SocInt	LanDis	LanGam	LanInt	JudDis	JudGam	JudInt
LeaDis	1.00											
LeaGam	0.86	1.00										
LeaInt	0.86	0.75	1.00									
SocDis	0.82	0.70	0.70	1.00								
SocGam	0.72	0.86	0.63	0.89	1.00							
SocInt	0.69	0.60	0.87	0.84	0.76	1.00						
LanDis	0.81	0.69	0.69	0.81	0.72	0.68	1.00					
LanGam	0.77	0.82	0.67	0.77	0.82	0.65	0.95	1.00				
LanInt	0.71	0.62	0.85	0.71	0.64	0.85	0.88	0.85	1.00			
JudDis	0.83	0.71	0.71	0.85	0.75	0.71	0.91	0.87	0.80	1.00		
JudGam	0.74	0.87	0.65	0.76	0.88	0.65	0.82	0.92	0.73	0.89	1.00	
JudInt	0.64	0.57	0.88	0.66	0.60	0.89	0.71	0.68	0.92	0.77	0.71	1.00

Tabelle 8

Korrelationsmatrix der Summen der True-score-Variablen
jeweils über die drei übungsspezifischen True-score-Variablen

	Leader	Social	Language	Judgment
Leader	1.00			
Social	.83	1.00		
Language	.81	.82	1.00	
Judgment	.84	.86	.91	1.00

Tabelle 9
Korrelationsmatrix der Stabsoffizier-spezifischen Traits

	Leader	Social	Language	Judgment
Leader	1.00			
Social	0.83	1.00		
Language	0.81	0.84	1.00	
Judgment	0.81	0.88	0.94	1.00

Tabelle 10

Durch Trait bzw. Methodenfaktoren erklärte Varianzanteile und Reliabilitäten der manifesten Variablen für die in Tabelle 4 angegebenen Daten

Varianzanteil erklärt durch	LeaSta	LeaOff	LeaPsy	SocSta	SocOff	SocPsy	LanSta	LanOff	LanPsy	JudSta	JudOff	JudPsy
Traits	0.84	0.53	0.55	0.71	0.52	0.46	0.87	0.54	0.53	0.85	0.57	0.51
Methodenfaktoren	--	0.30	0.23	--	0.25	0.29	--	0.25	0.28	--	0.23	0.33
Reliabilität	0.84	0.83	0.78	0.71	0.77	0.75	0.87	0.79	0.81	0.85	0.80	0.84

Tabelle 11

Korrelationsmatrix der True-score-Variablen der 12 über die Übungen gemittelten Urteile

	LeaSta	LeaOff	LeaPsy	SocSta	SocOff	SocPsy	LanSta	LanOff	LanPsy	JudSta	JudOff	JudPsy
LeaSta	1.00											
LeaOff	0.80	1.00										
LeaPsy	0.84	0.82	1.00									
SocSta	0.83	0.67	0.70	1.00								
SocOff	0.69	0.89	0.71	0.82	1.00							
SocPsy	0.65	0.69	0.89	0.78	0.80	1.00						
LanSta	0.81	0.64	0.68	0.84	0.69	0.66	1.00					
LanOff	0.67	0.87	0.69	0.70	0.89	0.70	0.83	1.00				
LanPsy	0.65	0.68	0.87	0.68	0.71	0.90	0.81	0.82	1.00			
JudSta	0.81	0.65	0.68	0.88	0.72	0.69	0.94	0.78	0.76	1.00		
JudOff	0.69	0.87	0.71	0.74	0.92	0.73	0.79	0.96	0.78	0.85	1.00	
JudPsy	0.63	0.67	0.87	0.68	0.72	0.93	0.73	0.76	0.96	0.78	0.81	1.00

Tabelle 12

Korrelationsmatrix der Summen der True-score-Variablen
jeweils über die drei Rater-spezifischen True-score-Variablen

	Leader	Social	Language	Judgment
Leader	1.00			
Social	0.86	1.00		
Language	0.83	0.86	1.00	
Judgment	0.84	0.90	0.94	1.00

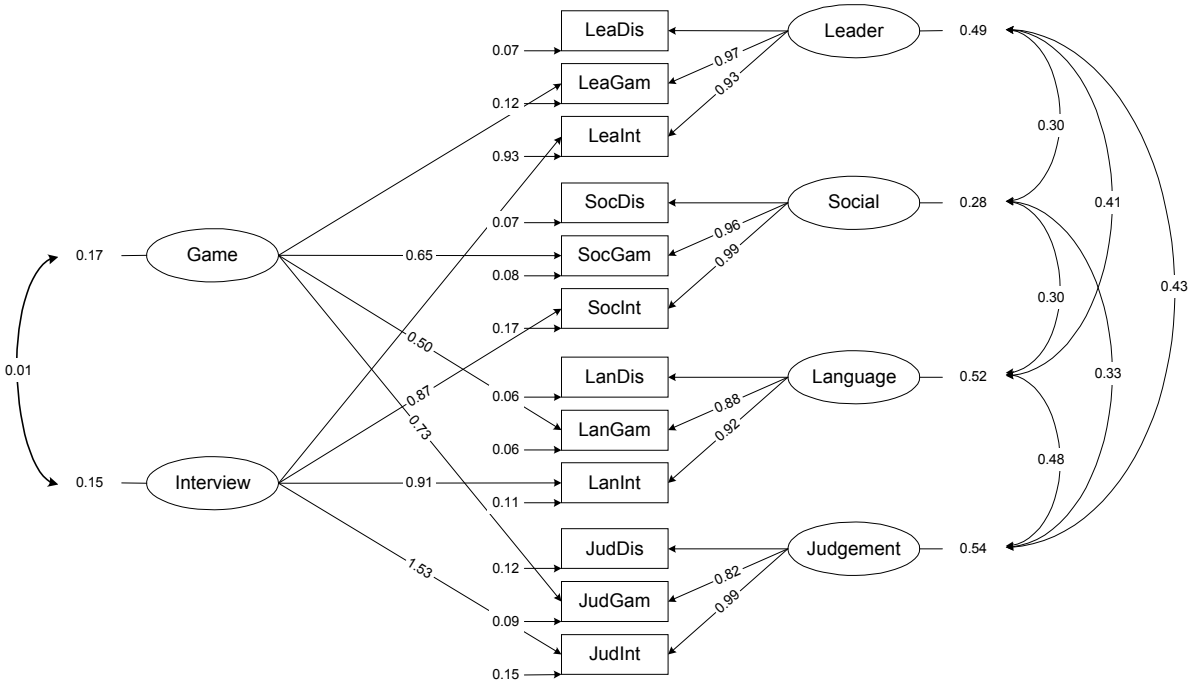

Abbildung 1. Pfaddiagramm des MTMM-Modells für die über die Rater gemittelten Beurteilungen. Der χ^2-Wert beträgt 54.83 bei 39 Freiheitsgraden. Der zugehörige *p*-Wert ist .048 und der RMSEA-Wert ist .037. Die eingetragenen Koeffizienten beziehen sich auf die nichtstandardisierte Lösung.

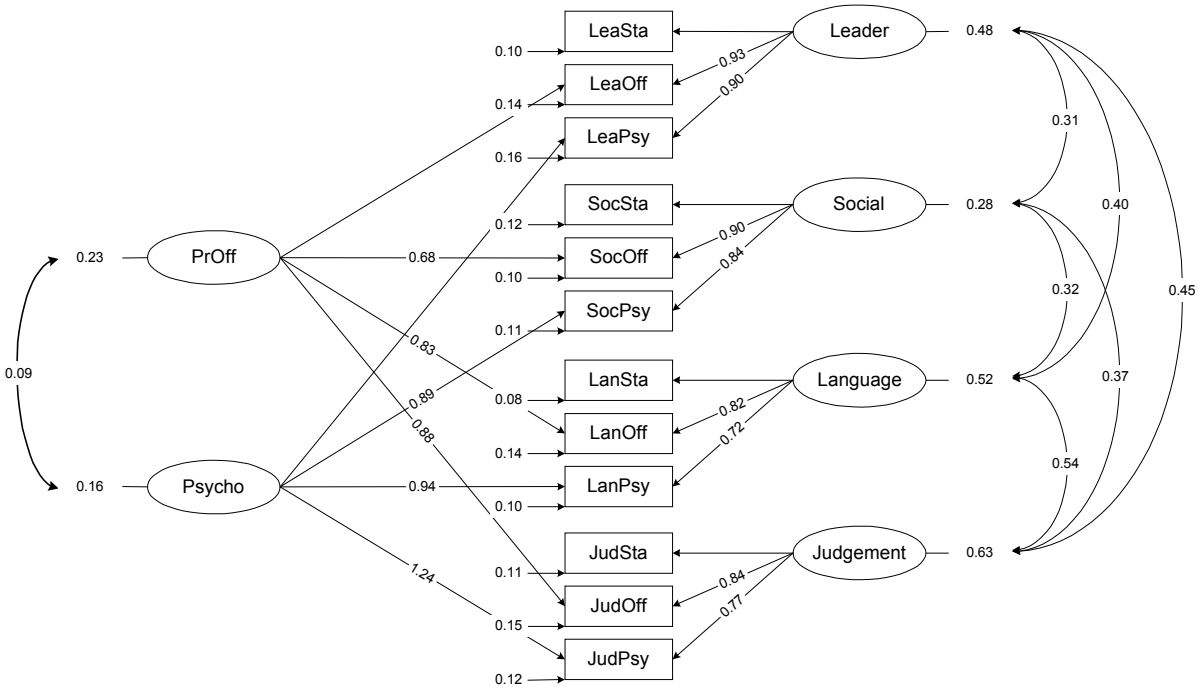

Abbildung 2. Pfaddiagramm des MTMM-Modells für die über die Übungen gemittelten Beurteilungen. Der χ^2-Wert beträgt 53.15 bei 39 Freiheitsgraden. Der zugehörige *p*-Wert ist .065 und der RMSEA-Wert ist .041. Die eingetragenen Koeffizienten beziehen sich auf die nichtstandardisierte Lösung.

Anhang

LISREL –Steuerkarten für das Modell der Abbildung 1

```
TI: 4 Traits mit 3 Methoden (Rater) aggregiert ueber drei Rater.
Die Ys haben freie Ladungen auf den Etas und freie Ladungen auf den
Methodenfaktoren. Die Fehlervarianzen haben keine Restriktionen.
DA     NI = 12  NO = 296  MA = CM
CM     FI = 4T3PA3U
! Die hier benötigte Kovarianzmatrix kann aus den in Tab. 3 angegebenen Daten
! berechnet werden
MO     NX=12  NK=6  LX=FU,FI  TD=SY,FI  PH=SY,FI
LK
Leader Social  Langua  Judgme  Game  IntVie
FR     PH(1,1)   PH(2,2)   PH(3,3)   PH(4,4)
FR     PH(2,1)   PH(3,1)   PH(4,1)   PH(3,2)   PH(4,2)
FR     PH(4,3)
FR     PH(5,5)   PH(6,6)  PH(6,5)
LA
LeaDis LeaGam LeaInt SocDis SocGam SocInt LanDis LanGam LanInt JudDis JudGam JudInt
VA 1.0 LX(1,1)  LX(4,2)  LX(7,3)  LX(10,4)
VA 1.0 LX(2,5)  LX(3,6)
FR     LX(2,1) LX(3,1) LX(5,2) LX(6,2) LX(8,3) LX(9,3) LX(11,4) LX(12,4)
FR     LX(5,5)  LX(8,5) LX(11,5)
FR     LX(6,6)  LX(9,6) LX(12,6)
ST 1.0 LX(2,1) LX(3,1) LX(5,2) LX(6,2) LX(8,3) LX(9,3) LX(11,4) LX(12,4)
ST 1.0 LX(5,5)  LX(8,5) LX(11,5)
ST 1.0 LX(6,6)  LX(9,6) LX(12,6)
FR     TD(1,1) TD(2,2) TD(3,3) TD(4,4) TD(5,5)
FR     TD(6,6) TD(7,7) TD(8,8) TD(9,9) TD(10,10)
FR     TD(11,11) TD(12,12)
PATH DIAGRAM
OUTPUT WP ND=3 SI=lisout.mat AD=OFF MI SE SC
```

LISREL –Steuerkarten für das Modell der Abbildung 1 mit expliziter Modellierung der True-score-Variablen und der Summen der True-score-Variablen

```
TI: 4 Traits mit 3 Methoden (Rater) aggregiert ueber drei Rater.
Die Ys haben freie Ladungen auf den Etas und freie Ladungen auf den
Methodenfaktoren. Die Fehlervarianzen haben keine Restriktionen.
True-score-Variablen werden explizit modelliert und die Summen der True-score-
Variablen über die jeweils drei Uebungen werden eingeführt

DA      NI = 12  NO = 296  MA = CM
CM      FI = 4T3PA3U
! Die hier benötigte Kovarianzmatrix kann aus den in Tab. 3 angegebenen Daten
! berechnet werden
MO NY=12 NE=16 NK=6 LY=FU,FI TE=SY,FI PS=SY,FI PH=SY,FI GA=FU,FI BE=FU,FI
LE
TY1 TY2 TY3 TY4 TY5 TY6 TY7 TY8 TY9 TY10 TY11 TY12 SUM1 SUM2 SUM3 SUM4
LK
Leader Social  Langua  Judgme  Game  IntVie
LA
LeaDis LeaGam LeaInt SocDis SocGam SocInt LanDis LanGam LanInt JudDis JudGam JudInt
VA 1.0 LY(1,1) LY(2,2) LY(3,3) LY(4,4) LY(5,5) LY(6,6)
VA 1.0 LY(7,7) LY(8,8) LY(9,9) LY(10,10) LY(11,11) LY(12,12)
FR     PH(1,1)  PH(2,2)  PH(3,3)  PH(4,4)  PH(5,5)  PH(6,6)
FR     PH(2,1)  PH(3,1)  PH(4,1)  PH(3,2)  PH(4,2)  PH(4,3)
FR     PH(6,5)
VA 1.0 GA(1,1) GA(4,2) GA(7,3) GA(10,4)

VA 1.0 GA(2,5) GA(3,6)
FR     GA(2,1) GA(3,1) GA(5,2) GA(6,2) GA(8,3) GA(9,3) GA(11,4) GA(12,4)
FR     GA(5,5) GA(8,5) GA(11,5) GA(6,6) GA(9,6) GA(12,6)
```

```
VA 1.0 BE(13,1)  BE(13,2)  BE(13,3)
VA 1.0 BE(14,4)  BE(14,5)  BE(14,6)
VA 1.0 BE(15,7)  BE(15,8)  BE(15,9)
VA 1.0 BE(16,10) BE(16,11) BE(16,12)
FREE   TE( 1, 1) TE( 2, 2) TE( 3, 3) TE( 4, 4) TE( 5, 5)
FREE   TE( 6, 6) TE( 7, 7) TE( 8, 8) TE( 9, 9) TE(10,10)
FREE   TE(11,11) TE(12,12)
PATH DIAGRAM
OUTPUT  ND=3 WP SI=lisout.mat AD=OFF MI SE SC
```

LISREL –Steuerkarten für das Modell der Abbildung 2

```
TI: 4 Traits mit 3 Methoden (Pruefarten) aggregiert ueber drei Uebungen.
Die Ys haben freie Ladungen auf den Etas und freie Ladungen auf den
Methodenfaktoren. Die Fehlervarianzen haben keine Restriktionen.

DA     NI = 12  NO = 216  MA = CM
CM     FI = 4T3UA3P
! Die hier benötigte Kovarianzmatrix kann aus den in Tab. 4 angegebenen Daten
! berechnet werden
MO     NX=12  NK=6  LX=FU,FI  TD=SY,FI  PH=SY,FI
LK
Leader Social  Langua  Judgme  PrOff  Psycho
FR     PH(1,1)   PH(2,2)   PH(3,3)   PH(4,4)
FR     PH(2,1)   PH(3,1)   PH(4,1)   PH(3,2)   PH(4,2)
FR     PH(4,3)
FR     PH(5,5)   PH(6,6)  PH(6,5)
LA
LeaSta LeaOff LeaPsy SocSta SocOff SocPsy LanSta LanOff LanPsy JudSt JudOff
JudPsy
VA 1.0 LX(1,1)  LX(4,2)  LX(7,3)  LX(10,4)
VA 1.0 LX(2,5)  LX(3,6)
FR     LX(2,1) LX(3,1) LX(5,2) LX(6,2) LX(8,3) LX(9,3) LX(11,4) LX(12,4)
FR     LX(5,5)  LX(8,5) LX(11,5)
FR     LX(6,6)  LX(9,6) LX(12,6)
ST 1.0 LX(2,1) LX(3,1) LX(5,2) LX(6,2) LX(8,3) LX(9,3) LX(11,4) LX(12,4)
ST 1.0 LX(5,5) LX(8,5) LX(11,5)
ST 1.0 LX(6,6) LX(9,6) LX(12,6)
FR     TD(1,1) TD(2,2) TD(3,3) TD(4,4) TD(5,5)
FR     TD(6,6) TD(7,7) TD(8,8) TD(9,9) TD(10,10)
FR     TD(11,11) TD(12,12)
PATH DIAGRAM
OUTPUT AL WP ND=3 SI=lisout.mat AD=OFF MI SE SC
```

LISREL –Steuerkarten für das Modell der Abbildung 2 mit expliziter Modellierung der True-score-Variablen und der Summen der True-score-Variablen

```
TI: 4 Traits mit 3 Methoden (Rater) aggregiert ueber drei Uebungen.
Die Ys haben freie Ladungen auf den Etas und freie Ladungen auf den
Methodenfaktoren. Die Fehlervarianzen haben keine Restriktionen.
True-score-Variablen werden explizit modelliert und die Summen der True-score-
Variablen über die jeweils drei Rater werden eingeführt

DA      NI = 12  NO = 216  MA = CM
CM      FI = 4T3UA3P
! Die hier benötigte Kovarianzmatrix kann aus den in Tab. 4 angegebenen Daten
! berechnet werden
MO NY=12 NE=16 NK=6 LY=FU,FI TE=SY,FI PS=SY,FI PH=SY,FI GA=FU,FI BE=FU,FI
LE
TY1 TY2 TY3 TY4 TY5 TY6 TY7 TY8 TY9 TY10 TY11 TY12 SUM1 SUM2 SUM3 SUM4
LK
Leader Social  Langua  Judgme  PrOff  Psycho
LA
```

```
LeaSta LeaOff LeaPsy SocSta SocOff SocPsy LanSta LanOff LanPsy JudSt JudOff
JudPsy
VA 1.0 LY(1,1) LY(2,2) LY(3,3) LY(4,4) LY(5,5) LY(6,6)
VA 1.0 LY(7,7) LY(8,8) LY(9,9) LY(10,10) LY(11,11) LY(12,12)
FR PH(1,1)  PH(2,2)  PH(3,3)  PH(4,4)  PH(5,5)  PH(6,6)
FR PH(2,1)  PH(3,1)  PH(4,1)  PH(3,2)  PH(4,2)  PH(4,3)
FR PH(6,5)
VA 1.0  GA(1,1) GA(4,2) GA(7,3) GA(10,4)
VA 1.0  GA(2,5) GA(3,6)
FR GA(2,1) GA(3,1) GA(5,2)  GA(6,2) GA(8,3) GA(9,3)  GA(11,4) GA(12,4)
FR GA(5,5) GA(8,5) GA(11,5) GA(6,6) GA(9,6) GA(12,6)
VA 1.0 BE(13,1) BE(13,2) BE(13,3)
VA 1.0 BE(14,4) BE(14,5) BE(14,6)
VA 1.0 BE(15,7) BE(15,8) BE(15,9)
VA 1.0 BE(16,10) BE(16,11) BE(16,12)
FR TE(1,1) TE(2,2) TE(3,3) TE(4,4) TE(5,5)
FR TE(6,6) TE(7,7) TE(8,8) TE(9,9) TE(10,10)
FR TE(11,11) TE(12,12)
PATH DIAGRAM
OUTPUT  ND=3 WP SI=lisout.mat AD=OFF MI SE SC
```