Evaluation von kompetenz- und übungsbasierten Assessment Centern in einem deutschen Unternehmen

Inaugural-Dissertation

zur Erlangung des Doktorgrades

der Philosophie an der Ludwig-Maximilians-Universität

München

vorgelegt von
Anita Maria Kittmann
aus München
2015

Erstgutachter: Prof. Dr. Markus Bühner
Zweitgutachter: Prof. Dr. Moritz Heene
Tag der mündlichen Prüfung: 06. Juli 2015

Inhaltsverzeichnis

Zusammenfassung IV

Einleitung 1

Stand der Forschung 4

Begriffsklärung und historische Entwicklung von Assessment Centern 4

Diagnostische Einordnung von Assessment Centern 6

Hauptgütekriterien diagnostischer Verfahren 9

Objektivität 9

Reliabilität. 11

Validität 16

Das Multimodale Auswahlverfahren 32

Die prädiktive Validität von ausgewählten Personalselektionsinstrumenten einzeln und in Kombination miteinander betrachtet 33

Intelligenz – Definition, Messinstrumente und Unterfacetten. 39

Das strukturierte Interview 46

Normen und Anforderungen an diagnostische Verfahren 47

Fragestellung und Methodik der empirischen Studie 50

Zielsetzung der vorliegenden Untersuchung und Forschungslücke 50

Unterschiede der eingesetzten Auswahlverfahren und Hypothesen 52

Eingesetzte Prüfverfahren und statistische Analysen 58

Datenerhebung und Stichprobe 62

Das Untersuchungsdesign der empirischen Studie 64

Die Zielgruppe 64

Die beiden klassischen Assessment Center 65

Das Trainee-Assessment Center. 66

Das Stabs-Assessment Center 68

Das Multimodale Auswahlverfahren 71

Zielsetzungen und Rahmenbedingungen. 71

Die Übungen des neuen Auswahlverfahrens 74

Die Vorgesetztenbeurteilung als Zielkriterium 77

Die Ergebnisse der empirischen Untersuchung 80

Die Konstruktvalidität des Trainee-Assessment Centers (Hypothese 1) 80

Die Konstruktvalidität des Stabs-Assessment Centers (Hypothese 2) 83

Trennschärfe der eingesetzten Assessment Center Übungen im Multimodalen Auswahlverfahren (Hypothese 3) 86

Überprüfung der Interrater-Korrelation in den eingesetzten Auswahlverfahren (Hypothese 4) 88

Überprüfung der prädiktiven Validität der eingesetzten Auswahlverfahren (Hypothese 5) 89

Gesamtdiskussion 99

Zusammenfassung und Interpretation der Ergebnisse 99

Limitationen der empirischen Studie und Implikationen für die Praxis 110

Literaturverzeichnis CXIV

Anhang..... CXXII

Anhang A: Korrelationskoeffizienten zur Konstruktvalidität des Trainee-Assessment Centers..... CXXII

Anhang B: Korrelationskoeffizienten zur Konstruktvalidität des Stabs-Assessment Centers..... CXXX

Anhang C: Interrater-Korrelationskoeffizienten in den eingesetzten Auswahlverfahren CXXXVI

Anhang D: Tabellen zur prädiktiven Validität der eingesetzten Auswahlverfahren CXXXIX

Zusammenfassung

In der hier vorliegenden empirischen Studie werden in einem deutschen Unternehmen zwei klassische Assessment Center und ein Multimodales Auswahlverfahren im Hinblick auf die Hauptgütekriterien diagnostischer Verfahren untersucht. Der gravierende Unterschied des Multimodalen Auswahlverfahrens gegenüber den beiden klassischen Assessment Centern ist neben dem Methodenmix das Beurteilungsprinzip.

Die Ergebnisse der Untersuchung haben deutlich gemacht, dass in den beiden klassischen Assessment Centern, hingegen der zugrundeliegenden Annahme, eine übungsbezogene statt eine dimensionsbezogene Beurteilung durch die Beobachter erfolgte. Durch die Veränderung des Beurteilungsprinzips und weitere Überarbeitungen der Übungen im Multimodalen Auswahlverfahren konnte eine höhere Trennschärfe erzielt werden. Eine Steigerung der Interrater-Korrelation im Multimodalen Auswahlverfahren gegenüber den beiden Assessment Centern wurde nicht erreicht. Jedoch konnte die Prognose des beruflichen Erfolgs der Kandidaten und somit die Zuverlässigkeit der Einstellungsentscheidung verbessert werden.

Einleitung

„Der Erfolg eines Unternehmens steht und fällt mit seinen Mitarbeitern“ (Hufnagl, 2001, S. 9). Die Rekrutierung von Mitarbeitern, insbesondere von Führungskräften oder auch potentiellen Nachwuchsführungskräften, verursacht hohe monetäre Investitionen. Zusätzlich sind auch nicht-monetäre Aspekte zu berücksichtigen. Nachlässigkeiten bei der Besetzung dieser Positionen rufen neben dem Gehalt hohe Kosten wie Anwerbungs-, Einarbeitungs- und Entlassungskosten hervor. Bei Trainees kommen darüber hinaus noch die Ausbildungskosten hinzu. Fehlentscheidungen führen neben den genannten finanziellen Einbußen auch zu einer Verunsicherung unter den Mitarbeitern, wenn sich beispielsweise herauskristallisiert, dass ein neu gewonnener Entscheidungsträger nicht zur Unternehmenskultur passt (Obermann, 2013, S. 4).

Um Fehlbesetzungen möglichst gering zu halten, gilt es für Unternehmen, möglichst valide Personalauswahlinstrumente einzusetzen (Hufnagl, 2001, S. 9).

Untersuchungen über Personalauswahlverfahren als Prädiktoren für späteren Berufserfolg und Lernen am Arbeitsplatz gibt es seit dem ersten Jahrzehnt des 20. Jahrhunderts (Schmidt & Hunter, 1998, S. 263). Gängige und untersuchte Auswahlinstrumente sind beispielsweise Intelligenztests, Integritätstests, strukturierte oder unstrukturierte Interviews, Arbeitsproben sowie Assessment Center (Schmidt & Hunter, 1998, S. 265).

Eines der nach wie vor am häufigsten eingesetzten Instrumente ist das Assessment Center. Der Arbeitskreis Assessment Center e.V. führte im Jahr 2008 eine Studie zum Einsatz von Personalauswahlverfahren in den DAX 100 Unternehmen durch. Das Ergebnis ist gegenüber der Vorgängerbefragung aus dem Jahr 2001 in etwa konstant geblieben. Der Einsatz von Assessment Center liegt bei den sehr mitarbeiterstarken Konzernen (mehr als 20.000 Mitarbeiter) nach wie vor bei etwa 90 Prozent (Obermann, 2013, S. 19).

In der Literatur wird ein immer größerer Einsatz von Multimodalen Auswahlverfahren – eine Kombination vielfältiger Methoden – gefordert. Dies findet auch zunehmend Berücksichtigung in der Praxis (Obermann, 2013, S. 21f; Schuler, 2007, S. 18f). In den letzten Jahren geht der Trend weg von klassischen Assessment Center Übungen, wie einer Präsentation, Rollenübung, Interview, Fallstudie, Gruppendiskussion und Postkorb, hin zu einem vermehrten Einsatz von Tests und Fragebögen, die kognitive Fähigkeiten oder Persönlichkeitsmerkmale erfassen (Obermann, 2013, S. 21f).

Die vorliegende Arbeit betrachtet den angesprochenen Trend vom klassischen Assessment Center hin zu einem Multimodalen Auswahlverfahren in der Unternehmenspraxis. Die ersten Kapitel beschäftigen sich mit der historischen Entwicklung und diagnostischen Einordnung von Assessment Centern. Daran knüpft eine genauere Betrachtung der bereits vorhandenen Forschungen zu den Hauptgütekriterien diagnostischer Verfahren an. Die wichtigsten Ergebnisse dieser Studien werden zusammengefasst. Im Anschluss daran werden mögliche Bausteine eines Multimodalen Auswahlverfahrens, insbesondere im Hinblick auf die prognostische Validität einzelner und in Kombination miteinander eingesetzter Auswahlinstrumente, betrachtet. Aufgrund der hohen Bedeutung von Intelligenz für beruflichen Erfolg liegt hier ein Fokus (Salgado, Anderson, Moscoso, Bertua & De Fruyt, 2003; Schmidt & Hunter, 1998).

Im empirischen Teil dieser Arbeit wird der Auswahlprozess von Referenten und Trainees in einem führenden deutschen Unternehmen untersucht. Bei den Zielpositionen handelt es sich um für das Unternehmen bedeutende Stellen in Stabsbereichen wie Personal, Recht, Controlling oder Mathematik. Diese Zielgruppe bildet die künftigen Nachwuchsführungskräfte bzw. Fachexperten des Unternehmens, weshalb der Selektionsprozess von entscheidender Bedeutung für das Unternehmen ist. Betrachtet wird zuerst ein klassisches Assessment Center – getrennt für die oben genannten Zielgruppen.

Im Anschluss daran wird das für beide Bewerbergruppen neu konzipierte Multimodale Auswahlverfahren untersucht. Der gravierende Unterschied des neuen Verfahrens gegenüber den beiden zuvor eingesetzten Assessment Centern ist, neben dem Methodenmix, das Beurteilungsprinzip. Den beiden klassischen Assessment Centern liegt eine dimensionsbezogene Beurteilung zu Grunde, dem Multimodalen Verfahren eine übungsbezogene. Der Unterschied in der Beurteilungsart ist ein entscheidender Einflussfaktor auf die im empirischen Teil untersuchten Hauptgütekriterien der drei eingesetzten Auswahlverfahren.

Letztendlich soll das Zielkriterium – die Mitarbeiterbeurteilung durch den Vorgesetzten nach einer Mindestverweildauer von sechs Monaten – Aufschluss über die prädiktive Validität der eingesetzten Instrumente geben.

Stand der Forschung

Dieses Kapitel startet mit einer Begriffsklärung und einem Einblick in die historische Entwicklung von Assessment Centern. Daran knüpft die diagnostische Einordnung von Assessment Centern. Im Anschluss werden die Hauptgütekriterien diagnostischer Verfahren, die *Objektivität, Reliabilität und Validität,* erläutert. Ausgewählte Studien zur Konstruktvalidität und der prädiktiven Validität von Assessment Centern werden genauer beschrieben. In einem nächsten Schritt werden die Gründe für eine Weiterentwicklung des klassischen Assessment Centers hin zu einem Multimodalen Auswahlverfahren behandelt. In diesem Zuge werden mögliche Bausteine fokussiert und deren prädiktive Validität betrachtet. Den Abschluss bilden Normen und Anforderungen an diagnostische Verfahren.

Begriffsklärung und historische Entwicklung von Assessment Centern

Bei Assessment Centern handelt es sich um eine Methode der Personalauswahl und Personalentwicklung. In diesem umfassenden und standardisierenden Verfahren kommen unterschiedlichste Beurteilungs- und Bewertungsverfahren zum Einsatz (Amelang & Schmidt-Atzert, 2006, S. 458).

Beispielhafte Übungen in einem Assessment Center sind zum einem individuell auszuführende Arbeitsproben und Aufgabensimulationen und zum anderen Gruppendiskussionen und sonstige Gruppenaufgaben mit Wettbewerbs- und/oder Kooperationscharakteristik. Darüber hinaus sind Vorträge und Präsentationen, Rollenspiele, strukturierte oder unstrukturierte Interviews, Selbstpräsentationen, computergestützte Entscheidungssimulationen und psychologische Tests Bestandteil des Auswahlverfahrens (Schuler, 2007, S. 6).

In diesem Verfahren beobachten und bewerten ausgebildete Beurteiler, die optimaler Weise nicht im direkten Vorgesetztenverhältnis zu den Teilnehmern stehen, die Kandidaten (Amelang & Schmidt-Atzert, 2006, S. 458). Die Beurteilungsergebnisse dienen im

nächsten Schritt als Grundlage personalpolitischer Entscheidungen, wie einer Einstellung, Beförderung, Versetzung oder Karriereplanung (Amelang & Schmidt-Atzert, 2006, S. 458 ff).

Ursprünglich geht das Assessment Center auf die deutsche Wehrmachtspsychologie zurück und wurde in den 1930er Jahren zur Auswahl von Offiziersnachwuchskräften der Reichswehr eingesetzt (Hufnagl, 2001, S. 10; Schuler, 2007, S. 7). Im industriellen Bereich wurden Assessment Center erstmals im Jahre 1956 bei der amerikanischen Gesellschaft *American Telephone and Telegraph Company* (AT&T) zu Forschungszwecken verwendet (Hufnagl, 2001, S. 10; Schuler, 2007, S. 8). Grundbaustein für die Verbreitung dieses Auswahlinstrumentes als Methode war die vom Unternehmen in den Jahren von 1956 bis 1966 durchgeführte *Management Progress Study* zur Führungskräftenachwuchsentwicklung (Amelang & Schmidt-Atzert, 2009, S. 462; Bray, Campbell & Grant, 1974).

Die positiven Ergebnisse dieser Studie verhalfen dem Assessment Center zum Durchbruch im industriellen Bereich (Amelang & Schmidt-Atzert, 2009, S. 462). In deutschen Unternehmen hat das klassische Assessment Center erst in den siebziger Jahren durch Tochterunternehmen amerikanischer Konzerne, wie der *International Business Machines Corporation* (IBM), Einzug gehalten. Das Verfahren setzte sich zunächst nur schwer durch und etablierte sich erst Ende der 80er Jahre (Obermann, 2013, S. 18).

Mittlerweile sind Assessment Center in deutschen Unternehmen ein gängiges Personenauswahlinstrument. Aufgrund des hohen zeitlichen und finanziellen Aufwands wurden Assessment Center ursprünglich nur für die Auswahl von Führungskräften, Spezialisten und Trainees eingesetzt. Erst im Jahr 1970 erfolgte die Weiterentwicklung in Richtung eines Personalentwicklungsinstruments. Die Bedeutung in diesem Bereich steigt auch heute noch (Amelang & Schmidt-Atzert, 2006, S. 458).

Diagnostische Einordnung von Assessment Centern

Die „Psychologische Eignungsdiagnostik besteht in dem Bemühen, Zusammenhänge zwischen menschlichen Merkmalen und beruflichem Erfolg zu entdecken bzw. Methoden zu entwickeln, um beides zu messen und zueinander in Beziehung zu setzen“ (Achouri, 2011, S. 79). Grundlage der Eignungsdiagnostik ist die traditionelle Klassifikation psychologischer Merkmale in Kenntnisse, Fähigkeiten und Fertigkeiten sowie Eigenschaften (Achouri, 2011, S. 79).

Beim Versuch, Assessment Center in die Eignungsdiagnostik einzuordnen, ergibt sich ein Spannungsfeld zwischen *eigenschafts- und situationsdiagnostischen* Überlegungen (Amelang & Schmidt-Atzert, 2006, S. 6, 2009, S. 10ff; Drees, 1994, S. 7). Dies ist darauf zurückzuführen, dass sich Assessment Center einerseits mittels der Anforderungsdimensionen personenbezogener Merkmale und Konstrukte bedienen, von denen man annimmt, dass diese über die Zeit hinweg stabil sind. Andererseits werden Situationen bzw. Übungen konstruiert, die ein bestimmtes Verhalten erzielen sollen (Drees, 1994, S. 7).

Im Rahmen der klassischen Eignungsdiagnostik, die sich in der Regel am *eigenschaftstheoretischen* Ansatz ausrichtet, werden psychische Eigenschaften bzw. Persönlichkeitsmerkmale von Menschen aus Verhaltensbeobachtungen erschlossen (Amelang & Schmidt-Atzert, 2009, S. 11, Drees, 1994, S. 7, 1994, S. 6). Für den Diagnostiker besitzen die fokussierten Persönlichkeitsmerkmale einen Prognosewert, wenn die skizzierten Voraussetzungen erfüllt sind:

- *Interindividuelle Varianz*: Zwischen zwei Personen bestehen deutliche Unterschiede in den Verhaltensweisen.

- *Transsituative Verhaltenskonsistenz*: Interindividuelle Verhaltensunterschiede müssen in gleicher oder zumindest ähnlicher Weise über verschiedene Beobachtungsgelegenheiten bestehen.
- *Verhaltensstabilität:* Eigenschaften sind dauerhaft, d.h. sie gelten nicht nur für verschiedene Situationen, sondern auch über die Zeit hinweg (Drees, 1994, S. 6; Schmitt, 1990).

Der *situationsdiagnostische* Ansatz hingegen verfolgt die Zielsetzung, eine Methodik zu entwickeln, die bei einer vorgegebenen Fragestellung Aussagen über psychologisch relevante Charakteristika von Situationen machen kann (Drees, 1994, S. 6). Diesem Grundbaustein zu Folge wird, anders als in der klassischen Eignungsdiagnostik, eine relative Konsistenz der Situationsmerkmale über Personen hinweg unterstellt (Drees, 1994, S. 7). Es stehen das beobachtbare Verhalten und die beobachtbaren situativen Bedingungen des Verhaltens im Vordergrund (Amelang & Schmidt-Atzert, 2009, S. 14). Für *situationsspezifisches* Verhalten und somit den *situationsdiagnostischen* Ansatz spricht, dass Eigenschaften oft so breit definiert sind, dass deren Vorhersagekraft für das Verhalten in einer spezifischen Situation eher gering ist (Amelang & Schmidt-Atzert, 2009, S. 12).

Forschungsergebnisse unterstützen hingegen das *Eigenschaftsmodell*. Die Grundlagen des Big Five Modells der Persönlichkeit sowie das Konzept der Allgemeinen Intelligenz sprechen für den eigenschaftstheoretischen Ansatz (Amelang & Schmidt-Atzert, 2009, S. 13). Zum Beispiel weisen Fleeson und Gallagher (2009) in einer Metaanalyse zum Big Five Modell eine Korrelation von aggregierten Verhaltensmaßen mit Persönlichkeitseigenschaften nach. Es konnten Korrelation zwischen $r = .42$ und $r = .56$ nachgewiesen werden (S. 1097). Darüber hinaus zeigt eine Metaanalyse von Strenze (2007) eine Korrelation von Intelligenz mit dem sozioökonomischen Erfolg eines Menschen. Relevante Stu-

dien zeigen eine Korrelation in Höhe von $r = .56$, $r = .45$ und $r = .23$ zwischen Intelligenz und dem akademischen bzw. beruflichen Status und des Einkommens (S. 411).

Neben diesen beiden Modellen, gibt es die Richtung des *Interaktionismus*. Dieser sieht sowohl die Situation, als auch die Persönlichkeitseigenschaft als Einflussfaktor auf das Verhalten von Menschen. Allerdings hat sich diese Überzeugung auf Dauer nicht durchgesetzt (Amelang & Schmidt-Atzert, 2009, S. 16).

Aus diagnostischer Sicht gibt es bisher kein einheitliches Konzept, das *allen* Assessment Centern zu Grunde liegt. Vielmehr hängt die diagnostische Konzeption von der Zielsetzung des Assessment Centers ab (Drees, 1994, S. 8). Legt der Diagnostiker einen stärkeren Fokus auf die *Anforderungsdimensionen*, geht man von der Verfolgung eines *eigenschaftstheoretischen* Ansatzes aus. Die Anforderungsdimensionen stellen die Erfassung der Persönlichkeitsmerkmale dar. Für die Beurteilung im Rahmen eines Assessment Centers geht man davon aus, dass die personenbezogenen Merkmale unabhängig von den unterschiedlichen Übungssituationen gleichermaßen beobachtet und beurteilt werden können. Deshalb sollten die Übungssituationen keinen entscheidenden Einfluss auf die Beurteilung haben (Drees, 1994, S. 8).

Die *Situationsdiagnostiker* hingegen konstruieren Situationen, in welchen bestimmte Verhaltensweisen bei Kandidaten provoziert werden. Je größer die Generalisierbarkeit der Situation, desto höher ist deren diagnostischer Wert. In diesem Fall ist die Gesamtkonstellation der Übung ausschlaggebend für die Beurteilung und nicht die Einzeldimensionen über verschiedene Situationen hinweg. Hiermit wird von einer *übungsbezogenen* Beurteilung gesprochen (Drees, 1994, S. 8 f).

Die beiden skizzierten Richtungen werden im empirischen Teil dieser Arbeit in die Praxis umgesetzt und miteinander verglichen. Daneben spielen die Grundlagen der beiden

Ansätze eine Rolle bei der Betrachtung der in den nachfolgenden Kapiteln skizzierten Gütekriterien diagnostischer Verfahren.

Hauptgütekriterien diagnostischer Verfahren

Gütekriterien machen eine Aussage über die Qualität eines psychologischen Tests oder Verfahrens (Amelang & Schmidt-Atzert, 2009, S. 130). Die Hauptgütekriterien diagnostischer Verfahren und im Rahmen der empirischen Untersuchung fokussiert, sind:

- *Objektivität:* Wie stark hängt das Ergebnis des Verfahrens davon ab, wer dieses durchführt, auswertet oder interpretiert?
- *Reliabilität:* Wie genau oder zuverlässig ist das Messergebnis? Wie stark verändert sich dieses zum Beispiel bei einer Testwiederholung?
- *Validität:* Wie gut erfolgt die Messung des Merkmals – und nicht eines anderen – mit genau diesem Test oder Verfahren (Amelang & Schmidt-Atzert, 2009, S. 131)?

In den nachfolgenden Kapiteln werden die genannten Gütekriterien genauer erläutert, wobei ein Fokus auf der Validität liegt, da diese das wichtigste Gütemerkmal eines Tests darstellt (Amelang & Schmidt-Atzert, 2009, S. 143).

Objektivität.

„Objektivität bedeutet, dass die Ergebnisse eines diagnostischen Verfahrens unabhängig davon zustande kommen, wer die Untersuchung, die Auswertung und die Interpretation durchführt“ (Amelang & Schmidt-Atzert, 2009, S. 133).

Zur Bestimmung der Objektivität eines diagnostischen Verfahrens werden Maßnahmen zur Standardisierung der *Durchführung*, *Auswertung* und *Interpretation* definiert. Diese Maßnahmen sind Bestandteil des diagnostischen Verfahrens und müssen dokumentiert sein. Eine Aussage über die Höhe der Objektivität kann durch eine Bewertung der genannten Maßnahmen erzielt werden (Amelang & Schmidt-Atzert, 2009, S. 133).

Die *Durchführungsobjektivität* stellt sicher, dass ein Verfahren immer auf die gleiche Weise durchgeführt wird. Dies bedeutet beispielsweise, dass alle Kandidaten das gleiche Übungsmaterial erhalten und die Übungen unter gleichen Bedingungen durchführen. Eine hohe Durchführungsobjektivität lässt sich erzielen, wenn alle Kriterien festgelegt sind, die sich erfahrungsgemäß auf das Verhalten der Kandidaten auswirken können. Eine vollkommene Standardisierung von Assessment Centern ist jedoch aufgrund der verschiedenen Interaktionen zwischen Teilnehmern und Beobachtern nicht möglich (Amelang & Schmidt-Atzert, 2009, S. 113f).

Die *Auswertungsobjektivität* bezieht sich darauf, dass gleiche Testteile den gleichen Antwortkategorien zugeordnet werden. D.h. anhand von klaren Anweisungen und Hilfsmitteln zur Auswertung werden die Beurteilungen vorgenommen (Fisseni & Preusser, 2007, S. 237).

Interpretationsobjektivität ist dann gegeben, wenn alle Testanwender den Rohwert eines Probanden gleich interpretieren (Amelang & Schmidt-Atzert, 2009, S. 136).

Die *Auswerter*- und *Interpretationsobjektivität* lassen sich in sich in einem Assessment Center Verfahren auf die gleiche Weise ermitteln. Erfasst wird die Korrelation zwischen den Ratings verschiedener Beobachter in der selben Dimension und Übung (Fisseni & Preusser, 2007, S. 238). Diese fällt in der Regel relativ hoch aus und kann durch ein gezieltes Beobachtertraining im Vorfeld deutlich gesteigert werden (Fisseni & Preusser, 2007, S. 239). Verschiedene Studien und Metaanalysen weisen Werte zwischen $r = .60$

und r = .90 nach (Fisseni & Preusser, 2007, S. 238). Weitere Studien zur *Auswerter- und Interpretationsobjektivität* werden im Kapitel zur Reliabilität noch im Detail aufgegriffen, da die Objektivität im Rahmen der Assessment Center Forschung oftmals als Teilaspekt der Reliabilität angesehen wird (Fisseni & Preusser, 2007, S. 239).

Reliabilität.

Reliabilität bezeichnet die Genauigkeit oder Zuverlässigkeit der Messung beziehungsweise deren Freiheit von Zufallsfehlern (Bortz, 2004, S. 10). Reliabilität ist eine notwendige, allerdings keine hinreichende Bedingung für Validität (Achouri, S. 82; Achouri, 2011, S. 82). Die Validität kann maximal so hoch sein, wie die Wurzel ihrer Reliabilität (Obermann, 2013, S. 286). Zwischen 20 und 37 Prozent der Varianz in den Beobachtungsurteilen gehen nicht auf Unterschiede zwischen den Kandidaten, sondern auf die Beobachterunterschiede zurück (Obermann, 2013, S. 286). Dies zeigt, welche hohe Bedeutung die Reliabilität und deren Einflussfaktoren auf diagnostische Verfahren haben.

Die Reliabilitätskoeffizienten liegen zwischen null und eins (Amelang & Schmidt-Atzert, 2009, S. 137). Je höher der Wert, desto besser ist auch die Reliabilität. Allerdings gibt es unterschiedliche Schätzmethoden zur Messung der Reliabilität von diagnostischen Verfahren. Die einzelnen Kennwerte liefern unterschiedliche Erkenntnisse und sind nicht austauschbar (Amelang & Schmidt-Atzert, 2009, S. 137). In der Literatur sind die beiden folgenden Schätzmethoden zur Erfassung der Reliabilität im Rahmen von Assessment Centern aufgeführt (Amelang & Schmidt-Atzert, 2009, S. 466; Obermann, 2013, S. 286; Zenglein, 2010, S. 22):

- Die *Interrater-Reliabilität*: Diese gibt an, wie gut die Beurteilungen der verschiedenen Beobachter übereinstimmen (Obermann, 2013, S. 286).

- Die *Retest-Reliabilität:* Diese trifft eine Aussage über die zeitliche Stabilität der Aussagen (Obermann, 2013, S. 286).

Die *Interrater-Reliabilität* stellt zugleich ein Maß der Objektivität – nämlich der eigentlichen Auswertungsobjektivität (Amelang & Schmidt-Atzert, 2009, S. 466) – dar, wie bereits im vorherigen Kapitel zur Objektivität skizziert. Sie ist eine wichtige Voraussetzung für die Zuverlässigkeit eines Assessment Centers (Fecker, 1989, S. 106f; Obermann, 2009, S. 277).

Ein Problem besteht in der Methodik, die *Interrater-Reliabilität* zu ermitteln, da es sich nicht immer um die gleichen Beurteiler und Kandidaten handelt. Meist wird die Ermittlung eines Korrelationskoeffizienten als Ausdruck der *Interrater-Reliabilität* herangezogen. Bei dieser Herangehensweise besteht allerdings das Problem, dass bei unterschiedlichen Bewertungen trotzdem eine nahezu perfekte Korrelation ermittelt werden kann, da nur die relativen Rangplätze der Einschätzungen berücksichtigt werden und nicht die absoluten Werte. Solange der Abstand der beobachteten Teilnehmer-Werte untereinander gleich bleibt, weisen Beobachter auch mit einem sehr verschiedenen Maßstab eine hohe *Interrater-Reliabilität* auf. Für eine hohe Übereinstimmung müssten die Beobachter zu gleichen Einschätzungen kommen, für eine hohe Interrater-Korrelation ist es ausreichend, wenn die Rangplätze zwischen den bewerteten Teilnehmern gleich sind (Obermann, 2013, S. 286).

In der Praxis ist eine andere empirische Ermittlung allerdings schwierig zu erzielen. Die verschiedenen Einschätzungen zwischen Beobachtern sind zwar einfach zu ermitteln, in der Regel haben jedoch nicht alle Beobachter jeden Teilnehmer gesehen. Eine Rotation der Teilnehmer über sämtliche Bedingungen ist nur im Forschungskontext möglich, um

damit die Varianz der Ergebniswerte auf Beobachter, Teilnehmer und deren statistische Interaktion zurück zu führen (Obermann, 2013, S. 287).

Einen Überblick über die *Interrater-Korrelation* einzelner Beurteilungsdimensionen im Rahmen von Assessment Centern gibt eine Metaanalyse von Howard (1974) und ist in Tabelle 1 dargestellt.

Tabelle 1

Summary of Interrater Reliability Studies of Assessment Procedures (eigene Darstellung in Anlehnung an (Howard, 1974, S. 121)

Source	Company	Variables	Assessors	Interrater Reliability (r)
Thomson (43) (N = 71)	SOHIO	13 dimensions	2 psychologists	Ratings, .73-.93, $\bar{r}$ = .85
Thomson (43) (N = 71)	SOHIO	13 dimensions	3 managers	Ratings, .78-.95, $\bar{r}$ = .89
Thomson (43) (N = 71)	SOHIO	Potential	2 psychologists	Ratings, .89
Thomson (43) (N = 71)	SOHIO	Potential	3 managers	Ratings, .93
McConnell& Parker (36) (N = 12)	AMA Clients	a)12 categories b) Potential	5 managers 5 managers	Ratings, .64-.90 Ratings, .83
McConnell& Parker (36) (N = 12-48)	6 AMA Clients	Overall manage-ment abilities	5 managers	Ratings, .85- .98
Greenwood& McNamara (26) (N = 288)	IBM	a) Task force game b) Leaderless group c) Mfg. Problem	All pairs of 3 alternating ob-servers	a) Ratings, .70 b) Ratings, .66 c) Ratings, .74
Bray & Grant (6) (N = 355)	AT&T	a) Leaderless group b) Mfg. Problem c) In basket	2 psychologists 2 psychologists 2 psychologists	a) Ratings, .75 b) Ratings, .60 c) Ratings, .92
Grant, Kato-vsky, & Bray (25) (N = 355)	AT&T	9 variables from projective tests	2 psychologists	Ratings, .85- .94
Grant & Bray (24) (N = 355)	AT&T	18 variables from interview data	2 psychologists	Median = .82 college, .72 non- college

Die meisten Studien zeigen Werte für einzelne Beurteilungsdimensionen zwischen $r = .60$ und $r = .94$. 50 Prozent der ermittelten Korrelationskoeffizienten liegen über $r = .75$ (Howard, 1974, S. 121). Thornton und Byham (1982) ermitteln ähnlich hohe Werte zwischen $r = .80$ für und $r = .70$. Richtet man den Fokus weg von den Beurteilungsdimensionen hin zu den einzelnen Übungen, so weist eine Studie von Bray und Grant (1966) *Interrater-Korrelations-Werte* zwischen $r = .60$ für eine führerlose Gruppendiskussion und $r = .92$ für eine Postkorbübung auf (Bray & Grant, 1966).

Drees (1994) hat zusätzlich die *Interrater-Korrelation* bei unterschiedlichen Beobachtertrainings untersucht. Zum einem wurde ein *kognitionsbezogenes* Beobachtertraining nach dem bereits skizzierten *eigenschaftstheoretischem* Ansatz vorangestellt und zum anderen ein *verhaltensbezogenes* Beobachtertraining nach dem *situationstheoretischen* Modell. Dabei lag die durchschnittliche *Interrater-Korrelation* beim verhaltensbezogenen Training mit einem durchschnittlichen Wert von $\bar{r} = .71$ deutlich über dem des kognitionsbezogenen Trainings mit $\bar{r} = .43$ (Drees, 1994, S. 112).

Eine hohe Urteilsübereinstimmung und somit eine hohe *Interrater-Korrelation* bzw. *-Reliabilität* bedeutet nicht unbedingt eine hohe Genauigkeit dieser Urteile, schließlich können alle Beobachter in einem Beobachterteam zwar hoch übereinstimmen, in ihrem Urteil jedoch alle falsch liegen. Eine hohe Übereinstimmung bedeutet also nicht unbedingt eine hohe Genauigkeit der Beurteilung im Rahmen von Assessment Centern (Obermann, 2013, S. 287). Eine falsche Beurteilung wird beispielsweise durch ein manipulatives Verhalten der Kandidaten verursacht. Diesem Effekt kann durch ein Beobachtertraining und den Einsatz erfahrener Beobachter entgegengewirkt werden (Lievens, 2002, S. 684, 2002, S. 684; Obermann, 2009, S. 278 ff). Bereits Lorenzo (1984) machte im Rahmen seiner Untersuchungen explizit darauf aufmerksam, dass die Erfahrung der eingesetzten Beobach-

ter eine wesentliche Voraussetzung für eine reliable Beobachtung ist (Lorenzo, 1984, S. 629ff).

Obermann (2013) weist in diesem Zusammenhang auf eine Studie von De Kock, Born und Lievens (2009) hin, in welcher die Determinanten der Beobachtergenauigkeit untersucht sind. Das Wissen der Beobachter, wie Persönlichkeitsfaktoren mit beobachtetem Verhalten verknüpft sind, ist mit $r = .52$ der wesentliche Einflussfaktor für die spätere Genauigkeit der Einschätzung. Dieses Wissen hängt wiederum mit der allgemeinen kognitiven Leistungsfähigkeit der Beobachter zusammen (Obermann, 2013, S. 290).

Studien, die die *Retest-Korrelation* als Schätzung der Reliabilität erfassen, sind aufgrund der hohen Kosten für die Durchführung von Assessment Centern eher selten (Amelang & Schmidt-Atzert, 2009, S. 466; Obermann, 2013, S. 290). Dennoch liegen in der Literatur einige Studien dazu vor und werden im Folgenden betrachtet.

Moses (1973) führte im Rahmen einer Studie zu Assessment Centern im Abstand von mehr als einem Monat Testwiederholungen durch. Hierbei wurde ein sehr hoher Wert von $r = .73$ ermittelt (Moses, 1973, S. 574). Dieser Wert wird jedoch kontrovers diskutiert, da ein Intelligenztest mit eingerechnet wurde, der alleine eine *Retest-Korrelation* von $r = .72$ aufweist (Obermann, 2009, S. 280; Schuler, 2007, S. 238; Zenglein, 2010, S. 24).

Kleinmann (1997) vergleicht in einer Studie zwei aufeinander folgende Assessment Center. Bei 63 Teilnehmern, die zwischen den beiden Assessment Center-Wiederholungen kein Feedback erhielten, lag die *Retest-Korrelation* bei $r = .65$. Für Teilnehmer, die insbesondere dimensionsbezogene Rückmeldungen erhielten, lag der Wert bei $r = .34$ (Kleinmann, 1997).

Kelbetz und Schuler (2002) führten eine Studie zur *Retest-Korrelation* in der Finanzdienstleistungsbranche durch. In dieser Studie wurden die Assessment Center Leistungen von 47 Teilnehmern eines internen Assessment Centers in einer Zeitspanne der Wiederho-

lung von zwei Jahren verglichen. Das Verfahren ist inhaltlich identisch geblieben, den Beobachtern war die Tatsache der Verfahrenswiederholung im Vorfeld nicht bekannt. Der Wert der *Retest-Korrelation* lag nach durchschnittlich zwei Jahren bei $r = .41$. Für den Übungsgewinn durch wiederholte Teilnahme ergab sich nach Korrektur des Regressionseffekts eine Effektstärke von mindestens $d = .40$. Sowohl die *Retest-Korrelation* der Einzelverfahren als auch die Leistungssteigerungen der Teilnehmer in diesen Einzelverfahren waren stark unterschiedlich. Ein systematischer Zusammenhang zwischen den *Retest-Korrelationen* und der Leistungssteigerungen der Kandidaten war allerdings nicht erkennbar (Kelbetz & Schuler, 2002, S. 4ff).

Validität.

Die Validität gibt Auskunft über die Gültigkeit eines eignungsdiagnostischen Verfahrens. Man versteht darunter den Grad der Genauigkeit, mit dem ein Test oder Verfahren das Persönlichkeitsmerkmal oder die Verhaltensweise, das bzw. die es messen soll, auch tatsächlich misst (Amelang & Schmidt-Atzert, 2009, S. 142; Lienert & Raatz, 1998, S. 16; Obermann, 2013, S. 291). Ein Test oder Verfahren „(…) ist demnach vollkommen valide, wenn seine Ergebnisse einen unmittelbaren und fehlerfreien Rückschluss auf den Ausprägungsgrad des zu erfassenden Persönlichkeits- oder Verhaltensmerkmals zulassen (…)“ (Lienert & Raatz, 1998, S. 16).

Validität ist das wichtigste Gütemerkmal eine Tests (Amelang & Schmidt-Atzert, 2009, S. 143). Je nach Verwendungszeck oder Konstruktionsprinzip kann die eine oder andere Art der Validität besonders wichtig sein. Im allgemeinen unterscheidet man jedoch zwischen drei wichtigen Arten von Validität, der *Inhalts-, Kriteriums- und Konstruktvalidität* (Amelang & Schmidt-Atzert, 2009, S. 144; Obermann, 2013, S. 292). Übertragen auf das Konstrukt Assessment Center ergibt sich daraus die im nachfolgenden skizzierte Erläuterung:

Die *Inhalts-* bzw. *Kontentvalidität* macht eine Aussage darüber, wie gut die eingesetzten Übungen bzw. Aufgaben in Assessment Centern die spätere berufliche Tätigkeit repräsentieren (Obermann, 2009, S. 282; Zenglein, 2010, S. 30).

Die *Konstruktvalidität* gibt an, inwieweit das Verfahren tatsächlich ein spezifisches Merkmal und nicht ein anderes erfasst. D.h. ob im Konstrukt Assessment Center tatsächlich bestimmte Persönlichkeitsdimensionen abgebildet werden (Obermann, 2013, S. 292).

Die *Kriteriums-* oder auch *prädiktive* bzw. *prognostische* Validität erfasst den Bezug zwischen dem Assessment Center Ergebnis und einem relevanten Außenkriterium (Amelang & Schmidt-Atzert, 2009, S. 146; Obermann, 2013, S. 313). Hierfür wird häufig die Leistungsbeurteilung durch den Vorgesetzten verwendet. Die prädiktive Validität ist die wichtigste Form und hat die höchste Bedeutung für die Beurteilung der Gültigkeit von Assessment Centern (Obermann, 2013, S. 292).

Eine hohe prädiktive Validität ist insbesondere bei der Personalauswahl ausschlaggebend, da eine Prognose im Hinblick auf den späteren Berufserfolg gegeben werden kann (Achouri, 2011, S. 84; Obermann, 2009, S. 282; Zenglein, 2010, S. 32f). Eine hohe Konstruktvalidität ist insbesondere für die Personalentwicklung von entscheidender Bedeutung. Grund hierfür ist, dass auf Basis der erfassten Anforderungsdimensionen Entwicklungsmaßnahmen definiert werden (Obermann, 2009, S. 282; Zenglein, 2010, S. 25).

Alle drei Formen von Validität werden in den nachfolgenden Kapiteln genauer erläutert und relevante Studien zusammengefasst. Der Fokus liegt aufgrund vielfältiger, kontroverser Diskussionen und zahlreicher Studien auf der Konstruktvalidität und aufgrund der hohen Bedeutung auf der prädiktiven Validität.

Die Inhaltsvalidität

Die Inhaltsvalidität oder auch Augenscheinvalidität liefert eine Aussage darüber, wie repräsentativ die im Assessment Center erfasste Verhaltensstichprobe für den Verhaltensbe-

reich ist, auf den man Rückschlüsse ziehen will. D.h. wie repräsentativ sind die Assessment Center Übungen für den beruflichen Alltag (Obermann, 2013, S. 310).

Ausschlaggebend und als Prüfkriterium zu verwenden sind nach Wernimont und Campbell (1986) insbesondere folgende zwei Prüffragen:

- Erfassen die Übungen des Assessment Centers alle Situationsmerkmale des Alltags, für deren Bewältigung bestimmte Verhaltenskompetenzen notwendig sind?
- Sind diese Übungen im Stande, erfolgskritische Verhaltensunterschiede der Kandidaten sichtbar zu machen? (Wernimont & Campbell, 1968, S. 373ff)

Schippmann, Hughes und Prien (1987) empfehlen einen Fragebogen einzusetzen, der eine Beschreibung der Jobinhalte enthält. Dieser wird dann von Experten oder Stelleninhabern auf deren Wichtigkeit bewertet. Darüber hinaus ist für die Konstruktion von Assessment Centern wichtig, welche Fähigkeiten die Kandidaten vorher mitbringen können und welche sie auch erst im Job erwerben können (J. S. Schippmann, G. L. Hughes & E. P. Prien, 1987, S. 355ff).

Allerdings kann alleine der Bezug der Assessment Center Übungen zur beruflichen Praxis keine hohe (Inhalts-)Validität erzielen, denn wie lässt sich sonst erklären, dass beispielsweise kognitive Fähigkeitsteste – die in einem späteren Kapitel noch genauer betrachtet werden – eine hohe prädiktive Validität besitzen, aber kaum Bezug zur Jobposition haben (Obermann, 2009, S. 301f).

Aus dem soeben aufgeführten Grund und der Tatsache, dass die Inhaltvalidität nur am Rande eine Rolle für die empirische Untersuchung spielt, wird im Rahmen dieser Arbeit im Detail nicht weiter eingegangen. Dennoch leiten sich aus den skizzierten Prüffra-

gen von Wernimont und Campbell (1968) teilweise die im späteren Kapitel beschriebenen Normen und Anforderungen an diagnostische Verfahren ab.

Die Konstruktvalidität

„Während man lange Zeit davon ausging, daß eine „augenscheinvalide“ Auswahl und Definition der Anforderungsdimensionen und eine entsprechende Entwicklung der Übungen ausreicht, um die Konstruktvalidität des AC-Verfahrens sicherzustellen, so wurde dies durch einige Untersuchungen deutlich in Frage gestellt“ (Drees, 1994, S. 12).

Keine andere Frage zum Konstrukt des Assessment Centers hat die Wissenschaft in den vergangenen dreißig Jahren mehr beschäftigt, als die Konstruktvalidität (Obermann, 2013, S. 292). Nach wie vor sind die Ergebnisse immer noch relativ ähnlich: Betrachtet man die Zusammenhänge der Beurteilungen zu den verschiedenen Dimensionen innerhalb einzelner Übungen, fällt deren Einschätzung durch die Beobachter sehr ähnlich aus und differenziert kaum. Vergleicht man hingegen die Beurteilungen der gleichen Dimensionen in unterschiedlichen Übungen, erhält man kaum konsistente Leistungen, obwohl es sich eigentlich um das gleiche Konstrukt handelt (Obermann, 2013, S. 292f).

Durch diese Erkenntnisse ist eine Grundannahme des Assessment Centers bedroht – nämlich dass sich innerhalb des Verfahrens stabile Konstrukte bzw. Dimensionen beobachten lassen, die über Situationen und Zeiten hinweg stabil sind (Obermann, 2013, S. 293). Fokussiert man sich in der Personalpraxis lediglich auf den Gesamtergebniswert des Assessment Centers, so hat diese konzeptionell wichtige Grundvoraussetzung wenig Bedeutung. Werden allerdings einzelne Dimensionen interpretiert, verglichen oder als Feedback herangezogen, kann dies gravierende Auswirkungen haben. Beispielweise werden Fehlentscheidungen in Bezug auf die Einstellung oder auf einzelne Entwicklungsmaßnahmen getroffen (Obermann, 2013, S. 293).

Im nachfolgenden erfolgt zunächst ein kurzer Einblick in methodische Grundlagen zur Erfassung der Konstruktvalidität.

Grundvoraussetzungen der Konstruktvalidität sind nach Campbell und Fiske (1959) die *konvergente* und *diskriminante* Validität. Von konvergenter Validität spricht man, wenn Messungen der gleichen Merkmale trotz verschiedener Methoden hoch korrelieren. Diese Korrelationen sollten höher ausfallen als die Korrelationen der Dimensionen innerhalb der Übungen. Dies ist eine Voraussetzung für die diskriminante Validität (Campbell & Fiske, 1959, S. 81f).

Das Konzept zur Überprüfung der Konstruktvalidität nach Campbell und Fiske (1959) ist die Multitrait-Multimethod-Matrix (MTMM). Nachfolgende Bedingungen müssen für eine vorhandene Konstruktvalidität erfüllt sein:

- Die *Monotrait-Heteromethod* Korrelationen sollten signifikant größer null sein (konvergente Validität).
- Die *Heterotrait-Heteromethod* Korrelationen sollten signifikant kleiner als die *Monotrait-Heteromethod* Korrelationen sein.
- Die *Heterotrait-Monomethod* Korrelationen sollten ebenso signifikant kleiner sein als die *Monotrait-Heteromethod* Korrelationen (diskriminante Validität).
- Die *Heterotrait-Monomethod* Korrelationen sollten das gleiche Muster wie für die *Heterotrait-Heteromethod* Korrelationen aufweisen (Campbell & Fiske, 1959, S. 82f).

Zahlreiche Studien haben sich bereits mit der Konstruktvalidität von Assessment Centern beschäftigt. Die ersten Untersuchungen gingen von Sackett und Dreher (1982) aus. Ein Ergebnis der Studie ist die Erkenntnis, dass innerhalb einzelner Übungen die Zusammen-

hänge der Beurteilungen zu den verschiedenen Dimensionen sehr ähnlich ausfallen und kaum differenzieren (geringe diskriminante Validität). Beispielhafte Werte für die durchschnittliche Korrelation verschiedener Dimensionen innerhalb der gleichen Übungen in den drei betrachteten Organisationen sind $\bar{r}$= .64, $\bar{r}$ = .40 und $\bar{r}$ = .65 (Sackett & Dreher, 1982, S. 408). Vergleicht man hingegen die Beurteilungen der gleichen Dimensionen in unterschiedlichen Übungen, so erhält man kaum konsistente Leistungen, obwohl es sich um das gleiche Konstrukt handelt (geringe konvergente Validität). In allen drei betrachteten Organisationen waren die Korrelationen innerhalb der Übungen deutlich stärker als die Korrelationen der gleichen Dimensionen in verschiedenen Übungen. Folglich ist die Konstruktvalidität nach Campbell und Fiske (1959) für diese Untersuchung nicht gegeben (Sackett & Dreher, 1982, S. 406). Zu dieser Zeit finden zahlreiche andere Studien ähnliche Ergebnisse heraus (Neidig & Neidig, 1984; Sackett & Dreher, 1984; Sackett & Harris, 1988; Turnage & Muchinsky, 1984).

Eine weitere Untersuchung von Drees (1994) wird im nachfolgenden detaillierter behandelt, da hier der bereits betrachtete Unterschied des *eigenschaftstheoretischen* und *situationstheoretischen* Ansatzes durch zwei verschiedene Arten von Beobachtertrainings dargestellt wird. Zum einen fand im Vorfeld des Assessment Centers ein *kognitionsbezogenes* Training der Beobachter, das dem eigenschaftstheoretischen Ansatz entspricht, statt. Zum anderen ein *verhaltensbezogenes* Beobachtertraining, das dem situationstheoretischen Ansatz entspricht. Im ersten Fall wird ein stärkerer Fokus auf die Beurteilung der Übungen gelegt, in zweiten Fall auf die Beurteilung der Anforderungsdimensionen.

Die Ergebnisse der Untersuchung sind in Tabelle 2 skizziert und werden im Anschluss genauer erläutert.

Tabelle 2

Übersicht zu den konvergenten und diskriminanten Validitäten (eigene Darstellung in Anlehnung an (Drees, 1994, S. 84))

		Kognitionsbezogenes Training	Verhaltensbezogenes Training
Konvergente Validität	gleiche Anforderungen/ versch. Übungen	.30	.35
Diskriminante Validität	versch. Anforderungen/ versch. Übungen	.16	.22
	versch. Anforderungen/ gleiche Übungen	.45	.59

Anmerkung. Bei den Korrelationskoeffizienten handelt es sich um den Rangkorrelations-koeffizienten Kendall's Tau.

Bei einem *kognitionsbezogenen* Beobachtertraining beträgt der durchschnittliche konvergente Validitätskoeffizient $\bar{r}$ = .30. Die durchschnittliche Korrelation unterschiedlicher Anforderungen in unterschiedlichen Übungen beträgt $\bar{r}$ = .16. Die durchschnittliche übungsbezogene Korrelation beträgt $\bar{r}$ = .45 und liegt somit deutlich über dem anforderungsbezogenen Wert mit $\bar{r}$ = .30. Dies bedeutet folglich, dass im Rahmen dieses Assessment Centers hingegen des vorangegangenen kognitiven Beobachtertrainings eher übungsbezogen als dimensionsbezogen beurteilt wird, was wiederum den Annahmen der Konstruktvalidität widerspricht (Drees, 1994, S. 82).

Bei einem *verhaltensbezogenen* Beobachtertraining liegen die Werte für die durchschnittliche konvergente Validität etwas höher mit $\bar{r}$= .35. Bei der diskriminanten Validität bestehen bei dieser Variante des Beobachtertrainings gravierende Unterschiede. Gemäß den Voraussetzungen zur diskriminanten Validität liegen die durchschnittlichen Korrelati-

onen zwischen verschiedenen Anforderungen und unterschiedlichen Übungen mit $\bar{r}$ = .22 niedriger als die durchschnittlichen konvergenten Korrelationskoeffizienten mit $\bar{r}$ = .35. Der entscheidende Unterschied und ein Verstoß gegen die Regeln der diskriminanten Validität und somit der Konstruktvalidität liegt in der hohen Korrelation der Anforderungen innerhalb einer Übung mit $\bar{r}$ = .59. Dieser Durchschnitt liegt deutlich über dem der konvergenten Korrelationskoeffizient mit $\bar{r}$ = .35. Die Übungen sind folglich auch hier stärker übungsbezogen als dimensionsbezogen beurteilt worden (Drees, 1994, S. 82f).

Insgesamt wird ersichtlich, dass in beiden Arten des Beobachtertrainings die konvergente Validität gering ist und die diskriminante Validität unzureichend und somit keine befriedigende Konstruktvalidität gewährleistet ist (Drees, 1994, S. 82f).

Shore, Shore und Thornton (1992) liefern erste Beweise für die Konstruktvalidität von Assessment Centern, wenn man die einzelnen Dimensionen zwei Überdimensionen zuordnet, den *interpersonellen* und den *performance* Dimensionen. Das Ergebnis zeigt auf, dass die Interkorrelationen der Dimensionen innerhalb der *interpersonellen* ($\bar{r}$ = .51) und *performance* ($\bar{r}$ = .59) Kategorien größer waren als die Korrelationen über die Übungsdimensionen hinweg. Eine Faktorenanalyse bestätigt die beiden Hauptfaktoren der Dimensionen (Shore et al., 1992, S. 109f).

Gaugler und Thornton (1989) wählten einen ähnlichen Ansatz und reduzierten die Anzahl der Assessment Center Dimensionen. Sie unterschieden zwischen drei, sechs und neun zu beobachteten Dimensionen. Letztendlich konnte zwar durch eine Minimierung der Anzahl der Dimensionen die konvergente Validität positiv beeinflusst werden, eine Steigerung der diskriminanten Validität und somit der gesamten Konstruktvalidität jedoch konnte nicht erzeugt werden (Gaugler & Thornton, 1989, S. 616).

Darüber hinaus haben sich verschieden Studien mit den Einflussfaktoren auf die Konstruktvalidität von Assessment Centern beschäftigt.

Eine Metaanalyse von Woehr (2003) zeigt auf, dass die Assessment Center Beurteilungen nur so gut sein können wie die Entwicklung, das Design und die Implementierung des Verfahrens (Woehr, 2003, S. 251). Methodische Einflussfaktoren, wie die Anzahl der zu beobachteten Dimensionen innerhalb der Übungen oder die Qualität und Länge des Beobachtertrainings, haben einen entscheidenden Einfluss auf die Konstruktvalidität (Woehr, 2003, S. 248). Melchers, Henggeler und Kleinmann (2007) bestätigen, dass der Beurteilungszeitpunkt, der Austausch von Informationen zwischen den Beobachtern oder eine Rotation der Beobachter ebenso Moderatoren der Konstruktvalidität sind (Melchers et al., 2007, S. 141).

Eine Metaanalyse von Bowler und Woehr (2006) kommt zu dem Ergebnis, dass es gravierende Unterschiede bei einzelnen Dimensionen gibt. Nicht alle Assessment Center-Dimensionen haben einen gleich hohen Einfluss auf die gesamte Konstruktvalidität (Bowler & Woehr, 2006, S. 1120ff). Sie empfehlen folglich, dass der Fokus nicht auf einzelnen Dimensionsbewertungen liegen soll, sondern auf dem Gesamturteil, da dieses relevant für weitere Schritte und das Feedback an den Kandidaten ist (Bowler & Woehr, 2006, S. 1123).

Kuncel und Sackett (2014) kommen zu einem ähnlichen Ergebnis wie Bowler und Woehr (2006). Sie raten, sich nicht auf die Bewertung der einzelnen Dimensionen in den Übungen zu fokussieren, sondern auf das Gesamturteil der Dimensionsbewertungen. Aufgrund der Tatsache, dass in klassischen Assessment Centern die einzelnen Dimensionen in einer Vielzahl von Übungen bewertet werden, sinkt die relative Wichtigkeit der übungsspezifischen Varianz und die Wichtigkeit der Varianz zwischen den Übungen steigt (Kuncel & Sackett, 2014, S. 44). Aggregiert man die Einzelbewertungen einer gegebenen Dimension zu einem Gesamturteil der Dimension, sinkt die Rolle der übungsspezifischen Varianz. Die Autoren empfehlen folglich, Gesamturteile als Basis für Personalentwick-

lungsmaßnahmen und ein Assessment Center Feedback zu wählen und nicht auf übungsbezogene Einzeldimensionsurteile einzugehen (Kuncel & Sackett, 2014, S. 46).

Trotz der soeben skizzierten Indizien und Hinweise auf die Konstruktvalidität von übergeordneten Dimensionen kann kein allumfassender Hinweis für eine Konstruktvalidität von Assessment Centern geliefert werden (Bowler & Woehr, 2006; Gaugler & Thornton, 1989; Kuncel & Sackett, 2014; Shore et al., 1992). Die Untersuchungen zeigen deutlich, dass Beobachter dazu geneigt sind, eher Pauschalurteile innerhalb einer Übung abzugeben, als Merkmale zwischen mehreren Übungen zu beurteilen (Drees, 1994; Neidig & Neidig, 1984; Sackett & Dreher, 1982, 1984; Sackett & Harris, 1988; Turnage & Muchinsky, 1984). Folglich stellt sich die Frage, ob Assessment Center nicht eher übungsbezogene Fertigkeiten erfassen, als verschiedene Dimensionen über mehrere Übungen hinweg. Demnach ist es eine offensive Alternative, auf die Dimensionen zu verzichten und das Modell der aufgabenorientierten Assessment Center einzusetzen (Obermann, 2013, S. 310).

Die prädiktive Validität

Die prädiktive Validität gibt den statistischen Zusammenhang zwischen dem Assessment Center Ergebnis und einem externen Kriterium wie den beruflichen Erfolg wieder. Diese wird angegeben als statistisches Korrelationsmaß. Sie ist das entscheidende Qualitätsmaß und hat die höchste Bedeutung für die Beurteilung der Gültigkeit des Verfahrens. Zur Messung des Berufserfolgs werden in der Regel Kriterien wie der Einkommenszuwachs, die Anzahl der Beförderungen, die Vorgesetztenbeurteilung oder der Zuwachs an Verantwortung herangezogen (Obermann, 2013, S. 313).

Die Vorgesetztenbeurteilung lässt sich in die nach Amelang und Schmidt-Atzert (2006) definierte Ebene der *Leistungsbeurteilung* eingliedern (S. 444). Diese dient in erster Linie der Personalentwicklung und des Personalmanagements. Durchführungsart ist in der

Regel ein Vier-Augengespräch zwischen Mitarbeiter und Vorgesetzten und sollte aus Objektivitätsgründen mindestens in halbstandardisierter Form ablaufen. Gewöhnlich werden im Rahmen des Gesprächs auch individuelle Entwicklungs- und Fördermaßnahmen vereinbart. Ergänzend zu einer Schulung des Vorgesetzten zur Gesprächsführung sollten dem Beurteiler auch klar definierte Einstufungsverfahren wie Skalen zur Verhaltensbeobachtung zur Verfügung stehen (Amelang & Schmidt-Atzert, 2006, S. 444f).

Die wohl bekannteste Studie, die sich mit der prognostischen Validität von Assessment Centern befasst, ist die bereits in der Einleitung erwähnte *Management Progress Study* von Bray und Grant (1966). Die American Telephone & Telegraph Company (AT&T) startete im Jahre 1956 eine Langzeitstudie, wobei die Daten vor den Vorgesetzten und Beurteilten geheim gehalten wurden, damit die Ergebnisse nicht die Karriere der betroffenen Personen beeinflusste. 422 Personen durchliefen damals ein Assessment Center und wurden im Zeitverlauf in ihrer Karriereentwicklung beobachtet. Im Rahmen des Auswahlverfahrens mussten die Beurteiler neben der Bewertung verschiedener Anforderungsdimensionen zusätzlich eine Prognose abgeben, inwiefern die Kandidaten in das mittlere Management aufrücken werden (Bray & Grant, 1966, S. 1ff).

Das Ergebnis zeigt, dass 21 Prozent der gesamten Teilnehmer im Jahr 1965 eine mittlere Management Position (Level 3-4) erreicht hatten. 52 Prozent erreichten die zweite Ebene und 27 Prozent blieben auf dem ersten Level (Bray & Grant, 1966, S. 17). Es wird ersichtlich, dass sich die Collegeabsolventen schneller entwickelten als die Teilnehmer ohne Collegeabschluss. 30 Prozent der Collegeabsolventen erreichten am Ende der Studie eine mittlere Management Position, bei den Teilnehmern ohne Abschluss hingegen nur 13 Prozent. Noch gravierender war der Unterschied auf dem ersten Level. 45 Prozent der Kandidaten ohne Abschluss sind auf dem gleichen Niveau geblieben, bei den Collegeabsolventen lag diese Zahl nur bei sechs Prozent (Bray & Grant, 1966, S. 17). Was die zu

Beginn der Studie abgegebenen Prognosen betraf, wird ersichtlich, dass von den 62 prognostizierten Teilnehmern mit Collegeabschluss 48 Prozent tatsächlich auch das mittlere Management erreicht haben. Bei den Personen ohne Collegeabschluss war die Prognose deutlich schlechter, so haben nur 32 Prozent der 41 prognostizierten Personen auch eine Position im mittleren Management erreicht. Die Prognose für ein Verbleiben auf dem Level 1 war für die Zielgruppe mit 60 Prozent der 103 vorausgesagten Personen deutlich besser (Bray & Grant, 1966, S. 17f).

Seit der *Management Progress Study* wurden eine Vielzahl von Studien zur prädiktiven Validität von Assessment Centern durchgeführt, allerdings selten in Europa (Obermann, 2013, S. 213). Moses (1973) führt eine Studie durch, in welcher der Zusammenhang zwischen einem Potential-Assessment Center und einem erneuten Performance-Assessment Center ermittelt wird (S. 572f). Die durchschnittliche Korrelation lag hier bei $\bar{r}$ = .73 zwischen den beiden Assessment Center Ergebnissen, unabhängig von Geschlecht und ethnischer Herkunft (Moses, 1973, S. 574). Dieser Wert ist sehr hoch, was u.a. darauf zurückzuführen ist, dass es sich bei Kriterium und Prädiktor um dieselbe Methode handelt. Hunter und Hunter (1984) ermittelten in einer Metaanalyse eine mittlere Korrelation von $\bar{r}$ = .43 für Assessment Center als prädiktives Instrument für Berufserfolg (S. 91). Klimoski und Brickner (1987) bestätigen die prädiktive Validität von Assessment Centern, weisen jedoch darauf hin, dass es nach wie vor viele ungeklärte Aspekte in Bezug auf die prädiktive Validität gibt, die u.a. auf die mangelnde Konstruktvalidität zurückzuführen sind (Klimoski & Brickner, 1987, S. 256).

Turnage und Muchinsky (1984) untersuchten die Validität von Assessment Centern für 799 Teilnehmer und ermittelten keinen direkten Zusammenhang mit Berufserfolg (S.595). Allerdings konnte ein substanzieller Zusammenhang mit Potential Rankings und Gehaltszuwachs aufgezeigt werden. Darüber hinaus wurde ein eindeutiger prädiktiver Zu-

sammenhang zwischen dem Assessment Center Ergebnis und der Förderung der Kandidaten durch Promotoren ermittelt (r = .14 bis r = .35). Die Förderung der Teilnehmer hing hingegen nicht mit dem tatsächlichen Berufserfolg zusammen. Dies bedeutet, dass die Personen, die ihren Job tatsächlich erfolgreich ausfüllen, nicht unbedingt ge- bzw. befördert werden (Turnage & Muchinsky, 1984, S. 600).

Eine weitere Studie von Goffin, Rothstein und Johnston (1996) untersuchten Assessment Center und Persönlichkeitstests in der Voraussagekraft des Führungserfolgs. Das Ergebnis der Studie war, dass neben Assessment Centern (r = .02 bis r = .40) auch Persönlichkeitstests (r = .02 bis r = .45) ein äquivalentes Instrument sind, den Berufserfolg vorauszusagen (Goffin et al., 1996, S. 746).

Eine Metaanalyse von Gaugler, Rosenthal, Thornton und Bentson aus dem Jahr 1987 bestätigt die Ergebnisse vorangegangener Studien, dass Assessment Center eine prädiktive Validität besitzen. Untersucht wurden 50 Studien zu Assessment Centern, die nach Korrekturen insgesamt 107 Validitätskoeffizienten ergaben (Gaugler, Rosenthal, Thornton & Bentson, 1987a, S. 497). Es wurde differenziert zwischen fünf Kategorien, die sich im Hinblick auf das Zielkriterium unterschieden haben und vier Kategorien, nach der Zielsetzung des Assessment Centers (Gaugler et al., 1987a, S. 494f). Der durchschnittliche, gesamte Validitätskoeffizient, korrigiert um Stichprobenfehler, lag bei $\bar{r}$ = .37. Die durchschnittlichen, korrigierten Validitätskoeffizienten für verschiedene *Assessment Center Zielsetzungen* variierten von $\bar{r}$ = .30 bis zu $\bar{r}$ = .48. Die korrigierten durchschnittlichen Validitätskoeffizienten für die *Vorhersage verschiedener Kriterien* variierten von $\bar{r}$ = .33 bis zu $\bar{r}$ = .53. Zieht man die untere Grenze des 90 prozentigen Vertrauensintervalls für einen durchschnittlichen, korrigierten Validitätskoeffizienten in allen Stichproben in Höhe von $\bar{r}$ = .21 heran, lässt sich auf eine allgemeine Validität von Assessment Centern schließen (Gaugler et al., 1987a, S. 503). Eine ausreichende Varianz wurde erzielt, nachdem eine

Korrektur der Schätzfehler erfolgte und Moderatoren ermittelt wurden. Die Validität stieg, wenn der Anteil an weiblichen Assessoren höher, der Umfang an Beurteilungsinstruktion umfassender, die Beobachter Psychologen statt Manager und die Studien methodisch stabil durchgeführt wurden. Das Alter der Teilnehmer, das Stattfinden eines Feedbackgesprächs, die Anzahl der Tage des Beobachtertrainings, die Dauer des Assessment Centers, der Prozentsatz an Teilnehmern, die einer Minderheit angehören und die Kontamination des Kriteriums stellten keine Moderatoren der Validität dar (Gaugler et al., 1987a, S. 504ff).

Eine Metaanalyse von Arthur, Day, McNelly und Edens aus dem Jahr 2003 fokussiert sich auf die prädiktive Validität einzelner Dimensionen, anstatt auf das Assessment Center-Gesamtergebnis. Eine Fokussierung auf das Gesamtergebnis ist laut der Autoren u.a. deshalb nicht sinnvoll, da es einige Dimensionen gibt, die stärker den Berufserfolg prognostizieren als andere (Arthur et al., 2003, S. 127). Aus 34 Studien, die die Validität bestimmter Dimensionen ermittelten, wurden 168 Assessment Center Dimensionen sieben Dimensionen zugeordnet. Diese Dimensionen lauteten: *consideration / awareness of others*, *communication*, *drive*, *influencing others*, *organizing and planning*, *problem solving* und *tolerance for stress / uncertainty*. Letztgenannte Dimension wurde am Ende aus den weiteren Untersuchungen ausgeschlossen, da bei der Zuordnung aller Assessment Center Dimensionen deutlich wurde, dass bei dieser Dimension zwei verschiedene Konstrukte erfasst wurden (Arthur et al., 2003, S. 132). Basierend auf den verbleibenden sechs Dimensionen ergab sich eine Datenbasis aus 258 Validitätskoeffizienten (Arthur et al., 2003, S. 137). Die Ergebnisse wiesen eine signifikante, kriterienorientierte Validität von $r = .25$ bis $r = .39$ auf. Eine Regressionsanalyse, basierend auf vier der sechs Dimensionen, erklärt mit 20 Prozent einen größeren Teil der Varianz als die Metaanalyse von Gaugler et al. (1987a) mit 15 Prozent (Arthur et al., 2003, S. 125).

Eine Metaanalyse von Meriac, Hoffman, Woehr und Fleisher (2008) greift die Dimensionen von Arthur et al. (2003) auf und bestätigt, dass diese neben den kognitiven Fähigkeiten und Persönlichkeitsfaktoren einen großen Anteil der Varianz von beruflichem Erfolg erklären (Meriac et al., 2008, S. 1042). Die aufgegriffenen Assessment Center-Dimensionen korrelierten beispielweise mit den kognitiven Fähigkeiten zwischen r = .20 und r = .32 ($\bar{r}$ = .27) (Meriac et al., 2008, S. 1047).

Betrachtet man die Korrelationen der Assessment Center-Dimensionen mit den Big Five Persönlichkeitsfaktoren *Extraversion, Gewissenhaftigkeit, Neurotizismus, Verträglichkeit* und *Offenheit für neue Erfahrungen*, variierten die Werte zwischen r = -.09 und r = .24 ($\bar{r}$ = .07) (Meriac et al., 2008, S. 1047). Darüber hinaus konnte durch das Hinzufügen der Assessment Center-Dimensionen zu den kognitiven Fähigkeiten und den Persönlichkeitsfaktoren eine signifikante Erhöhung der Varianz des beruflichen Erfolgs von 20 bis 30 Prozent erzielt werden. Des Weiteren bestätigen die Autoren für sechs der sieben Dimensionen von Arthur et al. (2003) einen, wenn auch sehr geringen, signifikanten (r < .05) Zusammenhang mit Berufserfolg (Meriac et al., 2008, S. 1048).

Auch Dilchert und Ones (2009) setzen sich in ihrer Untersuchung mit den sieben Assessment Center-Dimensionen von Arthur et al. (2003) auseinander. Ein großer Vorteil dieser Untersuchung ist, dass die Studie auf eine weitaus höhere Datenbasis (N = 4.985) zurückgreift, als vorangegangene Metaanalysen. Die Autoren lieferten die ersten stabilen Werte für einen Zusammenhang der Assessment Center-Dimensionen mit den allgemeinen kognitiven Fähigkeiten und den Big Five Persönlichkeitsfaktoren (Dilchert & Ones, 2009, S. 254). Addiert man die genannten Assessment Center-Dimensionen von Arthur et al. (2003) zu den Big Five Persönlichkeitsfaktoren erhöht sich die Prognose für Berufserfolg auf r = .47. Eine Addition zu den kognitiven Fähigkeiten steigert die Validität auf r= .68.

Kombiniert man alle drei Prädiktoren miteinander, lässt sich sogar eine operationale Validität von $r = .71$ erzielen (Dilchert & Ones, 2009, S. 261f).

Während die skizzierten Metaanalysen sich in erster Linie auf den englischsprachigen Raum beschränken, liefert die Studie von Becker, Höft, Holzenkamp und Spinath (2011) erste systematische metaanalytische Ergebnisse für die prädiktive Validität von Assessment Centern in deutschsprachigen Regionen (Becker et al., 2011, S. 61). Es wurden insgesamt 24 Validitätskoeffizienten aus 19 Studien ($N = 3.556$) erfasst, die eine korrigierte, durchschnittliche Validität in Höhe von $\bar{r} = .40$ ergaben. Bei einem Vertrauensintervall von 80 Prozent variierten die Werte zwischen $r = .24$ und $r = .56$. Die Zielgruppe (intern versus extern), das Durchschnittsalter der Kandidaten, der Anteil an Intelligenztests, die Anzahl der eingesetzten Instrumente, die Dauer des Assessment Centers und der Zeitraum zwischen dem Assessment Center und dem Zielkriterium moderierten die Validität (Becker et al., 2011, S. 61).

Insgesamt wird ersichtlich, dass die prädiktive Validität von Assessment Centern eher begrenzt ist und es bestimmte Assessment Center-Dimensionen (Arthur et al., 2003; Dilchert & Ones, 2009; Meriac et al., 2008) gibt, deren Prognosekraft höher ist als bei anderen. Daneben kann ein Hinzuziehen der kognitiven Fähigkeiten und bzw. oder von Persönlichkeitstests die Validität erhöhen (Dilchert & Ones, 2009; Meriac et al., 2008).

Neben der prädiktiven Validität gibt es noch zwei weitere Parameter, die Einfluss auf die Trefferquote haben, nämlich die *Selektionsrate* und die *Basisrate* (Obermann, 2013, S. 282). Die *Selektionsrate* ist definiert als die Anzahl der Personen, die in Relation zu der Gesamtteilnehmeranzahl im Assessment Center ausgewählt werden sollen. Die Trefferquote ist bei ansonsten konstanter Validität desto besser, je niedriger die Selektionsrate ist. D.h. wenn ohnehin nur die Besten aus einer Vielzahl von Bewerbern ausgewählt werden sollen, kann die Güte des Verfahrens relativ gesehen etwas schwächer sein. Müssen hinge-

gen nahezu alle Teilnehmer ausgewählt werden, dann muss die Validität höher sein, um eine vergleichbare Trefferquote zu erzielen (Obermann, 2013, S. 282). Es gibt zwei Möglichkeiten, die Selektionsrate zu beeinflussen. Zum einem kann ein hoher Cut-Off-Wert festgelegt werden, zum anderen können verhältnismäßig mehr Bewerber eingeladen werden. Die zweite Strategie besteht darin, durch eine Erhöhung der Anzahl an Assessment Center-Teilnehmern die Selektionsrate zu reduzieren und somit die Trefferwahrscheinlichkeit zu verbessern (Obermann, 2013, S. 282).

Zweiter Einflussfaktor auf die Trefferquote ist die *Basisrate*. Hierunter wird der Anteil der vermutlich geeigneten Personen verstanden. Mit einer höheren Basisrate kann die Trefferquote unter sonst gleichen Bedingungen gesteigert werden. Ein hochvalides Assessment Center hat aufgrund einer guten „Bewerberqualität" nur einen geringen Zusatznutzen. Die Basisrate kann durch eine höherwertigere Vorauswahl der Teilnehmer beeinflusst werden (Obermann, 2013, S. 282).

Das Multimodale Auswahlverfahren

Aufgrund der aufgeführten Herausforderungen in Bezug auf die Konstruktvalidität und einer eher geringen prädiktiven Validität von Assessment Centern als alleiniges Auswahlinstrument, hat sich ein Trend weg vom klassischen Assessment Center hin zu einem Multimodalen Auswahlverfahren entwickelt (Hufnagl, 2001, S. 16).

„Der Begriff Multimodales Auswahlverfahren leitet sich vom wissenschaftlichen Begriff „Multimodales Interview" ab und bedeutet in der Praxis, dass unterschiedlichste Einzelmodule in ein Bewerberauswahlverfahren integriert sind" (Hufnagl, 2001, S. 17). In den meisten Multimodalen Auswahlverfahren werden ein strukturiertes Interview, eine Präsentation sowie häufig ein Intelligenztest eingesetzt (Hufnagl, 2001, S. 17). Diese drei Bausteine werden auch in der späteren empirischen Untersuchung berücksichtigt.

In den nachfolgenden Unterkapiteln wird zuerst die prädiktive Validität von verschiedenen Personalauswahlinstrumenten alleine und im Anschluss in Kombination miteinander betrachtet. Aufgrund des besonderen Stellenwertes als Prädiktor für beruflichen Erfolg wird das Konstrukt *Intelligenz* genauer betrachtet (Amelang & Schmidt-Atzert, 2006, S. 200). Kurz erläutert ist das strukturierte Interview, da es nach wie vor ein sehr verbreitetes Selektionsinstrument und Bestandteil der meisten Multimodalen Auswahlverfahren ist (Amelang & Schmidt-Atzert, 2006, S. 17) .

Die prädiktive Validität von ausgewählten Personalselektionsinstrumenten einzeln und in Kombination miteinander betrachtet

Untersuchungen über Personalauswahlverfahren als Prädiktoren für späteren Berufserfolg und die Lernfähigkeit, beispielsweise im Rahmen von berufsbezogenen Trainingsprogrammen, gibt es seit dem ersten Jahrzehnt des 20sten Jahrhunderts.

In den 20er Jahren wurde schnell klar, dass die meisten Studien unterschiedliche Ergebnisse erzielen. So entwickelte sich in den 30er und 40er Jahren die Theorie, dass die Unterschiede auf die unterschiedlichen Anforderungen von Jobs zurückzuführen sind und folglich auch verschiedene, spezifische Auswahlverfahren notwendig sind (Schmidt & Hunter, 1998, S. 264).

In den 70ern entdeckte man dann, dass ein Großteil der unterschiedlichen Erkenntnisse auf statistische Fehler und Messfehler bzw. Messartefakte zurückzuführen sind (Schmidt & Hunter, 1977; Schmidt, Hunter, Pearlman & Shane, 1979). Das größte Artefakt wurde durch Stichprobenfehler hervorgerufen. So umfasste ein Großteil der Studien zu kleine Stichproben, oftmals wurden nur 40-70 Angestellte betrachtet (Schmidt & Hunter, 1998, S. 264).

Diese Erkenntnisse führten dazu, dass Metaanalysen durchgeführt wurden, die die Validitäten einzelner Studien zusammenfassen, um statistische und Messfehler beheben zu können. Die Ergebnisse der Metaanalyse von Schmidt und Hunter aus dem Jahre 1998 sind im nachfolgenden genauer beschrieben.

Schmidt und Hunter (1998) haben in ihrer Metaanalyse die Ergebnisse der letzten 85 Jahre Forschung zusammengefasst und neunzehn verschiedene Personalselektionsinstrumente als Prädiktor für Berufs- und Ausbildungserfolg untersucht. Zudem haben sie, aufgrund des hohen Stellenwertes für die Prognose von Berufserfolg, die generellen kognitiven Fähigkeiten als Hauptselektionsinstrument in Kombination mit anderen Instrumenten im Besonderen betrachtet. Berufserfolg wird in den vorliegenden Studien meistens durch die Vorgesetztenbewertung gemessen, teilweise auch durch Produktions- oder Verkaufsergebnisse (Schmidt & Hunter, 1998, S. 264).

Tabelle 3 zeigt einen Ausschnitt der Validitäten ausgewählter Personalselektionsinstrumente, alleine und in Kombination mit den generellen kognitiven Fähigkeiten. Die wichtigsten Ergebnisse sind im Anschluss genauer beschrieben.

Tabelle 3

Predictive Validity for Overall Job Performance of General Mental Ability (GMA) Scores Combined With a Second Predictor Using (Standardized) Multiple Regression (eigene Darstellung in Anlehnung an (Schmidt & Hunter, 1998, S. 265))

Personnell Measures	Validity (r)	Gain in validity from adding supplement	Standardized regression weights	
			GMA	Supplement
GMA tests	.51			
Work sample test	.54	.12	.36	.41
Integrity tests	.41	.14	.51	.41
Employment interviews (structured)	.51	.12	.39	.39
Employment interviews (unstructured)	.38	.04	.43	.22
Job knowledge tests	.48	.07	.36	.31
Job tryout procedure	.44	.07	.40	.20
Job experience (years)	.18	.03	.51	.18
Assessment Centers	.37	.02	.43	.15

Mit r = .54 liegt die Arbeitsprobe auf Platz Eins und besitzt folglich die höchste prognostische Validität der betrachteten Instrumente. Jedoch ist deren Einsatz in der Unternehmenspraxis beschränkt auf Personen, die bereits Berufserfahrung aufweisen oder die Inhalte der Position zuvor kannten und daraufhin trainiert wurden. Darüber hinaus verursacht diese

aufgrund des hohen spezifischen Vorbereitungsaufwands enorm hohe Kosten (Schmidt & Hunter, 1998, S. 267ff).

Das strukturierte Interview liegt mit einer Validität von r = .51 mit den allgemeinen kognitiven Fähigkeiten auf Platz Zwei. Assessment Center liegen mit einer Validität von r = .37 unverkennbar darunter und sind auch deutlich kostspieliger für Unternehmen (Schmidt & Hunter, 1998, S. 267ff).

General Mental Ability (GMA) wird im Folgenden als die allgemeinen kognitiven Fähigkeiten oder *Intelligenz* betitelt. Diese nimmt, wie bereits angedeutet, eine Sonderrolle unter den untersuchten Instrumenten ein. Ein Grund dafür ist, dass Intelligenztests für alle Berufe – egal welches Qualifikationsniveau bzw. Schwierigkeitsgrad – eingesetzt werden können und mit die geringsten Bewerbungs- bzw. Rekrutierungskosten verursachen. Zudem liegt die Validität mit r = .51. höher als bei vielen anderen betrachteten Selektionsinstrumenten und ist auf einem vergleichbaren Niveau wie die Arbeitsprobe und das strukturierte Interview (Schmidt & Hunter, 1998, S. 264).

Für Intelligenztests gibt es des Weiteren die meisten durchgeführten Untersuchungen. In der hier vorliegenden Metaanalyse von Schmidt und Hunter (1998) wurden mehr als 32.000 Angestellte in 515 sehr verschiedenen Ziviljobs in den USA untersucht. Bereits seit den 80er Jahren wurden in den USA immer wieder Studien zur prädiktiven Validität von General Mental Ability (GMA) veröffentlicht (Hunter & Hunter, 1984; Schmidt, Hunter & Pearlman, 1981). All diese Studien zeigen auf, dass die allgemeinen kognitiven Fähigkeiten valide Prädiktoren für Berufserfolg und Ausbildungserfolg sind. Beispielsweise wiesen Hunter und Hunter (1984) eine korrigierte, durchschnittliche prädiktive Validität von $\bar{r}$ = .53 für beruflichen Erfolg nach (S. 81).

Ein weiterer bedeutender Grund für den besonderen Stellenwert von Intelligenztests im Rahmen der Personalauswahl ist, dass Intelligenz die besten theoretischen Grundlagen

bzw. Konstrukte besitzt. Intelligenztheorien wurden von Psychologen seit Anfang des 20ten Jahrhunderts entwickelt und getestet. Folglich ist eindeutiger, was durch Intelligenz gemessen wird, als durch ein Interview oder Assessment Center (Schmidt & Hunter, 1998, S. 264).

Aufgrund der Sonderstellung von Intelligenz empfiehlt es sich, Intelligenztests als das primäre Selektionsinstrument für Einstellungsentscheidungen zu nutzen und weitere Instrumente ergänzend einzusetzen, um die gesamte Validität zu steigern (Schmidt & Hunter, 1998, S. 266). Die drei Kombinationen mit der höchsten multivariaten Validität und Prognosekraft für beruflichen Erfolg sind wie in Tabelle 3 (siehe Seite 35) dargestellt:

- GMA und eine Arbeitsprobe
- GMA und ein Integritätstest
- GMA und ein strukturiertes Interview (Schmidt & Hunter, 1998, S. 265).

Der höchste Wert wird durch eine Kombination eines Intelligenztests gepaart mit einem Integritätstest erzielt. Integritätstests messen in erster Linie die Persönlichkeitseigenschaft *Gewissenhaftigkeit* und bedeuten übertragen auf die Arbeitswelt, dass Personen mit einem hohen Wert gewissenhaft ihren Job erledigen und oft länger am Arbeitsplatz bleiben. Der hohe Wert kommt u.a. dadurch zustande, dass Integritätstests im Gegensatz zur Arbeitsprobe nicht mit den allgemeinen kognitiven Fähigkeiten korrelieren. (Schmidt & Hunter, 1998, S. 267,272).

Eine Steigerung der prädiktiven Validität um $r = 12$ bzw. 24 Prozent kann durch ein Hinzuziehen einer Arbeitsprobe zu einem Fähigkeitstest erzielt werden. Allerdings kann die Arbeitsprobe nur bei Personen mit Berufserfahrung eingesetzt werden (Schmidt & Hunter, 1998, S. 267).

Fügt man ein strukturiertes Interview zu den allgemeinen kognitiven Fähigkeiten hinzu, kann die prädiktive Validität um den gleichen Wert erhöht werden wie bei einer Arbeitsprobe. Strukturierte Interviews haben den Vorteil, dass diese auch für Personen ohne Berufserfahrung eingesetzt werden können (Schmidt & Hunter, 1998, S. 267).

Die Berufserfahrung in Jahren als Zusatz zu den allgemeinen kognitiven Fähigkeiten spielt eine unterschiedliche Rolle. Bei bis zu fünf Jahren Berufserfahrung erhöht sich der Berufserfolg linear mit steigender Erfahrung im Job. Nach diesem Zeitraum wird die Kurve stetig horizontaler, was heißt, dass eine höhere Berufserfahrung in Jahren die berufliche Leistung nur noch geringfügig erhöht. Dies bedeutet folglich, dass sogar unter idealen Umständen die Länge der Berufserfahrung nur für die ersten fünf Jahre als Prädiktor für Berufserfolg herangezogen werden kann (Schmidt & Hunter, 1998, S. 269).

Assessment Center als Zusatz zu den allgemeinen kognitiven Fähigkeiten haben wie die biografischen Daten keine zusätzliche Steigerung der inkrementellen Validität zu Folge. Der Grund dafür ist, dass Assessment Center an sich bereits hoch mit den allgemeinen kognitiven Fähigkeiten korrelieren, da sie typischerweise schon einen hohen Anteil der allgemeinen kognitiven Fähigkeiten im Konstrukt haben. Jedoch sind die erzielten Ergebnisse in einem Assessment Center ein guter Prädiktor für das relevante Wissen zur Ausführung der zukünftigen Jobposition (Schmidt & Hunter, 1998, S. 269).

Unter Betrachtung verschiedener Komplexitätsgrade der auszuführenden beruflichen Tätigkeiten liegt die Validität von den allgemeinen kognitiven Fähigkeiten für Managementpositionen bei $r = .58$, für hochkomplexe technische Berufe bei $r = .56$, für mittelmäßig komplexe Jobs bei $r = .51$, für geringqualifizierte bei $r = .40$ und für unqualifizierte Tätigkeiten bei $r = .23$. Dies bestätigt die Ergebnisse von Hunter und Hunter (1984), dass die Komplexität der auszuführenden Tätigkeit ein Moderator der allgemeinen kognitiven Fähigkeiten ist (Hunter & Hunter, 1984; Schmidt & Hunter, 1998, S. 264).

Aufgrund der soeben skizzierten, hohen Bedeutung der allgemeinen kognitiven Fähigkeiten für den Berufserfolg beschäftigen sich die nächsten Kapitel genauer mit dem Konstrukt Intelligenz. Zum besseren Verständnis der Bedeutung erfolgen zunächst eine Definition und sodann die Beschreibung eines möglichen Messinstruments. Darüber hinaus werden im nächsten Schritt einzelne Unterfacetten von Intelligenz und deren Zusammenhang mit beruflichem Erfolg betrachtet.

Intelligenz – Definition, Messinstrumente und Unterfacetten.

Ein Faktor, der lange Zeit sogar als alleiniges Merkmal für beruflichen Erfolg über alle Berufsgruppen hinweg galt, ist Intelligenz. Heutzutage gilt Intelligenz nicht als die einzige, aber immer noch als besonders wichtige Eigenschaft für beruflichen Erfolg (Achouri, 2011, S. 85). Die allgemeinen kognitiven Fähigkeiten werden in der Regel durch einen Intelligenzquotienten (IQ) angegeben und durch einen Intelligenztest gemessen (Amelang & Schmidt-Atzert, 2006, S. 200).

Intelligenztests gehören in der psychologischen Diagnostik aufgrund der hohen Prognosewerte in wichtigen Lebensbereichen und stabilen Werten im Zeitverlauf zu den erfolgreichsten Verfahren. Die Tearman Studie hat beispielsweise ergeben, dass bei weiterer Verfolgung des Lebenslaufes von 1.400 Kindern, diejenigen am erfolgreichsten waren, die eine überdurchschnittliche Intelligenz mit einem IQ von mindestens 135 aufwiesen (Amelang & Schmidt-Atzert, 2006, S. 200). Die Korrelationen mit Schul-, Ausbildungs- und Berufserfolg liegen wie bereits im vorherigen Kapitel genauer beschrieben im Bereich von $r = .50$ (Schmidt & Hunter, 1998).

Intelligenztests messen zum einen das allgemeine Intelligenzniveau und zum anderen spezifische Intelligenzfaktoren, wie z.B. räumliches Vorstellungs- und Abstraktionsvermögen, sprachliches, rechnerisches und logisches Denken. Das Anforderungsniveau der Stelle

bestimmt in der Regel, welche Inhalte im Rahmen von Intelligenztests zum Einsatz kommen (Hufnagl, 2001, S. 133).

Die Kriterienvalidität für spezifische kognitive Fähigkeiten wurde in den USA insbesondere im militärischen Umfeld geprüft. Fokussiert man sich auf spezifische Fähigkeiten, hat dies den Vorteil, dass Unterschiede in der Gruppe reduziert und positivere Reaktionen bei den Bewerbern hervorgerufen werden. Aus diesem Grunde wurde dann auch im beruflichen Kontext in den USA im Rahmen von Untersuchungen die prädiktive Validität von spezifischen Fähigkeiten wie verbale, numerische und räumliche Fähigkeiten betrachtet (Salgado et al., 2003, S. 574f).

Diese Ergebnisse ohne weiteres auf den europäischen Raum zu übertragen, erschien Salgado et al. (2003) aufgrund von unterschiedlichen Charakteristika für die Europäische Union nicht uneingeschränkt möglich. Aus diesem Grunde haben die Autoren eine Metaanalyse für den europäischen Raum durchgeführt, mit der Zielsetzung, die US-Amerikanischen Ergebnisse zu internationalisieren (Salgado et al., 2003, S. 574f).

Fokus der Studie von Salgado et al. (2003) war es, Ergebnisse zu liefern, die einerseits quantitative Aussagen zum Zusammenhang zwischen den allgemeinen kognitiven Fähigkeiten und Berufs- sowie Ausbildungserfolg liefern. Andererseits soll das Ausmaß der prädiktiven Validität von spezifischen Fähigkeiten für die Vorhersage von beruflichem und Ausbildungserfolg bestimmt werden (Salgado et al., 2003, S. 578). Folgende Kriterien haben die zu berücksichtigten Studien im Rahmen der Metaanalyse bei Salgado et al. (2003) gemeinsam:

1. Es muss ein valider Zusammenhang zwischen Berufs- und Ausbildungserfolg und den kognitiven Fähigkeiten bestehen.
2. Als Zielgruppe wurden Bewerber, Angestellte oder Trainees und keine Studenten gewählt.
3. Anders als in den US-Amerikanischen Studien werden nur zivile Erwerbstätige und keine Militärs berücksichtigt (Salgado et al., 2003, S. 579).

Insgesamt umfasst die Metaanalyse von Salgado 142 beispielhafte Studien mit dem Kriterium *Ausbildungserfolg* und 120 mit dem Kriterium *Berufserfolg*. Beide Kriterien sind in 14 Studien berücksichtigt. Die meisten Studien wurden im United Kingdom durchgeführt (Salgado et al., 2003, S. 579). Die wichtigsten Ergebnisse dieser Metaanalyse werden im nachfolgenden genauer beschrieben. Der Fokus liegt auf Berufserfolg, da dies die Basis für die im Rahmen dieser Arbeit durchgeführte empirische Untersuchung ist.

Im Fall von Berufserfolg wurde in der Mehrheit der Studien, inklusive der von Salgado et al. (2003), Beurteilungen von Trainern oder Vorgesetzten als Zielkriterium verwendet. Tabelle 4 zeigt die wichtigsten Ergebnisse auf.

Tabelle 4

Meta-Analysis of General Mental Ability and other Cognitive Ability Tests for Predicting Job Performance Ratings (eigene Darstellung in Anlehnung an (Salgado et al., 2003, S. 586)

Source	K	N	ρ	%VE	90% CV
GMA	93	9,554	.62	75	.37
Verbal	44	4,781	.35	53	.04
Numerical	48	5,241	.52	100	.52
Spatial-Mechanical	40	3,750	.51	52	.13
Perceptual	38	3,789	.52	73	.28
Memory	14	946	.56	100	.56

Anmerkung. K = Number of studies; N = total sample; ρ = operational validity; %VE = variance accounted for by artifactual errors; 90% CV = credibility value.

Die operationale Validität für die allgemeinen kognitiven Fähigkeiten zeigt mit einem Wert von $r = .62$, das kognitive Fähigkeitstests ein hervorragender Prädiktor für Berufserfolg sind. Der hier aufgeführte Wert ist der bis dahin in der Literatur höchste und bestätigt die Ergebnisse von Schmidt und Hunter aus dem Jahr 1998, dass die allgemeinen kognitiven Fähigkeiten der beste Prädiktor für beruflichen Erfolg sind. Das 90 prozentige Vertrauensintervall impliziert, dass die allgemeinen kognitiven Fähigkeiten eine generelle Validität über alle Beispiele, Berufsarten, Messungen und europäische Länder für das Kriterium Berufserfolg haben (Salgado et al., 2003, S. 585f).

Bei Betrachtung einzelner Facetten von Intelligenz ist die Merkfähigkeit nach den allgemeinen kognitiven Fähigkeiten mit $r = .56$ der zweitbeste Prädiktor für Berufserfolg. Die gesamte Varianz wurde hier durch künstliche Fehler erklärt. Folglich ist auch die Merkfähigkeit ein genereller Prädiktor für Berufserfolg über alle Beispiele, Berufe und Länder hinweg (Salgado, 2003a, S. 586).

Bei Betrachtung der verbalen und numerischen Fähigkeiten kommen unterschiedliche Ergebnisse heraus. Die operationale Validität liegt bei den verbalen Fähigkeiten lediglich bei $r = .35$, die erklärte Varianz nur bei 53 Prozent. Beigefügte Zusätze lassen darauf schließen, dass die Validität durch eine andere Variable moderiert wurde. Ein Grund dafür kann die unterschiedliche Sprache in den unterschiedlichen Ländern in der vorliegenden Studie sein (Salgado, 2003a, S. 587).

Im Falle der numerischen Fähigkeiten liegt der Wert bei $r = .52$, die beobachtete Varianz wurde komplett durch künstliche Fehler erklärt. Salgado et. al (2003) führen an, dass der Wert für die numerischen Fähigkeiten mit großer Wahrscheinlichkeit höher ist und hier unterschätzt wird, da die Stichprobe für diesen Test mit Abstand die geringste war. Folglich sind auch die numerischen Fähigkeiten ein guter Prädiktor für beruflichen Erfolg und haben Gültigkeit über die aufgeführten Beispiele und Länder hinweg (Salgado et al., 2003, S. 587f).

Zusammengefasst liegt das Ausmaß der prädiktiven Validität in den europäischen Metaanalysen deutlich über dem der amerikanischen Metaanalysen (Salgado et al., 2003). Der Wert für die allgemeinen kognitiven Fähigkeiten ist sogar höher als in vielen Metaanalysen durchleuchteten Personalselektionsinstrumenten. Insgesamt lässt sich auf eine internationale Generalisierung der Validität für kognitive Fähigkeitstests in den USA und der Europäischen Union schließen (Salgado et al., 2003, S. 589).

Die Ergebnisse für Ausbildungserfolg ergeben ähnlich hohe Werte wie bei Berufserfolg, sind allerdings im Gesamten etwas niedriger. Der höchste Wert für allgemeine kognitive Fähigkeitstests liegt bei r = .54. Die verbalen und numerischen Fähigkeiten lieferten mit r = .44 und r = .48 ähnlich hohe Ergebnisse. Bei einem 90 prozentigem Vertrauensintervall kann auch hier auf eine Generalisierung für die gesamte EU geschlossen werden. Wahrnehmung und Merkfähigkeit erzielten hier die geringsten Ergebnisse mit r = .25 und r =. 34.

Ähnlich wie bei den Ergebnissen für Berufserfolg kann eine internationale Generalisierung für die operationale Validität der allgemeinen kognitiven Fähigkeiten und Ausbildungserfolg geschlussfolgert werden. Das Gleiche gilt für die verbalen und numerischen Fähigkeiten (Salgado et al., 2003, S. 591).

Der gravierendste Unterschied zu den amerikanischen Studien ist, dass die allgemeinen kognitiven Fähigkeiten in der Europäischen Union den Berufserfolg stärker prognostizieren als den Ausbildungserfolg, die amerikanischen Studien hingegen kommen zu gegenteiligem Ergebnis. Der Unterschied ist aller Wahrscheinlichkeit auf eine unterschiedliche Definition und unterschiedliche Messweise der aufgeführten Zielkriterien zurückzuführen (Salgado et al., 2003, S. 593).

Hülsheger, Maier und Stumpp (2007) bestätigen im Rahmen einer deutschen Metaanalyse einen ähnlich hohen Wert zwischen den allgemeinen Fähigkeiten und Berufserfolg mit r = . 53 und einem r = .47 für Ausbildungserfolg (Hülsheger et al., 2007, S. 8). Auch die Metaanalyse von Hülsheger et al. (2007) gibt Implikationen für die internationale Generalisierung zur Validität von kognitiven Fähigkeitstests sowie für den praktischen Einsatz von kognitiven Fähigkeitstests als Personalauswahlinstrument.

Im nächsten Kapitel wird ein möglicher Test zur Messung der Intelligenz erläutert, der auch im Rahmen der empirischen Untersuchung eingesetzt wird.

Der IST-2000-R

Ein möglicher Test zur Messung von Intelligenz ist der IST-2000-R. Beim IST-2000R handelt es sich um eine Weiterentwicklung des in Deutschland früher am häufigsten angewandten Intelligenztests, dem IST-70. Die Weiterentwicklung des IST-2000-R besteht aus einer größeren Normierungsstichprobe, technischen Verbesserungen sowie einer Überarbeitung des Wissenstests. Der Test liegt in zwei Formen – Form A und B – vor, wobei die zweite ein Paralleltest zur ersten ist, exakt gleich aufgebaut und die gleichen Fähigkeiten messend. Seit 2007 liegt für das Grundmodul auch ein echter Paralleltest in Form C vor (Amelang & Schmidt-Atzert, 2009, S. 216f).

Inhaltlich erfasst der Test fünf der sieben Primärfaktoren von Thurstone – die verbale, numerische und figurale Intelligenz, Merkfähigkeit und schlussfolgerndes Denken (Reasoning) als Summenscore der drei erstgenannten Faktoren. Das Erweiterungsmodul enthält Wissensfragen verbaler, numerischer und figuraler Art (Amelang & Schmidt-Atzert, 2009, S. 216).

Die Autoren greifen das Modell von Horn und Catell (1966) auf, das die beiden Generalfaktoren fluide und kristallisierte Intelligenz unterscheidet. Bei der fluiden Intelligenz handelt es sich um die Fähigkeit, neuen Problemen oder Situationen gerecht zu werden, wobei frühere Lernerfahrungen hierbei keine bedeutende Rolle spielen. Die kristallisierte Intelligenz hingegen vereinigt kognitive Fähigkeiten, die aus kumulierten Effekten vergangenen Lernens bestehen. Mit dem Grundmodul wird ohne Merkfähigkeit eher die fluide und mit dem Erweiterungsmodul eher die kristallisierte Intelligenz erfasst (Amelang & Schmidt-Atzert, 2009, S. 216)

Wie die Zahlen zeigen, liegt die innere Konsistenz des Grundmoduls mit einem Wert von $\alpha = .96$ in einem sehr hohen Bereich (Amelang & Schmidt-Atzert, 2009, S. 219). Auch

die einzelnen Module weisen ein α über .80 auf. Zur Retest-Reliabilität liegen noch keine Daten vor (Amelang & Schmidt-Atzert, 2009, S. 218).

Die Normen basieren auf der Vorgabe des Grundmoduls. Die Stichprobe N = 3.484 Personen für die Form A und B sowie2.363 für die Form C. Normentabellen liegen für Gymnasiasten, Nicht-Gymnasiasten und die Gesamtgruppe vor. Am differenziertesten ist die Einteilung in acht Altersspannen, beginnend mit 15-16 und endend mit >50 (Amelang & Schmidt-Atzert, 2009, S. 219).

Insgesamt handelt es sich beim IST-2000-R um einen sehr sorgfältig konstruierten Test, der eine reliable Erfassung der fünf Primärfaktoren der Intelligenz und der beiden Sekundärfaktoren der fluiden und kristallisierten Intelligenz zulässt (Amelang & Schmidt-Atzert, 2009, S. 219).

Eine Rezension des Testkuratoriums nach dem Testbeurteilungssystem ermittelt insgesamt eine gute Bewertung. Dennoch wurden die Anforderungen an Reliabilität und Validität nicht als *voll*, sondern lediglich als *weitgehend erfüllt* eingestuft (Amelang & Schmidt-Atzert, 2009, S. 219).

Das strukturierte Interview.

Im Gegensatz zu den Ergebnissen der vorhandenen Studien und Metaanalysen zur prädiktiven Validität von Personalselektionsinstrumenten ist das Interview nach wie vor eines der häufigsten in Unternehmen eingesetzte und meist geschätzte Personalauswahlinstrument seitens der Bewerber und der Unternehmen (Achouri, 2011, S. 84). Wie bereits skizziert, ist die Validität mit r = .38 geringer als bei anderen Selektionsinstrumenten. Gesteigert werden kann dieser Wert auf r = .51, wenn man strukturierte Interviews durchführt (Schmidt & Hunter, 1998, S. 265). Dies bedeutet, die Interviews werden um situative Fra-

gen im Sinne von Wissensarbeitsproben und komplexen Biografie bezogenen Fragen erweitert (Achouri, 2011, S. 84).

Beim strukturierten Interview existiert im Vergleich zum unstrukturierten Interview ein Leitfaden. Der Bewerber wird mit bestimmten, situativen Fragen, die für die zu besetzende Position relevant sind, konfrontiert (Hufnagl, 2001, S. 104). Immer häufiger werden in der Unternehmenspraxis im Rahmen des strukturierten Interviews situative Fallbeispiele eingesetzt, die sich wiederum an den Anforderungsprofilen orientieren. In der Regel wird ein solches durch den Interviewer vorgelesen und zugleich eine Kopie dem Bewerber zum Mitlesen ausgeteilt (Hufnagl, 2001, S. 122). Vorteile des strukturierten Interviews sind eine hohe Objektivität und Vergleichbarkeit zwischen den Bewerbern sowie die Förderung einer bewerberorientierten Vorgehensweise (Hufnagl, 2001, S. 104).

Normen und Anforderungen an diagnostische Verfahren

Dieses Kapitel stellt zu berücksichtigende Normen aus der Wissenschaft an diagnostische Verfahren - im speziellen an Assessment Center - dar.

Initiative ergriff hierzu der Berufsverband Deutscher Psychologinnen und Psychologen (BDP), unterstützt durch die Deutsche Gesellschaft für Psychologie (DGPs). Veröffentlicht wurden die *Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen DIN 33430* im Jahr 2002 (Amelang & Schmidt-Atzert, 2009, S. 453).

Zweck der DIN 33430 ist es, einen Leitfaden für die Planung und Durchführung von Eignungsbeurteilungen für die Anbieter entsprechender Dienstleistungen zu haben. Zudem soll sie Personalverantwortliche unterstützen und eine Qualitätssicherung sowie -optimierung von Personalentscheidungen erzielen. Personen, deren Eignung beurteilt

wird, sollen vor unsachgemäßer oder missbräuchlicher Anwendung von Verfahren geschützt werden (Amelang & Schmidt-Atzert, 2009, S. 454).

Im nachfolgenden werden aus der DIN 33430 abgeleitete Prinzipien und in der Literatur aufgeführten Anforderungen an Assessment Centern skizziert:

- *Anforderungsbezug*: Im Vorfeld ist eine genaue Anforderungsanalyse relevant, damit im Rahmen der Assessment Center Übungen Merkmale oder Verhaltensweisen erfasst werden, die für die zu besetzende Position relevant sind.
- *Simulation:* Im Rahmen der Assessment Center Übungen werden möglichst realistische Situationen simuliert, die sich an den zukünftigen Jobanforderungen orientieren.
- *Methodenvielfalt:* In einem Assessment Center wird jedes Anforderungsmerkmal in verschiedenen Übungen erfasst. Durch die Aggregation der Beurteilungen über mehrere Übungen hinweg gleichen sich die Stärken und Schwächen einzelner Übungen für die Teilnehmer aus und die Reliabilität erhöht sich.
- *Mehrfachbeurteilung:* Die einzelnen Übungen müssen mehrfach und unabhängig voneinander beobachtbar sein, um zuverlässige Beurteilungen vornehmen zu können. I.d.R. wird jeder Teilnehmer von mehreren Personen abwechselnd in verschiedenen Übungen beobachtet und beurteilt. Hierdurch werden Beobachtungs- und Beurteilungsfehler einzelner Personen ausgeglichen. Das Verhältnis der Teilnehmer zu den Beobachtern sollte bei 2:1 liegen.
- *Transparenz:* Beginnend bei der Konstruktion als auch der späteren Durchführung sind die Anforderungen und Beobachtungskriterien den Beobachtern und Teilnehmern im Vorfeld transparent zu machen. Im Anschluss an das Verfahren ist ein individuelles Feedback an die Kandidaten mit Bezug auf das konkrete Anfor-

derungsprofil unabdingbar. Durch diese Offenheit und Transparenz wird die Akzeptanz des Assessment Centers bei den Teilnehmern gestärkt (Amelang & Schmidt-Atzert, 2006, S. 459, 2009, S. 463f; Obermann, 2009, S. 10).

Darüber hinaus gilt für die Auswertung der Beobachtungen eine strikte Trennung zwischen der Beobachtung einerseits und der Beurteilung andererseits. In der ersten Phase werden lediglich Beobachtungen gesammelt und notiert, die in der zweiten Phase dann den Beurteilungskriterien zugeordnet werden (Amelang & Schmidt-Atzert, 2009, S. 466; Obermann, 2013, S. 165).

Fragestellung und Methodik der empirischen Studie

In diesem Kapitel werden die Zielsetzung der vorliegenden empirischen Studie und die Forschungslücke beschrieben. Daran knüpfen eine Beschreibung der Unterschiede in den eingesetzten Auswahlverfahren und die Aufstellung der Hypothesen. Den Abschluss bilden eine Beschreibung des Prüfverfahrens und die Datenerhebung.

Zielsetzung der vorliegenden Untersuchung und Forschungslücke

Diese empirische Studie baut auf den Ergebnissen der im Theorieteil skizzierten Erkenntnisse zu diagnostischen Verfahren – im speziellen zu Assessment Centern und Multimodalen Auswahlverfahren – auf.

Die betrachteten Studien zur Konstruktvalidität von Assessment Centern haben überwiegend deutlich gemacht, dass innerhalb einzelner Übungen die Zusammenhänge der Beurteilungen zu den verschiedenen Dimensionen wenig differenzieren (geringe diskriminante Validität) und die Beurteilungen der gleichen Dimensionen in unterschiedlichen Übungen kaum konsistente Leistungen aufzeigen (geringe konvergente Validität) (Neidig & Neidig, 1984; Sackett & Dreher, 1984; Sackett & Harris, 1988; Turnage & Muchinsky, 1984).

In der vorliegenden empirischen Untersuchung wird genau dieses Problem der Konstruktvalidität von Assessment Centern untersucht. Für zwei Assessment Center, die vom Aufbau und dem Beurteilungsprinzip her sehr ähnlich sind, wird geprüft, ob das in der Literatur beschriebene Problem zutrifft und die verschiedenen Anforderungsdimensionen innerhalb einer Übung hoch miteinander korrelieren, die gleichen Dimensionen zwischen verschiedenen Übungen hingegen kaum miteinander korrelieren. Die beiden Assessment Center sind von den Übungen und vom Aufbau her ähnlich konstruiert und unterscheiden

sich in erster Linie in der Zielgruppe (Trainees und Referenten). Beiden Assessment Centern liegt eine dimensionsbezogene Beurteilung zu Grunde.

Der im Theorieteil betrachtete Unterschied einer dimensionsbezogenen und einer übungsbezogenen Beurteilung wird dann im nächsten Schritt aufgegriffen. Die beiden betrachteten Assessment Center weisen, wie erwähnt, eine dimensionsbezogene Beurteilung auf. In einem neu konzipierten Multimodalen Auswahlverfahren, das für beide Zielgruppen zugleich gelten soll, wird ein Teil der Übungen überarbeitet und ein übungsbezogenes Beurteilungsprinzip eingeführt. Darüber hinaus werden in diesem Verfahren weitere Bausteine wie ein Intelligenztest und ein strukturiertes Interview hinzugefügt. Zielsetzung war es, einen Methodenmix an verschiedenen Auswahlinstrumenten einzusetzen und eine möglichst hohe Trennschärfe der einzelnen Bausteine zu erreichen. Daneben soll die prädiktive Validität des neuen Verfahrens gegenüber den beiden alten Verfahren erhöht werden. Zur Überprüfung der prädiktiven Validität wird die Vorgesetztenbeurteilung in allen drei Fällen als Zielkriterium verwendet.

Vorangegangene Kapitel zur prädiktiven Validität verschiedener Auswahlinstrumente haben deutlich gemacht, dass die Validität von Auswahlverfahren gesteigert werden kann, wenn Intelligenztests mit anderen Bausteinen, wie dem strukturierten Interview oder einer Arbeitsprobe kombiniert werden (Schmidt & Hunter, 1998, S. 265). Eine Limitation der beschriebenen Metaanalyse von Schmidt und Hunter (1998) war, dass die Autoren lediglich eine Kombination von zwei Auswahlinstrumenten untersuchten. Dieser Einschränkung wird in der hier vorliegenden empirischen Studie entgegnet. Wie im Rahmen des Untersuchungsdesigns genauer beschrieben ist, werden im Multimodalen Auswahlverfahren ein Intelligenztest, ein strukturiertes Interview und drei ausgewählte Assessment Center Übungen eingesetzt. Augenmerk liegt bei der Auswahl darauf, dass die eingesetzten

Bausteine zur Zielgruppe passen und eine hohe prädiktive Validität besitzen. Weitere Details dazu werden im Rahmen des Untersuchungsdesigns erläutert.

Unterschiede der eingesetzten Auswahlverfahren und Hypothesen

Trotz der im Theorieteil skizzierten Indizien und Hinweise auf die Konstruktvalidität von übergeordneten Dimensionen konnte kein allumfassender Hinweis für eine Konstruktvalidität von Assessment Centern geliefert werden (Bowler & Woehr, 2006; Gaugler & Thornton, 1989; Kuncel & Sackett, 2014; Shore et al., 1992). Die Untersuchungen zeigten deutlich, dass Beobachter dazu geneigt sind, eher Pauschalurteile innerhalb einer Übung abzugeben, als Merkmale zwischen mehreren Übungen zu beurteilen (Drees, 1994; Neidig & Neidig, 1984; Sackett & Dreher, 1982, 1984; Sackett & Harris, 1988; Turnage & Muchinsky, 1984).

Grundvoraussetzungen für die Konstruktvalidität sind nach Campbell und Fiske (1959) die *konvergente* und *diskriminante* Validität. Von konvergenter Validität spricht man, wenn Messungen der gleichen Merkmale trotz verschiedener Methoden hoch korrelieren. Diese Korrelationen sollten signifikant höher ausfallen als die Korrelationen der Dimensionen innerhalb der Übungen – eine Voraussetzung für die diskriminante Validität (Campbell & Fiske, 1959, S. 81f).

Zieht man die Ergebnisse der betrachteten Studien zur Konstruktvalidität von Assessment Centern und die Kriterien zur Überprüfung von Campbell und Fiske (1959) heran, lassen sich die ersten beiden Hypothesen formulieren. Es liegt die Vermutung nahe, dass innerhalb der in der vorliegenden empirischen Untersuchung eingesetzten Assessment Center (Trainee- und Stabs-Assessment Center) die verschiedenen Anforderungsdimensionen in den gleichen Übungen hoch miteinander korrelieren. Die gleichen Anforderungsdi-

mensionen zwischen verschiedenen Übungen hingegen kaum miteinander korrelieren. Hypothese 1 und 2 lauten wie folgt:

Hypothese 1: *Die Korrelationen der verschiedenen Dimensionen innerhalb einer Übung sind im Trainee-Assessment Center im Durchschnitt deskriptiv höher als die Korrelationen der gleichen Dimensionen über verschiedene Übungen hinweg.*

Hypothese 2: *Die Korrelationen der verschiedenen Dimensionen innerhalb einer Übung sind im Stabs-Assessment Center im Durchschnitt deskriptiv höher als die Korrelationen der gleichen Dimensionen über verschiedene Übungen hinweg.*

Im nächsten Schritt wird das neu konzipierte Multimodale Auswahlverfahren betrachtet. In diesem Verfahren werden die im Rahmen des Untersuchungsdesigns beschriebenen Assessment Center Übungen aus dem bisherigen Trainee-Assessment Center dem Stabs-Assessment eingesetzt. Die eingesetzten Übungen wurden u.a. im Hinblick auf die Beurteilungsart verändert. Im Multimodalen Auswahlverfahren existiert nur mehr eine übungsbezogene und keine dimensionsbezogene Beurteilung. Eine weitere Zielsetzung der Überarbeitung der genannten Assessment Center Übungen und der gesamten eingesetzten Bausteine im Multimodalen Auswahlverfahren war es, möglichst trennscharfe Bausteine zu konzipieren, die möglichst wenigen Überschneidungen aufweisen. Durch die aufgeführten Veränderungen innerhalb dieser Übungen wird die Vermutung angestellt, dass die neuen Bausteine so trennscharf formuliert sind, dass es geringere Überschneidungen gibt als zuvor. Daraus wird Hypothese 3 wie folgt formuliert:

Hypothese 3: „Die eingesetzten Assessment Center-Übungen als Bausteine des Multimodalen Auswahlverfahrens korrelieren signifikant niedriger miteinander als die vergleichbaren Übungen im Trainee-Assessment Center."

Eine im Theorieteil skizzierte Studie von Drees (1994) zur Interrater-Reliabilität von Assessment Centern bei unterschiedlichen Beobachtertrainings hat deutlich gemacht, dass ein verhaltensbezogenes Beobachtertraining die Interrater-Reliabilität gegenüber einem kognitionsbezogenen Beobachtertraining steigern kann. In der Untersuchung wurde ein kognitionsbezogenes Beobachtertraining nach dem bereits skizzierten eigenschaftstheoretischen Ansatz vorangestellt und zum anderen ein verhaltensbezogenes Beobachtertraining nach dem situationstheoretischen Ansatz. Dabei lag die durchschnittliche Interrater-Korrelation beim verhaltensbezogenen Training mit einem durchschnittlichen Wert von $\bar{r}$ = .71 deutlich über dem des kognitionsbezogenen Trainings mit $\bar{r}$ = .43 (Drees, 1994, S. 112).

Dieser Aspekt des unterschiedlichen Beobachtertrainings wird auch in der hier vorliegenden empirischen Untersuchung aufgegriffen. Die beiden betrachteten Assessment Center (Trainee- und Stabs-Assessment Center) weisen eine dimensionsbezogene Beurteilung auf. Im Vorfeld findet ein dimensionsbezogenes Beobachtertraining statt. In dem neu konzipierten Multimodalen Auswahlverfahren findet hingegen eine übungsbezogene Beurteilung statt. Die Beobachter werden im Vorfeld übungsbezogen geschult.

Darüber hinaus müssen die Beobachter innerhalb einer Assessment Center Übung in der Regel bis zu fünf verschiedene Dimensionen bewerten, deren Operationalisierung, wie im Untersuchungsdesign noch näher beschrieben wird, nicht sehr umfänglich, sondern eher global gehalten ist. Aufgrund der schlechten Operationalisierung der Anforderungsdimensionen wird vermutet, dass die Beurteiler eine geringere Übereinstimmung in der Bewer-

tung erzielen, als in einer stark operationalisierten, übungsbezogenen Beurteilung, wie sie im Multimodalen Auswahlverfahren eingesetzt wird.

Aus den soeben aufgeführten Gründen liegt die Vermutung nahe, dass die Übereinstimmung der Beurteilungen zwischen den verschiedenen Beobachtern im Multimodalen Auswahlverfahren höher ausfällt als im Trainee- und Stabs-Assessment Center. Dies mündet in die vierte Hypothese:

Hypothese 4: *Die Interrater-Korrelation ist im Multimodalen Auswahlverfahren deskriptiv höher als im Trainee- und Stabs-Assessment Center.*

Folgende Unterhypothesen sind zur Überprüfung erforderlich:

Hypothese 4.1.: *Die Interrater-Korrelation ist im Multimodalen Auswahlverfahren deskriptiv höher als im Trainee-Assessment Center.*

Hypothese 4.2.: *Die Interrater-Korrelation ist im Multimodalen Auswahlverfahren deskriptiv höher als im Stabs-Assessment Center.*

Eine eher geringe, prädiktive Validität von Assessment Centern (Bray & Grant, 1966; Hunter & Hunter, 1984; Klimoski & Brickner, 1987; Moses, 1973; Turnage & Muchinsky, 1984) – insbesondere gegenüber anderen Personalauswahlinstrumenten, wie Intelligenztests – wurde bereits ausführlich behandelt. Eine Steigerung der prognostischen Validität bei einer Kombination eines Intelligenztests mit andern Instrumenten, wie einem strukturierten Interview, wurde ebenso aufgezeigt (Schmidt & Hunter, 1998).

Die drei Kombinationen mit der höchsten multivariaten Validität und Prognosekraft für beruflichen Erfolg nach Schmidt und Hunter (1998) sind GMA in Kombination mit einer Arbeitsprobe, einem Integritätstest oder einem strukturierten Interview (Schmidt & Hunter, 1998, S. 265). Assessment Center hingegen weisen als einzelnes Instrument lediglich eine prognostische Validität von $r = .37$ auf (Schmidt & Hunter, 1998, S. 265).

In der hier vorliegenden empirischen Untersuchung wird im Multimodalen Auswahlverfahren, wie noch ausführlich beschrieben wird, ein Intelligenztest (verbale und numerische Fähigkeiten), ein strukturiertes Interview und zwei Assessment Center Übungen eingesetzt. Betrachtet man die Validität des strukturierten Interviews mit $r = .51$ (Schmidt & Hunter, 1998, S. 265) und die der beiden Teile des Intelligenztests (verbal: $r = .35$ und numerisch: $r = .53$) (Salgado et al., 2003, S. 586), so sollte folglich bereits eine hohe multivariate Validität existieren. Addiert man noch zwei Assessment Center Übungen hinzu, sollte sich die Validität noch weiter steigern lassen, wenn auch in einem geringen Maße, da die Übungen höchstwahrscheinlich mit den Bausteinen des Intelligenztests korrelieren. Jedoch sind die erzielten Ergebnisse in den eingesetzten Assessment Centern ein guter Prädiktor für das relevante Wissen zur Ausführung der zukünftigen Jobposition (Schmidt & Hunter, 1998, S. 269).

Beruflicher Erfolg wird meistens durch die Leistungsbeurteilung des Vorgesetzten gemessen (Obermann, 2013, S. 292). Auch in der hier vorliegenden empirischen Untersuchung wird die Vorgesetztenbeurteilung als Zielkriterium zur Messung der prädiktiven Validität eingesetzt. Die Vorgesetztenbeurteilung wird im Rahmen der Beschreibung des Untersuchungsdesigns noch genauer erläutert.

Aus den soeben aufgeführten Überlegungen sollte das Multimodale Auswahlverfahren eine bessere Prognose über den beruflichen Erfolg der eingestellten Kandidaten

geben als die zuvor eingesetzten klassischen Assessment Center. Dies führt zur fünften Hypothese, die wie folgt lautet:

Hypothese 5: *Der Zusammenhang des Gesamtdurchschnitts der Vorgesetztenbeurteilung mit dem Gesamtergebnis des Multimodalen Auswahlverfahrens ist signifikant höher ausgeprägt als mit dem Gesamtergebnis des Stabs- und Trainee-Assessment Centers.*

Folgende Unterhypothesen sind zur Überprüfung von Hypothese 5 erforderlich:

Hypothese 5.1: *Im Trainee-Assessment Center existiert ein signifikanter Zusammenhang des Assessment Center-Gesamtergebnisses mit dem Gesamtdurchschnitt der Vorgesetztenbeurteilung.*

Hypothese 5.2: *Im Stabs-Assessment Center existiert ein signifikanter Zusammenhang des Assessment Center-Gesamtergebnisses mit dem Gesamtdurchschnitt der Vorgesetztenbeurteilung.*

Hypothese 5.3: *Im Multimodalen Auswahlverfahren existiert ein signifikanter Zusammenhang des Assessment Center-Gesamtergebnisses mit dem Gesamtdurchschnitt der Vorgesetztenbeurteilung.*

Hypothese 5.4: *Der Zusammenhang des Gesamtdurchschnitts der Vorgesetztenbeurteilung mit dem Gesamtergebnis des Multimodalen Auswahlverfahrens ist signifikant höher ausgeprägt als mit dem Gesamtergebnis des Trainee-Assessment Centers.*

Hypothese 5.5: *Der Zusammenhang des Gesamtdurchschnitts der Vorgesetztenbeurteilung mit dem Gesamtergebnis des Multimodalen Auswahlverfahrens ist signifikant höher ausgeprägt als mit dem Gesamtergebnis des Stabs-Assessment Centers.*

Im nächsten Kapitel werden die statistischen Analysen zur Überprüfung der soeben aufgestellten Hypothesen genauer beschrieben.

Eingesetzte Prüfverfahren und statistische Analysen

Die Auswertung der Daten erfolgte mit dem statistischen Datenauswertungsprogramm *IBM SPSS Statistics 22.*

Zur Überprüfung der Hypothesen 1 und 2, die sich mit der Konstruktvalidität der beiden Assessment Center befassen, wird jeweils ein Korrelationskoeffizient ermittelt. Eine Korrelation spiegelt den Zusammenhang zwischen zwei Variablen oder Merkmalen wider. Dies bedeutet, ob die Ausprägung einer Variable X mit der Ausprägung einer Variable Y korrespondiert. Der Wertebereich für Korrelationen liegt zwischen minus eins und plus eins, wobei Werte nahe minus oder plus eins einen hohen Zusammenhang der beiden Variablen und Werte nahe an null einen niedrigen Zusammenhang suggerieren. Liegt der Korrelationswert im negativen Bereich, bedeutet dies, dass hohe Merkmalsausprägungen der Variable X mit niedrigen Werten der Variabel Y einhergehen. Liegt der Korrelationswert im positiven Bereich, heißt dies hingegen, dass eine hohe Merkmalsausprägung der Variable X auch mit einem hohen Wert der Variable Y einhergeht (Bühner & Ziegler, 2009, S. 586).

Als statistisches Maß wird der zweiseitige Korrelationskoeffizient *r* nach Pearson verwendet, da es sich um intervallskalierte Daten handelt (Bühner & Ziegler, 2009, S. 597). Zur Beurteilung der Stärke des Zusammenhangs zwischen zwei intervallskalierten

Merkmalen für das Effektstärkemaß *r* gilt nach Cohen (1988) folgende Orientierung: |*r*| ~ .10: schwacher Effekt, |*r*| ~ .30: mittlerer Effekt und |*r*| ~ .50: starker Effekt (Field, 2013, S. 270; Wirtz, 2002, S. 107). Die genannten Werte dienen als Orientierung. Zur Beurteilung der Güte einer Korrelation muss daneben auch der jeweilige Forschungsbereich (z.B. Berufseignungsdiagnostik) herangezogen werden (Wirtz, 2002, S. 107). Um zu vermeiden, dass ein gemessener Effekt in der Stichprobe nicht zufallsbedingt zustande gekommen ist, gilt als grobe Richtlinie, dass ab einer Stichprobengröße von 20 eine Korrelation von |*r*| = .30 als bedeutsam und nicht zufällig angesehen wird (Wirtz, 2002, S. 108).

Die Stichprobenverteilung von Pearsons Korrelationskoeffizient *r* folgt nicht der Normalverteilung. Deshalb kann ein Vergleich der Korrelationskoeffizienten für verschiedene Merkmale direkt nebeneinander zu Fehlinterpretationen führen. Die sogenannte Fisher-z-Transformation wandelt Pearsons *r* in eine normalverteilte Variable *r'* um. Die transformierten z-Werte ermöglichen es, verschieden Merkmale direkt zu vergleichen, da für jede Merkmalsausprägung angegeben wird, wie sie relativ zu den übrigen gemessenen Werten in der Gesamtheit liegen (Wirtz, 2002, S. 88). Für den Fishers-z-transformierten Korrelationskoeffizient lassen sich im Anschluss dann der Standardfehler und das Konfidenzintervall berechnen (Bühner & Ziegler, 2009, S. 598).

Die Effektstärke zur Messung der Korrelationsunterschiede nach Cohen (1988) wird herangezogen nach folgender Formel ermittelt: $q = r_1' - r_2'$.

q = .10 : gering,

q = .30 : mittel (moderat),

q = .50 : groß (Bühner & Ziegler, 2009, S. 613f).

Zur Überprüfung der Hypothese 3 werden ebenso zwei Korrelationskoeffizienten ermittelt und miteinander verglichen. Es gelten dabei die soeben aufgeführten Vorgehensweisen und Gütekriterien. Darüber hinaus wird diese Hypothese auf Signifikanz getestet.

Als Prüfgröße zur Entscheidung, ob die gerichtete Alternativhypothese angenommen werden kann, wird der z-Wert ermittelt. Dieser errechnet sich durch folgende Formel:

$$z = \frac{r1' - r2'}{\sqrt{\frac{1}{n1-3} + \frac{1}{n2-3}}} \text{ mit } \quad r_i' = (0{,}5) * \ln \left|\frac{1+ri}{1-ri}\right|$$

r_1 = Korrelation aus Stichprobe 1

r_2 = Korrelation aus Stichprobe 1

n_1 = Größe Stichprobe 1

n_2 = Größe Stichprobe 1

r_i' = Fischer-z-transformierte Korrelation für die Stichprobe i

z = z-Wert

ln = natürlicher Logarithmus zur Basis e

r_i = Korrelation in Stichprobe i

Der kritische z-Wert bei einer α-Fehlerwahrscheinlichkeit von 5 Prozent einer zweiseitigen Testung lautet 1.96. Ist der Wert höher, wird die Nullhypothese ($\rho_1 = \rho_2$) abgelehnt und die gerichtete Alternativhypothese ($\rho_1 > \rho_2$) angenommen (Bühner & Ziegler, 2009, S. 612f).

Im Rahmen aller drei betrachteten Auswahlverfahren wird die Interrater-Korrelation als Schätzmethode zu Erfassung der Reliabilität der eingesetzten Auswahlverfahren ermittelt. Der Grund, warum der Interrater-Korrelationskoeffizient ermittelt wird und nicht eine andere Methodik herangezogen wird, ist, dass es sich bei den Beobachtern immer um verschiedene und nicht die gleichen handelt. Auch die Teilnehmer rotieren nicht in der Art und Weise das prinzipiell jeder Beobachter alle Teilnehmer gleich oft sehen und beurteilen kann. Das weitere Vorgehen ist das Gleiche wie in den ersten beiden Hypothesenprüfungen.

Für die Überprüfung der fünften Hypothese wird eine lineare Regressionsanalyse durchgeführt. Eine Regressionsanalyse trifft eine Aussage darüber wie sich ein Merkmal Y

aus einem korrelierten Merkmal X am besten vorhersagen lässt. D.h. welche Transformation der x-Werte liefert eine möglichst genaue Schätzung der Ausprägung der y-Werte (Wirtz, 2002, S. 98).

Eine notwendige Voraussetzung für eine aussagekräftige Regressionsgleichung ist die Existenz eines Zusammenhangs der Merkmale X und Y. Je höher dieser Zusammenhang ist, desto präziser wird auch die Vorhersage sein (Wirtz, 2002, S. 98). Um eine kausale Interpretation (für die Grundgesamtheit) vornehmen zu können, müssen folgende Voraussetzungen erfüllt sein:

- Es gibt einen signifikanten Zusammenhang zwischen X und Y (gemessen durch einen Korrelationskoeffizienten).
- Prädiktor und Kriterium müssen zeitlich aufeinander folgen oder eine strikte Theorie schließt aus, dass die Korrelation durch eine Drittvariable bedingt ist (Bühner & Ziegler, 2009, S. 594).

Um in der hier vorliegenden empirischen Untersuchung und der Überprüfung von Hypothese 5 einen kausalen Zusammenhang der Vorgesetztenbeurteilung mit dem Gesamtergebnis des Assessment Centers zu überprüfen, wird folglich zuerst der Korrelationskoeffizient zwischen dem erzielten Gesamtergebnis im eingesetzten Auswahlverfahren und dem Gesamtergebnis der Vorgesetztenbeurteilung ermittelt. Voraussetzung zwei ist in der empirischen Untersuchung gegeben, da die Vorgesetztenbeurteilung zeitlich versetzt (mindestens 6 Monate später) zum Assessment Center stattfindet.

Folgender Term veranschaulicht eine Regressionsgleichung:

$\hat{y}_i$ *(vorhergesagter Wert des Probanden i)* = a_y *(Konstante)* + b_{yx} *(Steigungskoeffizient)* * x_i *(Wert x des Probanden i)* + *Fehler* (Bühner & Ziegler, 2009, S. 591).

Bei der Regressionsrechnung als Fehler betrachtet, werden die Effekte aller Störvariablen, die in der Vorhersagegleichung nicht berücksichtigt werden und Messungenauigkeiten, die bei der Bestimmung von X und Y entstehen (Wirtz, 2002, S. 98).

Um die Korrelationsunterschiede der prädiktiven Validität der eingesetzten Auswahlverfahren auf Signifikanz zu prüfen, wird wie in der dritten Hypothesenprüfung der z-Wert ermittelt.

Datenerhebung und Stichprobe

Die Datenerhebung fand im Zeitraum vom 29.04.2009 bis zum 08.08.2013 statt. Das Trainee-Assessment Center haben im Zeitraum vom 29.04.2009 bis zum 03.02.2012 insgesamt 70 Personen durchlaufen. Von diesen Personen haben 41 Personen das Assessment Center bestanden, was eine Bestehens-Quote von 58,57 Prozent bedeutet. Von diesen 41 Personen wurden wiederum ca. zwei Drittel, also 28 Personen eingestellt. Das Stabs-Assessment Center haben im Zeitraum vom 22.06.2010 bis zum 06.01.2012 eine ähnliche Anzahl, nämlich 84 Personen, durchlaufen. Von diesen haben 61 Personen das Assessment Center bestanden, 28 Kandidaten wurden eingestellt. Die Bestehens-Quote liegt hier mit 73,49 Prozent deutlich über dem Trainee-Assessment Center und ist insgesamt sehr hoch. Von den Personen, die das Assessment Center bestanden haben, wurde mit 45,90 Prozent nur knapp die Hälfte dann auch eingestellt, was sehr gering ist. Am Multimodale Auswahlverfahren haben im Zeitraum vom 28.02.2012 bis zum 08.08.2013 insgesamt 90 Personen teilgenommen. Die Bestehens-Quote ist auch hier, mit 82,22 Prozent und 74 Personen, sehr

hoch. Von diesen Personen wurden letztendlich etwas mehr als die Hälfte, nämlich 41 Personen, eingestellt.

Die Vorgesetztenbeurteilung liegt insgesamt von 76 Personen vor. Bei den Referenten liegt von den 28 eingestellten Personen bei 23 Personen eine Beurteilung vor. Bei den Trainees liegt von 28 eingestellten Personen bei 21 Personen eine Beurteilung vor. Von den 41 eingestellten des Multimodalen Auswahlverfahrens liegen 31 Beurteilungen vor. Die Ursachen für eine fehlende Beurteilung liegen entweder in einem Vorgesetztenwechsel bzw. einer Nichterreichung des Mindestbeurteilungszeitraums von sechs Monaten. Sofern mehr als eine Beurteilung vorliegt, wurde der Durchschnitt gebildet.

Das Untersuchungsdesign der empirischen Studie

Die nachfolgenden Kapitel haben eine Beschreibung des Untersuchungsdesigns zum Inhalt. Zu Beginn wird die betroffene Zielgruppe erläutert. Daran knüpft eine Darstellung der eingesetzten Personalauswahlverfahren. Zuerst werden die beiden Assessment Center skizziert, die getrennt für beide Zielgruppen im Einsatz waren. Im Nachgang erfolgt eine detaillierte Erläuterung des neuen, Multimodalen Auswahlverfahrens, das für beide Zielgruppen neu konzipiert wurde, unter Berücksichtigung der skizzierten wissenschaftlichen Erkenntnisse. Die Vorgesetztenbeurteilung als Zielkriterium zur Erfassung der prädiktiven Validität der eingesetzten Auswahlverfahren bildet den Abschluss des Kapitels.

Die Zielgruppe

Bei der im Rahmen der Untersuchung betrachteten Zielgruppe handelt es sich um potentielle Kandidaten für Referenten- und Trainee-Positionen in einem deutschen Unternehmen. Beide Personengruppen kommen für gehobene Positionen im Unternehmen in Frage. Die gehobenen Positionen sind in den Stabs-Bereichen des Unternehmens angesiedelt. Typische Bereiche hierfür sind Personal, Recht, Controlling, Marketing, Produktvertriebsmanagement oder Mathematik.

Die ausgeführten Tätigkeiten zeichnen sich durch eine hohe Komplexität und Verantwortung aus. Je nach betrachtetem Bereich, müssen beispielsweise Aktuare in der Mathematik hoch komplexe Simulationen erstellen. Ähnliches gilt für das Controlling, in dem exakte Hochrechnungen sowie ein Planungsprozess für das ganze Unternehmen erstellt werden.

Die Zielgruppen unterscheiden sich in erster Linie durch die Einstiegsart in das Unternehmen. Bei der Zielgruppe Trainees handelt es sich um Hochschulabsolventen, die in der Regel keine Berufserfahrung aufweisen. Diese durchlaufen für die Dauer von 1,5 Jah-

ren drei Stationen im Unternehmen. Im Anschluss der Trainee-Tätigkeit übernimmt das Unternehmen in der Regel die eingestellten Personen als Referenten in einem der Stabs-Bereiche. Darüber hinaus ist auch ein direkter Einstieg als Referent in das Unternehmen möglich. Diese Zielgruppe weist in der Regel drei bis fünf Jahre Berufserfahrung auf. Aufgrund der hohen Verantwortung, der mit der Position verbundenen Tätigkeiten, stellen die Referenten in der Regel Führungsnachwuchskräfte dar oder nehmen später Expertenfunktionen ein.

Die beiden klassischen Assessment Center

Im nachfolgenden erfolgt eine Beschreibung der eingesetzten Assessment Center, die für beide Zielgruppen getrennt im Einsatz waren. Bei beiden Auswahlverfahren handelt es sich um eintägige Gruppen-Assessment Center, die mit zwei bis sechs Teilnehmern durchgeführt wurden. Die Übungsbausteine sind in beiden Verfahren ähnlich. In beiden Assessment Centern wird innerhalb der Übungen dimensionsbezogen beurteilt. Auch das Feedback erfolgt kompetenz- und nicht übungsbezogen. In den Beobachterteams sitzen jeweils eine erfahrene Beobachterin aus dem Personalbereich und ein geübter Beobachter aus dem Fachbereich. Im Vorfeld erfolgt in beiden Fällen eine ausführliche Beobachterschulung.

Den Teilnehmern werden zu Beginn des Assessment Centers im Rahmen einer Teilnehmereinführung die Bausteine des Auswahlverfahrens sowie das dahinterstehende Kompetenzmodell bekannt gemacht. Darüber hinaus erhalten sie Verhaltensempfehlungen. Am Durchführungstag steht den Teilnehmern für Fragen zum Ablauf eine Assistenz zur Seite. Die Übungen finden pro Teilnehmer abwechselnd in den einzelnen Beobachterteams statt, einzig die Gruppendiskussion erfolgt in einer Großgruppe.

Die Teilnehmer fungieren in beiden Assessment Centern als Mitarbeiter eines fiktiven Unternehmens. Die einzelnen Übungen sollen einen möglichst realistischen Bezug zur

Branche sowie zum Anforderungsprofil der Funktion als Trainee- oder Stabs-Mitarbeiter haben. Hierfür wird den Teilnehmern eine Rahmenhandlung zur Verfügung gestellt, die das fiktive Unternehmen skizziert.

Das Trainee-Assessment Center.

Im Rahmen des Trainee-Assessment Centers nehmen Kandidaten teil, die zuvor im Rahmen eines teilstrukturierten Bewerbungsgesprächs eine Empfehlung durch die Fachbereichsführungskraft sowie die Traineeverantwortliche bekommen haben. Die im nachfolgenden beschriebenen fünf Übungen werden absolviert.

Der Teilnehmer soll sich zu Beginn im Rahmen einer *Selbstpräsentation* den Beobachtern im Hinblick auf vorher definierte Aspekte selbst vorstellen. Zielsetzung dieser anfänglichen Übung ist es, herauszufinden, wie der Kandidat die Beobachter von seiner Person überzeugen kann.

Basis der *Projektpräsentationsübung* sind zahlreiche Materialien zu einem Projekt, das sich mit der Rekrutierung neuer Mitarbeiter beschäftigt. Die Erwartung an den Teilnehmer ist, dass er in der Kürze der Zeit die Unterlagen sichtet, diese kurz und prägnant zusammenfasst und seine Ergebnisse den Beobachtern präsentiert. Ebenso wichtig ist in dieser Übung der Umgang mit kritischen Fragen seitens der Beobachter.

Im Rahmen der *Gruppendiskussion* bilden alle Teilnehmer ein Projektteam und sollen gemeinsam Vorschläge und Maßnahmen zur Planung und Organisation des Projektes erarbeiten. Zielsetzung dieser Übung ist es, zu sehen, wie der Teilnehmer im Team agiert und wie groß die Beteiligung am Teamprozess ist.

Aus der Rolle des Projektleiters führt der Teilnehmer im *Mitarbeitergespräch* ein Gespräch mit einem Projektmitarbeiter, der seine projektbezogenen Aufgaben nicht wie erforderlich wahrnimmt. Zielsetzung ist, die kritischen Punkte im Rahmen einer offenen

und wertschätzenden Gesprächsatmosphäre zu benennen, aber auch Veränderungsbedarf aufzeigen und den Mitarbeiter dafür zu gewinnen, wieder aktiv mit zu arbeiten.

Im Rahmen des *Postkorbs* agiert der Teilnehmer als Projektleiter. Er muss nach seinem mehrwöchigen Urlaub eine Reihe von Vorgängen bearbeiten und Entscheidungen treffen. Der Teilnehmer erhält Unterlagen zu verschiedenen, miteinander verzahnten Vorgängen und muss diese erkennen.

Abbildung 1 zeigt die im Trainee-Assessment Center erfassten Dimensionen, die sich am Kompetenzmodell des Konzerns orientieren und dem des Stabs-Assessment Centers ähneln. Allerdings gibt es in der Beschreibung der einzelnen Dimensionen als auch in den einzelnen Verhaltensankern Unterschiede gegenüber dem Stabs-Assessment Center, wie im nächsten Kapitel deutlich wird. Jedes Kriterium wird mindestens in zwei Übungen erfasst und bis zu fünf verschiedene Kriterien werden innerhalb einer Übung durch den Beobachter beurteilt. Die Beurteilungsskala stellt eine sechsstufige Skala dar, die in Abbildung 2 ersichtlich ist.

Soziale Kompetenzen	Persönliche Kompetenz	Unternehmerische Kompetenz	Zusätzliche Kompetenzen
Kommunikationsfähigkeit (KOMM)	Initiative und Eigenverantwortung (IE)	Entscheidungsfähigkeit (ENT)	Überzeugungskraft (Ü)
Kooperationsfähigkeit (KOOP)	Stabilität und Selbstvertrauen (SUS)	Strategisches Denken und Handeln (SUH)	Auffassungsgabe (A)
Konflikt- u. Kritikfähigkeit (KONF/KRI)			

Abbildung 1. Erfasste Kompetenzen im Trainee-Assessment Center.

Beobachter: Teilnehmer:

+++	++	+	-	--	---
ausgesprochen positiv	deutlich positiv	leicht positiv	leicht entwicklungsbedürftig	deutlich entwicklungsbedürftig	stark entwicklungsbedürftig

Strategisches Denken und Handeln					
+++	++	+	-	--	---
O	O	O	O	O	O
• Setzt bei seinen Entscheidungen stets unternehmerisch sinnvolle Prioritäten • Baut seine Argumentation und Präsentation stets logisch und nachvollziehbar auf • Kann sich auf Nachfragen hin durchgängig verbessern			• Unternehmerisch sinnvolle Prioritäten sind nicht erkennbar • Argumentation und Präsentation sind unlogisch und nicht nachvollziehbar aufgebaut • kann auf Nachfragen keine richtigen Antworten finden		

Entscheidungsfähigkeit					
+++	++	+	-	--	---
O	O	O	O	O	O
• alle Entscheidungen sind richtig und plausibel begründet • nutzt als Entscheidungsgrundlage alle notwendigen Informationen • kann sich bei optionalen Lösungsmöglichkeiten für eine Alternative eindeutig entscheiden und diese überzeugend argumentieren			• Entscheidungen sind falsch und die Begründungen sind nicht plausibel • nutzt die gegebenen Informationen nicht als Entscheidungsgrundlage • kann bei optionalen Lösungsmöglichkeiten keine Entscheidungen treffen		

Abbildung 2. Auszug eines Beurteilungsbogens aus dem Trainee-Assessment Center.

Das Stabs-Assessment Center

Die Teilnehmer des Stabs-Assessment Centers sind Bewerber, die zuvor im Rahmen eines teilstrukturierten Bewerbungsgesprächs eine Empfehlung durch die Fachbereichsführungskraft sowie den betreuenden Personalreferenten bekommen haben. Im nachfolgenden sind die Bausteine des Stabs-Assessment Centers dargestellt. Die Übungsauswahl und der Grundgedanke der Übungen ähneln den Übungen des Trainee-Assessment Centers.

Im Rahmen der *Selbstpräsentation* ist der Teilnehmer angehalten, über den beruflichen Werdegang hinaus weitere Aspekte einzubauen, die seine Motivation und Stärken für die jeweilige Position aufzeigen.

In der *Präsentationsübung* ist es die Aufgabe des Teilnehmers, als Fachexperte im Rahmen eines Umstrukturierungsprojektes den Beobachtern, als fiktiven Unternehmensmitgliedern, erste Untersuchungsergebnisse vorzustellen und mit kritischen Fragen zu seinem Konzept umzugehen.

In der anschließenden *Gruppendiskussion* erarbeiten alle Teilnehmer gemeinsam Maßnahmenpakete zum Projekt, die anschließend dem Vorstand präsentiert werden.

Das variable *Rollenspiel* gliedert sich in drei mögliche Bausteine auf: Das *Beratungsgespräch*, das *Motivationsgespräch* und das *Kollegengespräch*. Der Fachbereich entscheidet hierbei unter Beratung der Personalentwicklung, welchen Baustein er in das Verfahren integriert. Ausschlaggebend für diese Entscheidung ist das zukünftige Tätigkeitsfeld des Bewerbers.

Im *Beratungsgespräch* fungiert der Teilnehmer als Stellvertreter des Projektleiters und führt ein Gespräch mit seinem internen Auftraggeber. Seine Aufgabe ist es hierbei, die genauen Rahmenbedingungen des Projektes zu klären.

Im *Motivationsgespräch* muss der Teilnehmer einen ihm lediglich fachlich und nicht disziplinarisch untergebenen Projektmitarbeiter dazu bewegen, ihn bei einem Aufgabenpaket im Rahmen des Projektes kurzfristig zu unterstützen.

Der Teilnehmer hat im *Kollegengespräch* ähnlich wie im Trainee-Assessment Center die Aufgabe, ein Gespräch mit einem Projektmitarbeiter zu führen, der seit einigen Wochen bei der Bewältigung seiner Projektaufgaben negativ auffällt. Zielsetzung ist auch hier, die kritischen Punkte zu benennen, aber auch Veränderungsbedarf aufzuzeigen und den Mitarbeiter dafür zu gewinnen, wieder aktiv mitzuarbeiten.

Im *Kick-Off-Meeting* soll der Teilnehmer für das neu fusionierte und umstrukturierte Unternehmen mit Vertretern aus unterschiedlichen Unternehmensbereichen einen Leitbildprozess aufsetzen. Er hat die Aufgabe, die Projektmitarbeiter über die wesentlichen Aspekte zu informieren, sie zur aktiven Teilnahme zu motivieren und sicherzustellen, dass sie sich mit dem Projekt identifizieren.

Der elektronische *Postkorb* hat zur Zielsetzung, dass der Teilnehmer Aufgaben in Form von E-Mails erhält, auf die er in unterschiedlicher Form reagieren muss. Zudem steht

ihm ein Kalender zur Organisation seiner Termine und Aufgaben zur Verfügung. Die Übung ist unbeobachtet. Das Ergebnis wird auf Basis der produzierten Antworten des Teilnehmers automatisch computergesteuert berechnet.

Abbildung 3 zeigt das zugrundliegende Kompetenzmodell des Stabs-Assessment Centers, das sich wiederum an dem des Konzerns orientiert. Es wird deutlich, dass die Kompetenzen denen im Trainee-Assessment Center ähneln, nur zum Teil anders benannt oder zugeordnet sind. Ein exemplarischer Beurteilungsbogen ist in Abbildung 4 ersichtlich. Dieser ist ähnlich wie im Trainee-Assessment Center aufgebaut, die Skala ist ebenso sechsstufig.

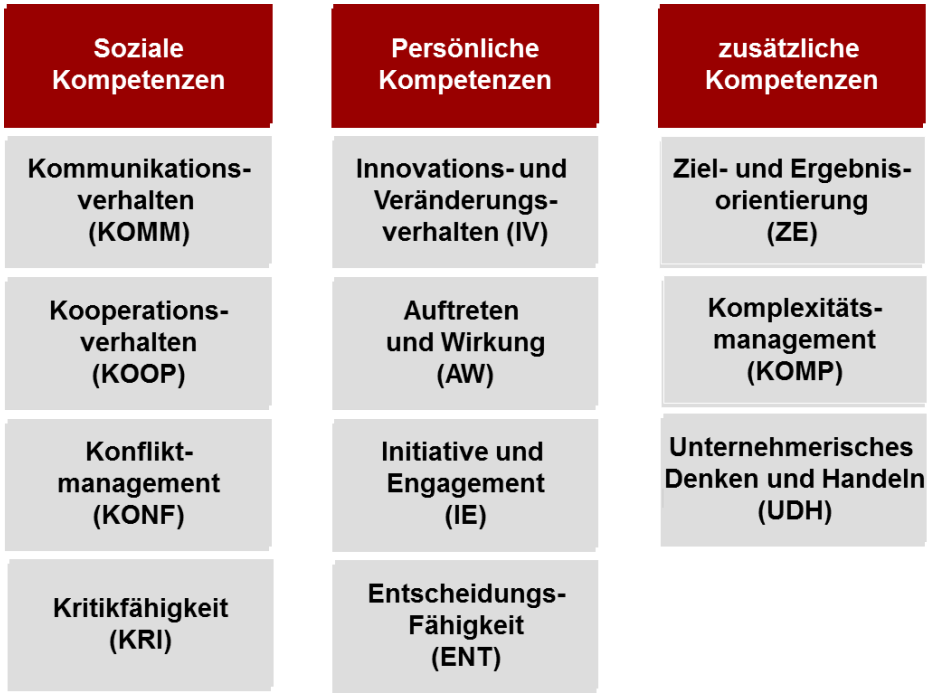


Abbildung 3. Erfasste Kompetenzen im Stabs-Assessment Center.

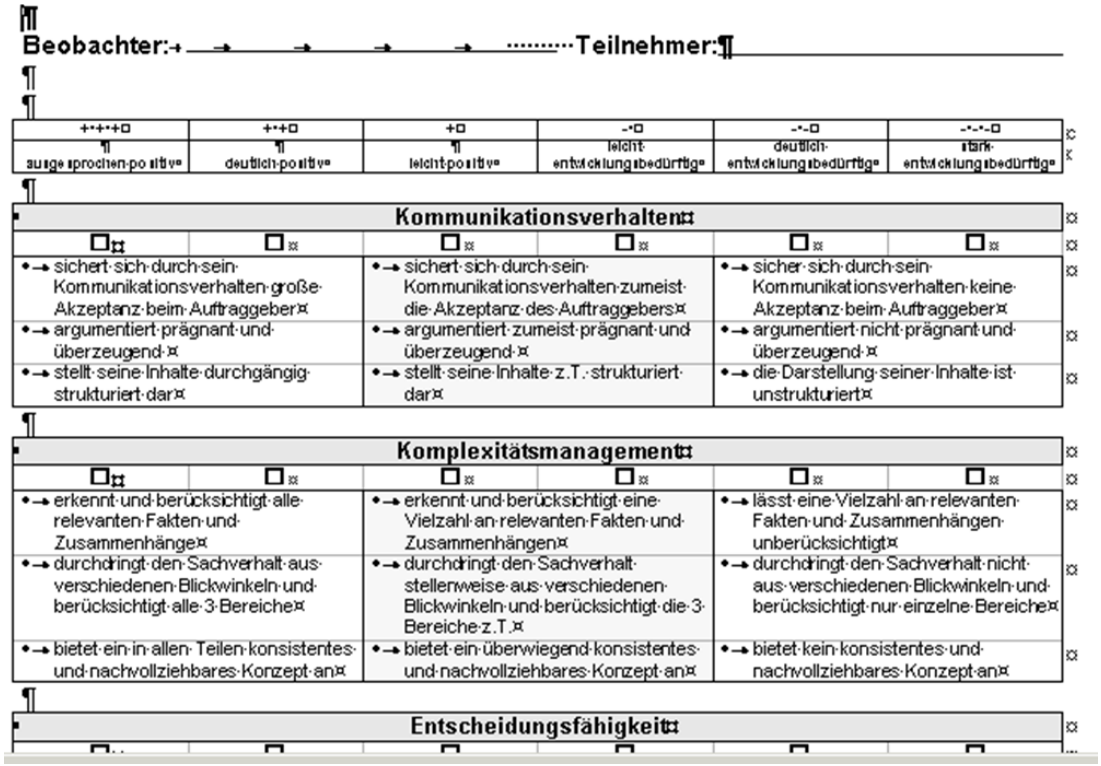


Abbildung 4. Auszug aus dem Beurteilungsbogen des Stabs-Assessment Centers.

Das Multimodale Auswahlverfahren

Zielsetzung bei der Konzeption des Multimodalen Auswahlverfahrens war es, einerseits die skizzierten wissenschaftlichen Erkenntnisse einfließen zu lassen und andererseits die Rahmenbedingungen seitens des Unternehmens einzuhalten.

Zielsetzungen und Rahmenbedingungen.

Zielsetzung seitens des Unternehmens war die Konzeption eines Personalauswahlverfahrens, das für die beiden Zielgruppen gilt. Darüber hinaus sollten Synergien durch die Zusammenlegung der beiden Auswahlverfahren geschaffen werden. Die Dauer sollte auf vier bis fünf Stunden gekürzt werden, um den Aufwand für die Beobachter und Bewerber zu reduzieren und zugleich die Bewerberorientierung zu verbessern. Eine geringere Komple-

xität des Verfahrens sowie eine zeitliche Reduzierung in der Erfassung, Vorbereitung, Durchführung und Nachbereitung galt es zu erreichen, um eine geringere Fehleranfälligkeit und einen geringeren organisatorischen Aufwand zu erzielen. Des Weiteren sollte die Validität des gesamten Verfahrens und der einzelnen Übungsbausteine gesteigert werden, um ein verbessertes Selektionskriterium zu haben, dass das Risiko von Fehleinstellungen reduziert.

Die aus der DIN 33430 geforderten Kriterien *Anforderungsbezug*, *Simulation*, *Methodenvielfalt*, *Mehrfachbeurteilung und Transparenz* (Amelang & Schmidt-Atzert, 2009, S. 463) wurden bei der Konzeption des Auswahlverfahrens berücksichtigt. Das Verfahren erfasst Merkmale oder Verhaltensweisen, die für die zu besetzenden Positionen relevant sind (*Anforderungsbezug*). Diese wurden im Rahmen diverser Konzeptionsworkshops detailliert mit Hilfe von Stellenbeschreibungen, Anforderungsprofilen und eigener, langjähriger Erfahrung ermittelt. Darüber hinaus simulieren die einzelnen Übungsbausteine möglichst realistische Situationen, die sich an den zukünftigen Jobanforderungen orientieren (*Simulation*). Eine *Methodenvielfalt* wird dadurch erreicht, dass durch die Kombination eines Intelligenztests, des strukturierten Interviews und zwei verschiedenen Assessment Center Übungen verschiedene Personalauswahlinstrumente miteinander kombiniert werden. Das Prinzip der *Mehrfachbeurteilung* wird dadurch erzielt, dass jeder Teilnehmer von mehreren Personen abwechselnd in verschiedenen Übungen beobachtet und beurteilt wird. Darüber hinaus galt es für die Zusammenstellung der Beobachterteams neben einem Beobachter aus dem Fachbereich jeweils einen erfahrenen Beobachter aus dem Personalbereich hinzuzuziehen. Ein Hinzuziehen von ausgebildeten Psychologen als Beobachtern ist sowohl aus Reliabilitäts- als auch Validitätsgründen sinnvoll (Obermann, 2013, S. 185). Eine ausführliche Beobachterschulung sowie Teilnehmereinführung sorgt für die notwendige *Transparenz* des Verfahrens. Im Anschluss an das Verfahren erfolgt ein ausführliches

und individuelles Feedback an die Kandidaten mit Bezug auf das konkrete Anforderungsprofil. Durch diese Offenheit und Transparenz kann die Akzeptanz des Assessment Centers bei den Teilnehmern gestärkt werden (Amelang & Schmidt-Atzert, 2006, S. 459, 2009, S. 463f; Obermann, 2009, S. 10).

Des Weiteren wurde im Rahmen des Konstruktionsprozesses ein großer Fokus auf die Übungsinstruktionen gelegt. D.h. zum einem muss den Kandidaten die genaue Aufgabenstellung inkl. Vorbereitungszeit, Durchführungszeit, Materialeinsatz etc. klar sein. Zum anderen muss den Beobachtern bewusst sein, worauf sie in den Übungen achten sollen. Ebenso entscheidend für den Konstruktionsprozess war, dass der Beobachtungsbogen die konkreten Anforderungsdimensionen sowie deren Operationalisierungen mit konkreten Verhaltensbeispielen enthält (Drees, 1994, S. 18f). Abbildung 5 zeigt beispielhaft einen Auszug aus der neuen Beurteilungsliste.

Beobachter: **Teilnehmer:**

• gibt Informationen und Fakten unvollständig und unsicher wieder • gibt eine nicht geordnete Darstellung der Kernpunkte (kein roter Faden)	• Argumentiert widersprüchlich	• Gespräch enthält Einleitung und Abschluss • Argumentiert konsequent für die ursprünglich angedachte Strategie (er kommt immer wieder auf sie zurück)	• Fasst am Ende das Gesprächsergebnis zusammen und erzielt ein gemeinsames Verständnis • Begründet die Herleitung der Entscheidung prägnant und überzeugend
• kann sich nicht auf Entscheidungen festlegen • Schiebt bei Einwänden die Verantwortung auf seinen Vorgesetzten und sagt, er fände die	• kann sich nur in den Punkten festlegen, die unstrittig sind • kann nicht alle Entscheidungen plausibel begründen • verweist darauf, dass	• Stellt pro und contra der Handlungsalternativen dar • begründet bei Fragen seine	• Beurteilt die Maßnahmen
0	**1**	**2**	**3**
stark entwicklungsbedürftig	leicht entwicklungsbedürftig	Kein Entwicklungsbedarf	Stärke
Zusätzliche Beobachtungen			

Abbildung 5. Auszug aus einem Beurteilungsbogen des Multimodalen Auswahlverfahrens.

Die einzelnen Übungsbausteine des Multimodalen Auswahlverfahrens werden im nächsten Kapitel genauer beschrieben.

Die Übungen des neuen Auswahlverfahrens

Die im Theorieteil bereits skizzierten, besten Prädiktoren für Berufserfolg sind nach Schmidt und Hunter (1998) GMA und eine Arbeitsprobe, GMA und ein Integritätstest oder GMA und ein strukturiertes Interview (Schmidt & Hunter, 1998, S. 265). Diese wissenschaftlichen Erkenntnisse sind im Rahmen der unternehmensseitigen Zielsetzungen in den Aufbau des Multimodalen Auswahlverfahrens eingeflossen. Wie im nachfolgenden genauer skizziert wird, werden als Bausteine ein Intelligenztest, ein strukturiertes Interview und überarbeitete Assessment Center-Übungen mit realistischem Bezug zu den zukünftigen Tätigkeiten eines Trainee- oder Stabs-Mitarbeiters eingesetzt.

Die einzelnen Bausteine sind in Abbildung 6 dargestellt und werden im Anschluss ausführlicher beschrieben.

Übungsbeschreibung	Dauer
Teilnehmereinführung und Vorstellung Beobachter	**15 Min. Gesamtdauer.**
Selbstpräsentation - Auftrag in Einladung, Vorbereitung zu Hause - Keine Bewertung - Zielsetzung: Wertschätzung, Teilnehmer lernen sich und die Beobachter kennen	**5-10 Minuten pro Teilnehmer**
Generisches Interview 1. Sammlung von Informationen 2. Umgang mit Stress	**15 Minuten (2 x 7 Minuten) Durchführung**
1. Beratungsgespräch (Stabs-Mitarbeiter) - in Anlehnung an die Übung des bisherigen Stabs-ACs, gekürzt, gering modifiziert **2. Kollegengespräch (Trainees)** - in Anlehnung an die Übung des bisherigenTrainee-ACs, gekürzt, gering modifiziert	**20 Minuten Vorbereitung** **15 Minuten Durchführung**
Projektpräsentation - in Anlehnung an die Übung des bisherigen Trainee-ACs, gekürzt, gering modifiziert	**30 Minuten Vorbereitung** **15 Minuten Durchführung**
Test (numerische und verbale Fähigkeiten)	**ca. 90 Minuten**
Beobachterkonferenz und Feedbackgespräche	**ca. 15-30 Min Beo.konf. pro TN** **ca. 10-15 Min Feeback pro TN**

Abbildung 6. Die eingesetzten Bausteine des Multimodalen Auswahlverfahrens.

Zu Beginn des Auswahlverfahrens finden eine kurze Vorstellungsrunde und eine *Teilnehmereinführung* statt. Die Einführung gibt wie in den vorangegangenen Assessment Centern einen Ausblick auf die Bausteine des Auswahlverfahrens und Verhaltensempfehlungen.

Der Teilnehmer ist angehalten im Rahmen der *Selbstpräsentation* über seinen Werdegang hinaus weitere Aspekte einzubauen, die seine Motivation für eine Trainee- oder Stabsposition und seine Stärken, die er in der genannten Position im Unternehmen einbringen kann, ausdrücken. Diese Übung geht nicht in das statistische Gesamtergebnis ein, es dient lediglich dem Zweck, dass die Teilnehmer das Setting kennen lernen und die Beobachter einen ersten Eindruck der Bewerber erhalten.

Beim zweiten Baustein, dem *generischen Interview,* handelt es sich um ein strukturiertes Interview. Hier werden dem Teilnehmer ohne Vorbereitungszeit hintereinander zwei Aufgabenstellungen skizziert, die dieser sofort beantworten muss. Die erste Aufgabenstellung beinhaltet die *Sammlung von Informationen*. Der Teilnehmer soll in diesem Fall beschreiben, wie er bei der Recherche zu einem Produktvergleich vorgeht und wie er die relevanten Informationen in einer Power Point Präsentation zusammenfasst. Zielsetzung dieser Aufgabe ist, eine möglichst genaue Beschreibung der Informationsbeschaffung und deren Quellen, sowie die formalen Kriterien einer Präsentation zu erhalten. Die zweite Aufgabe handelt vom *Umgang mit Stress*. Der Teilnehmer soll hierbei beschreiben, wie er vorgeht, wenn ihm im Rahmen seiner Stelle neue Aufgaben übertragen werden. Es handelt sich dabei um einen neuen Aufgabenbereich, den er schon immer übernehmen wollte, wohlwissend, dass dies einen erheblichen Mehraufwand bedeutet und ihn stark fordern wird. Aufgrund der kurzen Unternehmenszugehörigkeit möchte der Teilnehmer positiv in Erscheinung treten. Diese Übung hat zum Ziel, dass der Teilnehmer möglichst genau beschreibt, wie er in den nächsten Wochen vorgeht und mit dieser Mehrbelastung umgeht.

Dieses strukturierte Interview stellt eine realistische Alltagssituation beider Zielgruppen (Trainees und Stabs-Mitarbeiter) dar.

Um der unterschiedlichen Berufserfahrung der beiden Zielgruppen gerecht zu werden, unterscheidet die *Gesprächsübung* zwischen den Trainees und Referenten. Bei dieser Übung handelt es sich um eine klassische Assessment Center-Übung. Den Referenten wird aufgrund der Berufserfahrung eine Gesprächssituation aus dem beruflichen Alltag vorgelegt, den Trainees hingegen eine typische Bewerbersituation.

Bei den Referenten lehnt sich die Übung an das *Beratungsgespräch* aus dem *Stabs-Assessment Center* an. Dieses wird an die in der Literatur skizzierten Normen angepasst. Die Übung wird insgesamt gekürzt und die Aufgabenstellung des Gesprächs konkretisiert, sodass die Zielsetzung der Übung für den Bewerber deutlicher ist. Darüber hinaus wird für die Beobachter das Beurteilungsprinzip verändert. Wie bereits erwähnt, gilt es in diesem Baustein, eine übungsbezogene Beurteilung heranzuziehen anstatt einer kompetenzbezogenen wie in der Vergangenheit. Abgesehen von den wissenschaftlichen Aspekten hat das neue Beurteilungsprinzip für die Beobachter den Vorteil, dass die Entscheidungsfindung und somit eine Beurteilung jeder einzelnen Übung durch dieses Beurteilungsprinzip erleichtert wird. Auch das Feedback an die Teilnehmer ist für die Beobachter leichter möglich und für die Teilnehmer besser nachvollziehbar.

Für die Zielgruppe Trainees wird die Übung an das bisher eingesetzte *Kollegengespräch* angelehnt. Bei den vorgenommenen Anpassungen gilt das Gleiche wie beim Beratungsgespräch.

Die *Projektpräsentation* stellt ebenso eine klassische Assessment Center Übung dar. Diese lehnt sich an die Übung des bisherigen Trainee-Assessment Centers an, um beiden Zielgruppen gerecht zu werden. Auch diese wurde gekürzt, fokussierter erläutert und an das neue Beurteilungsprinzip angepasst.

Als Testbaustein ist in diesem Verfahren der bereits im Theorieteil beschriebene IST-2000-R eingesetzt. Für das vorliegende Auswahlverfahren werden lediglich die Bausteine numerische und verbale Intelligenz ausgewählt. Ein Grund hierfür ist, dass diese insbesondere bei der Ausführung der Tätigkeit für alle betrachteten Personengruppen ausschlaggebend sind. Nach der DIN 33430 sollte ein Bezug zur ausführenden Tätigkeit der Zielgruppe gegeben sein. Darüber hinaus haben die numerischen Fähigkeiten, wie bereits beschrieben, mit r = .52 ähnlich hohe Gesamtprognosewerte wie die allgemeinen kognitiven Fähigkeiten (Salgado et al., 2003, S. 586). Geht man davon aus, dass der geringere Validitätswert für die verbalen Fähigkeiten auf sprachliche Unterschiede zurückzuführen ist, wird dieses Problem in der zu Grunde liegenden Studie dadurch ausgeschlossen, da es sich in der vorliegenden Zielgruppe ausschließlich um deutschsprachige Bewerber handelt. Die anderen Bausteine des IST-2000-R wurden aus den vorgegebenen Zeitgründen seitens des Unternehmens und einer insgesamt gewünschten Methodenvielfalt unter starken zeitlichen Restriktionen nicht berücksichtigt. Die Vermutung liegt nahe, dass die anderen Facetten von Intelligenz teilweise indirekt durch andere Bausteine des Multimodalen Auswahlverfahrens abgedeckt sind. Der Grund dafür ist, dass Assessment Center an sich bereits hoch mit den allgemeinen kognitiven Fähigkeiten korrelieren, da sie typischerweise schon einen hohen Anteil der allgemeinen kognitiven Fähigkeiten im Konstrukt haben (Schmidt & Hunter, 1998, S. 269).

Die Vorgesetztenbeurteilung als Zielkriterium

Die Beurteilungskriterien der Vorgesetztenbeurteilung orientieren sich wie die Kriterien des Stabs- und Trainee-Assessment Centers an dem Konzernkompetenzmodell des Konzerns, enthalten allerdings Unterschiede im Wortlaut und der Operationalisierung. Eine professionelle Schulung der Vorgesetzten ist sichergestellt. Darüber hinaus haben die Füh-

rungskräfte zuvor als Beobachter im Trainee- oder Stabs-Assessment Center bzw. dem Multimodalen Auswahlverfahren teilgenommen.

Der Beurteilungszeitraum findet bei den direkt eingestellten Referenten einmal im Jahr, im Zeitraum von Anfang Februar bis Ende April, statt. Voraussetzung ist, dass der Vorgesetzte den Mitarbeiter mindestens sechs Monate geführt hat. Die Trainees hingegen werden alle sechs Monate nach jeder Station beurteilt, der Beurteilungsbogen ist derselbe. Nach der Zeitdauer von 18 Monaten mit drei Fachbereichsstationen werden die Trainees in der Regel als Referent von einem durchlaufenen Fachbereich übernommen und gliedern sich dem Beurteilungssystem der oben aufgeführten Referenten ein.

Tabelle 6 zeigt die erfassten Kompetenzen der Vorgesetztenbeurteilung.

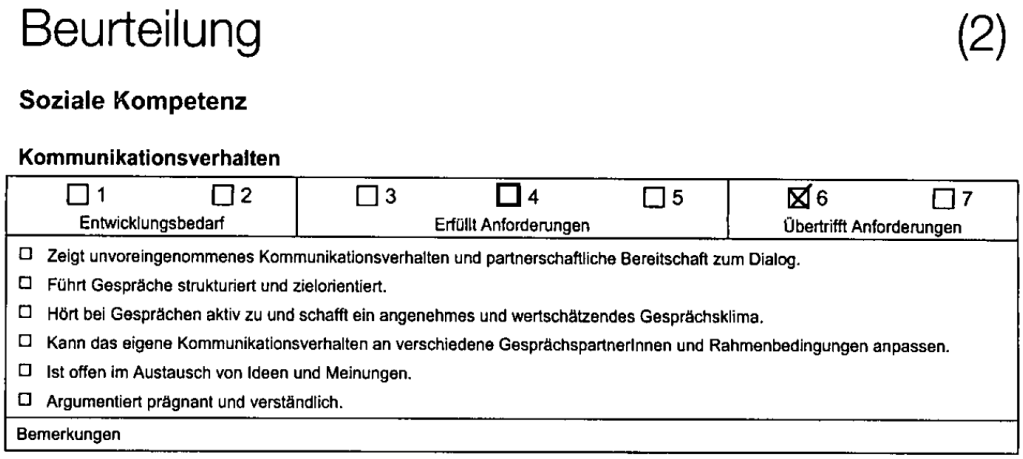
Tabelle 6

Die erfassten Kompetenzen im Rahmen der Vorgesetztenbeurteilung

Fach- und Methoden-kompetenz	**Soziale Kompetenz**	**Persönliche Kompetenz**	**Kunden,- Service- und Vertriebs-orientierung**
Kenntnisse und Fertigkeiten	Kommunikations-verhalten	Initiative und Engagement	Kunden- & Service-orientierung
Arbeitsqualität	Kooperations-verhalten	Entscheidungs-fähigkeit	Vertriebsorientierung
Arbeitsquantität	Konflikt- und Kritikfähigkeit		
Planung und Organisation			

Wie beim Blick auf die zu beurteilenden Kompetenzen deutlich wird, existieren deutliche Unterschiede in der Betitelung und Operationalisierung gegenüber der beiden Auswahlverfahren des Trainee- und Stabs-Assessment Centers. Aus diesem Grund kann bei der Berechnung des Zusammenhangs zwischen dem Assessment Center Ergebnis und der Vorgesetztenbeurteilung nicht auf Zusammenhänge einzeln bewerteter Kompetenzen eingegangen, sondern es können lediglich die Gesamturteile verglichen werden.

Bei der Skala handelt es sich um eine siebenstufige Skala. Die einzelnen Beurteilungsstufen sind klar definiert und die einzelnen Dimensionen wie in dem in Abbildung 7 dargestellten Beispiel zu Kommunikationsverhalten operationalisiert.



Beurteilung (2)

Soziale Kompetenz

Kommunikationsverhalten

☐ 1 ☐ 2	☐ 3 ☐ 4 ☐ 5	☒ 6 ☐ 7
Entwicklungsbedarf	Erfüllt Anforderungen	Übertrifft Anforderungen

- ☐ Zeigt unvoreingenommenes Kommunikationsverhalten und partnerschaftliche Bereitschaft zum Dialog.
- ☐ Führt Gespräche strukturiert und zielorientiert.
- ☐ Hört bei Gesprächen aktiv zu und schafft ein angenehmes und wertschätzendes Gesprächsklima.
- ☐ Kann das eigene Kommunikationsverhalten an verschiedene GesprächspartnerInnen und Rahmenbedingungen anpassen.
- ☐ Ist offen im Austausch von Ideen und Meinungen.
- ☐ Argumentiert prägnant und verständlich.

Bemerkungen

Abbildung 7. Auszug aus der Vorgesetztenbeurteilung.

Am Ende der Beurteilung gibt es die Möglichkeit, nächste Entwicklungsschritte festzuhalten. Grundlage der Beurteilung bildet die jeweilige Stellenbeschreibung.

Die Ergebnisse der empirischen Untersuchung

Dieses Kapitel befasst sich mit den Ergebnissen der mittels SPSS ausgewerteten Daten. Die einzelnen Hypothesen werden nacheinander betrachtet und die Ergebnisse kurz skizziert. Eine abschließende Gesamtdiskussion erfolgt im nächsten Kapitel.

Die Konstruktvalidität des Trainee-Assessment Centers (Hypothese 1)

Hypothese 1 bezieht sich auf die Konstruktvalidität des Trainee-Assessment Centers und lautete: *Die Korrelationen der verschiedenen Dimensionen innerhalb einer Übung sind im Trainee-Assessment Center im Durchschnitt deskriptiv höher als die Korrelationen der gleichen Dimensionen über verschiedene Übungen hinweg.* Tabelle 8 zeigt zunächst die einzelnen Korrelationen der Dimensionen innerhalb der Übungen auf. Diese werden im Anschluss kurz erläutert.

Tabelle 8

Durchschnittliche Korrelationskoeffizienten der Dimensionen innerhalb der Übungen im Trainee-Assessment Center

Statistische Maße	Übungen					
	SP	PP	GD	MA	PK	Gesamt
$\bar{r}$	.78	.73	.72	.76	.87	.77
r‘	1.05	.93	.91	1.00	1.33	1.02
$SE_{r'}$	.12	.12	.12	.12	.12	.12
$KI_{r'}$	[.80;1.27]	[.96;1.17]	[.67;1.15]	[.76;1.24]	[1.09;1.57]	[.78;1.25]
KI_r	[.66;86]	[.74;.82]	[.58;.82]	[.64;.85]	[.80;.92]	[.65;.85]

Anmerkung. $\bar{r}$ = durchschnittlicher zweiseitiger Korrelationskoeffizient nach Pearson, r‘ = Fisher-z-transformierter Korrelationskoeffizient, $SE_{r'}$ = Standardfehler von r‘, $KI_{r'}$ = 95 prozentiges Konfidenzintervall von r‘, KI_r = 95 prozentiges Konfidenzintervall von r, SP= Selbstpräsentation, PP= Präsentation, GD= Gruppendiskussion, MA= Mitarbeitergespräch, PK = Postkorb.

Die Korrelationen aller Anforderungsdimensionen in einer Übung liegen im Durchschnitt über alle Übungen hinweg bei $\bar{r}$ = .77 (*r‘* = 1.02). Der höchste Wert wird mit $\bar{r}$ = .87 (*r‘* = 1.33) innerhalb des Postkorbs erzielt, der niedrigste Wert mit $\bar{r}$ = .72 (*r‘* = .91) in der Gruppendiskussion.

Tabelle 9 zeigt die Korrelationen der einzelnen Dimensionen über die verschiedenen Übungen hinweg auf, die dann im Anschluss kurz erläutert werden.

Tabelle 9

Durchschnittliche Korrelationskoeffizienten der Dimensionen über die Übungen hinweg im Trainee-Assessment Center

Statistische Maße	Anforderungsdimensionen					
	KM	KP	KN	IE	SUS	Ü
r̄	.50	.36	.35	.59	.56	.55
r‘	.55	.38	.37	.68	.63	.62
$SE_{r‘}$	.12	.12	.12	.12	.12	.12
$KI_{r‘}$	[.31; .79]	[.15;.62]	[13;.61]	[.44;92]	[.43;87]	[.38;.86]
KI_r	[.30; .66]	[.13; .55]	[13;.54]	[.41;.1.86]	[.41;.70]	[.36;.70]

Statistische Maße	Anforderungsdimensionen				
	A	ENT	SUH	SG	Gesamt
r̄	.15	.11	.27	.47	.39
r‘	.15	.11	.28	.51	.41
$SE_{r‘}$	.12	.12	.12	.12	.12
$KI_{r‘}$	[-09;.39]	[-.13;.35	[.04;.52]	[.27;75	[.17;.65]
KI_r	[-.09;.37]	[-.13;.34]	[.08;.48]	[.26;64]	[.17;57]

Anmerkung. r̄ = durchschnittlicher zweiseitiger Korrelationskoeffizient nach Pearson, r‘ = Fisher-z-transformierter Korrelationskoeffizient, SEr‘ = Standardfehler von r‘, KIr‘ = 95 prozentiges Konfidenzintervall von r‘, KIr = 95 prozentiges Konfidenzintervall von r, KM = Kommunikationsfähigkeit, KP = Kooperationsfähigkeit, KN = Konflikt- und Kritikfähigkeit, IE = Initiative und Engagement, SUS = Stabilität und Selbstvertrauen, Ü = Überzeugungskraft, A = Auftreten und Wirkung, ENT = Entscheidungsfähigkeit, SUH = Strategisches Denken und Handeln, SG = Subjektiver Gesamteindruck.

Die durchschnittliche Korrelation aller Dimensionen über die Übungen hinweg liegt bei $\bar{r}$= .39 ($r‘$ = .41). Wobei ersichtlich wird, dass ein großer Unterschied zwischen den durch-

schnittlichen Korrelationskoeffizienten der einzelnen Dimensionen existiert. Die Werte liegen zwischen $\bar{r}$= .11 (r' = .11) für die Dimension Entscheidungsfähigkeit (ENT) und $\bar{r}$= .59 (r' = .68) für die Dimension Initiative und Engagement (IE). Die Dimensionen Auftreten und Wirkung (A), Entscheidungsfähigkeit (ENT) sowie Strategisches Denken und Handeln (SUH) erzielen im Vergleich zu den anderen Dimensionen relativ niedrige Werte. Die Dimensionen Kommunikationsfähigkeit (KM), Initiative und Engagement (IE), Stabilität und Selbstvertrauen (SUS) und Überzeugungskraft (Ü) weisen sehr hohe Werte mit $r > .50$ auf.

Fasst man die Ergebnisse zusammen, so liegt im Trainee-Assessment Center der durchschnittliche Korrelationskoeffizient der verschiedenen Anforderungsdimensionen innerhalb der gleichen Übung mit $\bar{r}$ = .77 (r' = 1.02) deutlich über dem durchschnittlichen Korrelationskoeffizienten zwischen den gleichen Dimensionen über verschiedene Übungen hinweg mit $\bar{r}$ = .39 (r' = .41). Die Effektstärke beträgt q = .61 und ist nach Cohen (1988) groß.

Die Konstruktvalidität des Stabs-Assessment Centers (Hypothese 2)

Hypothese 2 befasst sich mit der Konstruktvalidität des Stabs-Assessment Centers. Wobei die Herangehensweise und Art der Auswertung wie im Trainee-Assessment Center erfolgt. Die Hypothese wird im nachfolgenden überprüft und sodann ein kurzer Vergleich zu den Werten des Trainee-Assessment Centers genommen.

Hypothese 2 lautete: *Die Korrelationen der verschiedenen Dimensionen innerhalb einer Übung sind im Stabs-Assessment Center im Durchschnitt deskriptiv höher als die Korrelationen der gleichen Dimensionen über verschiedene Übungen hinweg.*

Die Ergebnisse der Korrelationen der einzelnen Dimensionen innerhalb der Übungen sind in Tabelle 10 ersichtlich und werden im Anschluss kurz erläutert.

Tabelle 10

Durchschnittliche Korrelationskoeffizienten der Dimensionen innerhalb der Übungen im Stabs-Assessment Center

Stat.	Übungen						
Maße	SP	PP	GD	RS	KO	PK	Gesamt
$\bar{r}$	.76	.67	.72	.85	.79	.25	.76
r‘	1.00	.81	.91	1.26	1.07	.26	1.00
$SE_{r'}$	.12	.12	.12	.12	.12	.12	.12
$KI_{r'}$	[.76;1.24]	[.57;1.05]	[.67;1.15]	[1.02;1.50]	[.83;1.31]	[.04;.48]	[.76;1.24]
KI_{r}	[.64;.85]	[.52;.78]	[.58;.82]	[.77;.91]	[.68;91]	[.04;.45]	[.64;.85]

Anmerkung. $\bar{r}$ = durchschnittlicher zweiseitiger Korrelationskoeffizient nach Pearson, r‘ = Fisher-z-transformierter Korrelationskoeffizient, SEr‘ = Standardfehler von r‘, KIr‘ = 95 prozentiges Konfidenzintervall von r‘, KIr = 95 prozentiges Konfidenzintervall von r, SP= Selbstpräsentation, PP= Präsentation, GD= Gruppendiskussion, RS = Rollenspiel, KO = Kick-off-Meeting, PK = Postkorb.

Die Korrelationen der verschiedenen Dimensionen innerhalb der gleichen Übung liegen im Durchschnitt bei $\bar{r}$ = .67. Extrahiert man aufgrund des sehr abweichenden Wertes und der elektronischen Erfassung den Postkorb, liegt der durchschnittliche Korrelationskoeffizient bei $\bar{r}$ = .76 (r' = 1.00). Das Rollenspiel (RO) erzielt mit $\bar{r}$ = .85 (r' = 1.26) den höchsten Intra-Korrelationskoeffizienten, nimmt man den Postkorb aus, so ergibt sich für die Präsentationsübung mit $\bar{r}$ = .67 (r' = .81) der geringste Korrelationskoeffizient. In Tabelle 11 sind die Korrelationen zwischen den einzelnen Dimensionen über die verschiedenen Übungen hinweg aufgeführt.

Tabelle 11

Durchschnittliche Korrelationskoeffizienten der Dimensionen über die Übungen hinweg im Stabs-Assessment Center

Stat. Maße	Anforderungsdimensionen						
	KM	KP	KN	KRI	IV	AW	IE
$\bar{r}$	.18	.39	.31	.15	.17	.48	.15
r‘	.18	.41	.32	.15	.17	.52	.15
$SE_{r'}$	.11	.11	.11	.11	.11	.11	.11
$KI_{r'}$	[-.04;.40]	[.19;.32]	[.10;.54]	[-.07;.37]	[-.05;.39]	[.30;.73]	[-.07;.37]
KI_{r}	[-.04;.38]	[.19;.31]	[.10;.49]	[-.07;.34]	[-.05;.37]	[.29;.62]	[-.07;.36]

Stat. Maße	Anforderungsdimensionen						
	IE	ENT	ZE	KMP	UDH	SG	Gesamt
$\bar{r}$	.15	-.02	.28	.16	.25	.34	.24
r‘	.15	-.02	.29	.16	.26	.35	.25
$SE_{r'}$	.11	.11	.11	.11	.11	.13	.11
$KI_{r'}$	[-.07;.37]	[-.24;.20]	[.07;.51]	[-.06;.38]	[.04;.48]	[.13;.57]	[.03;.57]
KI_{r}	[-.07;.36]	[-.24;.20]	[.07;.47]	[-.06;.36]	[.04;.45]	[.13;.52]	[.03;.52]

Anmerkung. $\bar{r}$ = durchschnittlicher zweiseitiger Korrelationskoeffizient nach Pearson, r‘ = Fisher-z-transformierter Korrelationskoeffizient, SEr‘ = Standardfehler von r‘, KIr‘ = 95 prozentiges Konfidenzintervall von r‘, KIr = 95 prozentiges Konfidenzintervall von r, KM = Kommunikationsverhalten, KP = Kooperationsverhalten, KN = Konfliktmanagement, KRI = Kritikfähigkeit, IV = Innovations- und Veränderungsverhalten, AW = Auftreten und Wirkung, IE = Initiative und Engagement, ENT = Entscheidungsfähigkeit, ZE = Ziel- und Ergebnisorientierung, KMP = Komplexitätsmanagement, SG = Subjektiver Gesamteindruck.

Die Korrelationen der gleichen Dimensionen liegen übungsübergreifend im Durschnitt bei $\bar{r}$ = .24 (r' = .25). Sehr auffällig ist, dass die Bandbreite der durchschnittlichen Korrelationskoeffizienten zwischen $\bar{r}$ = -.02 (r' = -.02) für die Dimension Entscheidungsfähigkeit (ENT) und $\bar{r}$ = .48 (r' = .52) für die Dimension Auftreten und Wirkung (AW) liegt.

Fasst man die Ergebnisse zusammen, so liegt im Stabs-Assessment Center der durchschnittliche Korrelationskoeffizient der verschiedenen Anforderungsdimensionen innerhalb der gleichen Übung mit $\bar{r}$ = .76 (r' = 1.00) deutlich über dem durchschnittlichen Korrelationskoeffizienten zwischen den gleichen Dimensionen über verschiedene Übungen hinweg mit $\bar{r}$ = .24 (r' = .25). Die Effektstärke beträgt q = .75, ist nach Cohen (1988) groß und liegt über dem des Trainee-Assessment Centers (q = .61).

Trennschärfe der eingesetzten Assessment Center Übungen im Multimodalen Auswahlverfahren (Hypothese 3)

Hypothese 3 lautete: *Die eingesetzten Assessment Center Übungen als Bausteine des Multimodalen Auswahlverfahrens korrelieren signifikant niedriger miteinander als die vergleichbaren Übungen im Trainee-Assessment Center.*

Ein Vorher – Nachher Vergleich ist lediglich zwischen der Präsentationsübung und dem Kollegengespräch aus dem Trainee Assessment Center möglich. Tabelle 12 zeigt die Korrelationskoeffizienten zwischen der Präsentationsübung und dem Kollegengespräch vor und nach den Überarbeitungen.

Tabelle 12

Korrelationskoeffizienten vor und nach der Überarbeitung der Präsentationsübung und des Kollegengesprächs

Statistische Maße	Trainee Assessment Center: Präsentation & Kollegengespräch (N= 70)	Multimodales Auswahlverfahren: Präsentation & Kollegengespräch (N = 90)
r	.34**	.03
r‘	.35	.03
$SE_{r'}$	.12	.11
$KI_{r'}$	[.12;.59]	[-.19;.25]
KI_{r}	[.12;.53]	[-.19;.25]

Anmerkung. r = zweiseitiger Korrelationskoeffizient nach Pearson. ** Die Korrelation ist auf dem Niveau .01 signifikant. r‘ = Fisher-z-transformierter Korrelationskoeffizient, SEr‘ = Standardfehler von r‘, KIr‘ = 95 prozentiges Konfidenzintervall von r‘, KIr = 95 prozentiges Konfidenzintervall von r.

Sieht man sich den Korrelationskoeffizienten der Präsentationsübung und des Kollegengesprächs im Multimodalen Auswahlverfahren an, so existiert mit $r = .03$ kein signifikanter Zusammenhang der beiden Übungen. Im Trainee-Assessment Center hingegen wird deutlich, dass die beiden Übungen mit $r = .34$ ($\rho < .01$) eine signifikante Korrelation aufweisen.

Die Prüfgröße z der Fishers-z-transformierten Korrelationskoeffizienten liegt bei $z = 1.97$ und somit geringfügig über der kritischen Prüfgröße $z = 1.96$. Bertachtet man die Effektstärke $q = .32$, so ist diese nach Cohen (1988) als moderat einzustufen.

Überprüfung der Interrater-Korrelation in den eingesetzten Auswahlverfahren (Hypothese 4)

Hypothese 4 lautete: *Die Interrater-Korrelation ist im Multimodalen Auswahlverfahren deskriptiv höher als im Trainee- und Stabs-Assessment Center.*

Tabelle 13 gibt einen Überblick über die durchschnittlichen Interrater-Korrelationskoeffizienten im Stabs- und Trainee-Assessment Center sowie im Multimodalen Auswahlverfahren. Die detaillierten Werte sind im Anhang C aufgeführt.

Tabelle 13

Die durchschnittliche Interrater-Korrelation im Stabs- und Trainee-Assessment Center sowie im Multimodalen Auswahlverfahren

Statistische Maße	Trainee-Assessment Center (N= 70)	Stabs-Assessment Center (N= 84)	Multimodales Auswahlverfahren (N = 90)
$\bar{r}$	.70	.76	.70
r‘	.87	1.00	.87
$SE_{r'}$	.12	.12	.11
$KI_{r'}$	[.46;.94]	[.76;1.24]	[.48;.92]
KI_{r}	[.43;.74]	[.64;.85]	[.45;73]

Anmerkung. $\bar{r}$ = durchschnittlicher zweiseitiger Korrelationskoeffizient nach Pearson, r‘ = Fisher-z-transformierter Korrelationskoeffizient, SEr‘ = Standardfehler von r‘, KIr‘ = 95 prozentiges Konfidenzintervall von r‘, KIr = 95 prozentiges Konfidenzintervall von r.

Die Interrater-Korrelation liegt im Trainee-Assessment Center über alle Kompetenzen und Übungen hinweg im Durchschnitt bei $\bar{r}$ = .70 (*r‘* = .87). Im Stabs-Assessment Center liegt diese im Durchschnitt mit $\bar{r}$ = .76 (*r‘* = 1.00) leicht höher. Im Multimodalen Auswahlverfahren wird der gleiche Wert wie im Trainee-Assessment Center mit $\bar{r}$ = .70 (*r‘* = .87) er-

zielt. In allen drei eingesetzten Auswahlverfahren ist die Interrater-Korrelation im Durchschnitt sehr hoch, die Werte liegen sehr eng beieinander.

Hypothese 4.1. lautete: *Die Interrater-Korrelation ist im Multimodalen Auswahlverfahren deskriptiv höher als im Trainee-Assessment Center*. In beiden Auswahlverfahren ist die durchschnittlich erzielte Interrater-Korrelation mit $\bar{r}$ = .70 (r' = .87) gleich hoch. Somit konnte keine Steigerung der Interrater-Korrelation im Multimodalen Auswahlverfahren erzielt werden.

Hypothese 4.2. lautete: *Die Interrater-Korrelation ist im Multimodalen Auswahlverfahren deskriptiv höher als im Stabs-Assessment Center*. Betrachtet man die durchschnittliche Interrater-Korrelation in beiden Verfahren, so wird deutlich, dass der durchschnittliche Korrelationskoeffizient mit $\bar{r}$ = .76 (r' = 1.00) im Stabs-Assessment Center – anders als vermutet – höher als im Multimodalen Auswahlverfahren mit $\bar{r}$ = .70 (r' = .87) ist.

Insgesamt konnte keine Steigerung der Interrater-Korrelation im Multimodalen Auswahlverfahren gegenüber den beiden Assessment Centern erzielt werden.

Überprüfung der prädiktiven Validität der eingesetzten Auswahlverfahren (Hypothese 5)

Hypothese 5 lautete: *Der Zusammenhang des Gesamtdurchschnitts der Vorgesetztenbeurteilung mit dem Gesamtergebnis des Multimodalen Auswahlverfahrens ist signifikant höher ausgeprägt als mit dem Gesamtergebnis des Stabs- und Trainee-Assessment Centers.*

Zur Überprüfung von Hypothese 5 werden zuerst die einzelnen Unterhypothesen im Einzelnen betrachtet. Die ersten drei Unterhypothesen sind erforderlich, um zu ermitteln, ob überhaupt ein Zusammenhang der Gesamtergebnisse der eingesetzten Auswahlverfahren mit der Vorgesetztenbeurteilung besteht.

Die erste Unterhypothese 5.1. lautete: *Im Trainee-Assessment Center existiert ein signifikanter Zusammenhang des Assessment Center-Gesamtergebnisses mit dem Gesamtdurchschnitt der Vorgesetztenbeurteilung.*

Der im Anhang D ersichtliche Korrelationskoeffizient zwischen dem Trainee-Assessment Center-Gesamtergebnis und der Vorgesetztenbeurteilung gesamt zeigt, dass mit $r = .14$ zwar ein schwacher, jedoch mit $p = .54$ nicht signifikanter Zusammenhang besteht (Field, 2013, S. 270; Wirtz, 2002, S. 107).

Die zweite Unterhypothese 5.2 lautete: *Im Stabs-Assessment Center existiert ein signifikanter Zusammenhang des Assessment Center-Gesamtergebnisses mit dem Gesamtdurchschnitt der Vorgesetztenbeurteilung.*

Der ebenso im Anhang D ersichtliche Korrelationskoeffizient des Stabs-Assessment Center-Gesamtergebnisses und der Vorgesetztenbeurteilung zeigt, dass dieser mit $r = .25$ etwas höher als im Trainee-Assessment Center ist und von einem schwachen bis mittleren Zusammenhang gesprochen werden kann (Field, 2013, S. 270; Wirtz, 2002, S. 107). Jedoch ist dieser mit $p = .26$ auch hier nicht signifikant.

Die dritte Unterhypothese lautete: *Im Multimodalen Auswahlverfahren existiert ein signifikanter Zusammenhang des Assessment Center Gesamtergebnisses mit dem Gesamtdurchschnitt der Vorgesetztenbeurteilung.*

Sieht man sich den Korrelationskoeffizienten in Anhang D an, wird deutlich, dass mit $r = .32$ ein über den beiden Assessment Centern liegender, moderater Zusammenhang des Assessment Center-Gesamtergebnisses mit dem Gesamtergebnis der Vorgesetztenbeurteilung besteht (Field, 2013, S. 270; Wirtz, 2002, S. 107). Das Signifikanzniveau liegt bei $p = .08$ und ist somit auch nicht signifikant.

Tabelle 17 fasst die Modelle der Regressionsgleichungen zusammen. Im Anschluss werden diese kurz erläutert und die einzelnen Regressionsgeraden dargestellt.

Tabelle 17

Modellzusammenfassung zur prädiktiven Validität der eingesetzten Auswahlverfahren

Statistische Maße	Trainees (N= 21)	Referenten (N= 23)	Multimodal (N= 31)
R	.14	.25	.32
R^2	.02	.06	.10
Angepasstes R^2	-.03	.02	.07
Standardfehler der Schätzung	.73	.62	.66

Anmerkung. *R* = Korrelation zwischen Prädiktor (Gesamtergebnis Assessment Center) und abhängiger Variable (Vorgesetztenbeurteilung), R^2 = Determinationskoeffizient.

Betrachtet man die Werte im Trainee-Assessment Center, so liegt das multiple *R* bei .14 und spiegelt den Zusammenhang zwischen Prädiktor (Gesamtergebnis Assessment Center) und Kriterium wider (Vorgesetztenbeurteilung). Der Determinationskoeffizient ist mit R^2 = .02 sehr gering und bedeutet, dass nur zwei Prozent der Variation des Gesamtergebnisses der Vorgesetztenbeurteilung durch den Regressor (Gesamtergebnis im Assessment Center) vorhergesagt werden. Das angepasste R^2 wird sogar leicht negativ. Der Standardfehler der Schätzung liegt bei .73 und bedeutet, dass der geschätzte Wert von Y (Gesamtergebnis der Vorgesetztenbeurteilung) im Mittel um .73 Punkte vom wahren Wert abweicht. Sieht man sich die in Abbildung 8 dargestellte Regressionsgerade an, wird deutlich, dass der geringe Zusammenhang zwischen dem Gesamtergebnis des Assessment Centers und der Vorgesetztenbeurteilung auch nicht durch Ausreißer zustande gekommen ist.

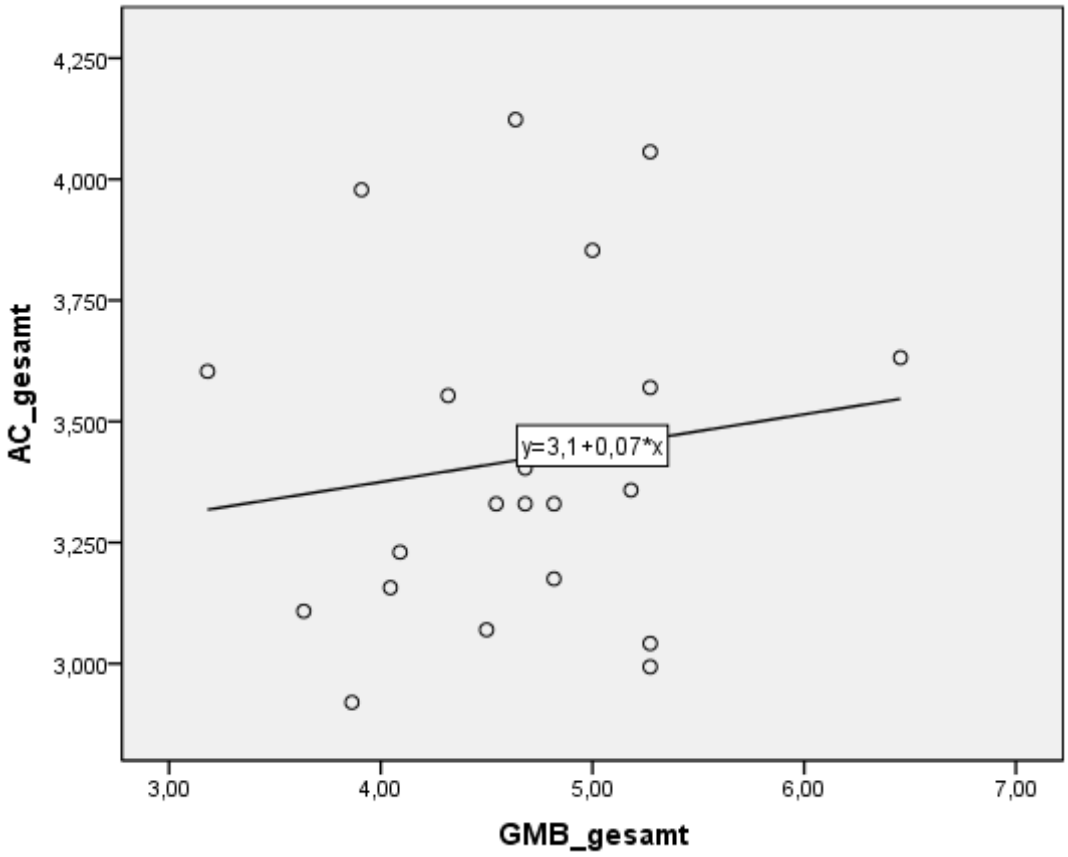


Abbildung 8. Regressionsgerade zur prädiktiven Validität des Trainee-Assessment Centers. Auf der Y-Achse sind die Werte des Assessment Center-Gesamtergebnisses ersichtlich. Auf der X-Achse die Gesamtergebnisse der Vorgesetztenbeurteilung.

Die im Anhang D ersichtlichen, standardisierten Koeffizienten und Beta-Gewichte der einzelnen Übungen im Trainee-Assessment Center zeigen deutlich, dass die Selbstpräsentation mit β = .25 und die Gruppendiskussion mit β = .27 die höchsten, allerdings nicht signifikanten Werte aufweisen. Das Mitarbeitergespräch (Kollegengespräch) liegt mit β = .14 sowie der Postkorb mit β = .03 deutlich darunter. Die Präsentationsübung weist mit β = -.28 sogar einen negativen Wert auf. Die Korrelationen nullter Ordnung sowie die partiellen Korrelationen weisen ein ähnliches Bild auf. Die Selbstpräsentation und die Gruppendiskussion erzielen die höchsten Werte, gefolgt von dem Mitarbeitergespräch. Der Korrelationskoeffizient für die Präsentationsübung bleibt auch unter Betrachtung der Korrelation nullter Ordnung und der partiellen Korrelation auf einem ähnlich hohen Niveau. Auf-

fällig ist, dass die Korrelation nullter Ordnung für den Postkorb negativ wird, die partielle Korrelation hingegen ist wiederum positiv.

Die in Tabelle 17 dargestellten statistischen Maße zeigen, dass im Stabs-Assessment Center das Multiple R = .25 und der Determinationskoeffizient mit R^2 = .06 geringfügig höher als im Trainee-Assessment Center sind. Dies bedeutet, dass zum einem der Zusammenhang zwischen Prädiktor und Kriterium höher ist und in diesem Fall sechs Prozent der Variation des Gesamtergebnisses der Vorgesetztenbeurteilung durch den Regressor (Gesamtergebnis im Assessment Center) vorhergesagt werden. Das angepasste R^2 liegt nur bei .02. Der Standardfehler der Schätzung liegt bei .62 und bedeutet, dass der geschätzte Wert von Y (Gesamtergebnis der Vorgesetztenbeurteilung) im Mittel um .62 Punkte vom wahren Wert abweicht. Sieht man sich die in Abbildung 9 dargestellte Regressionsgerade an, wird deutlich, dass es zwar drei Ausreißer gibt, die aber aufgrund mangelnder Hebelwirkung nicht zu einer starken Beeinflussung der Regressionsgleichung führen.

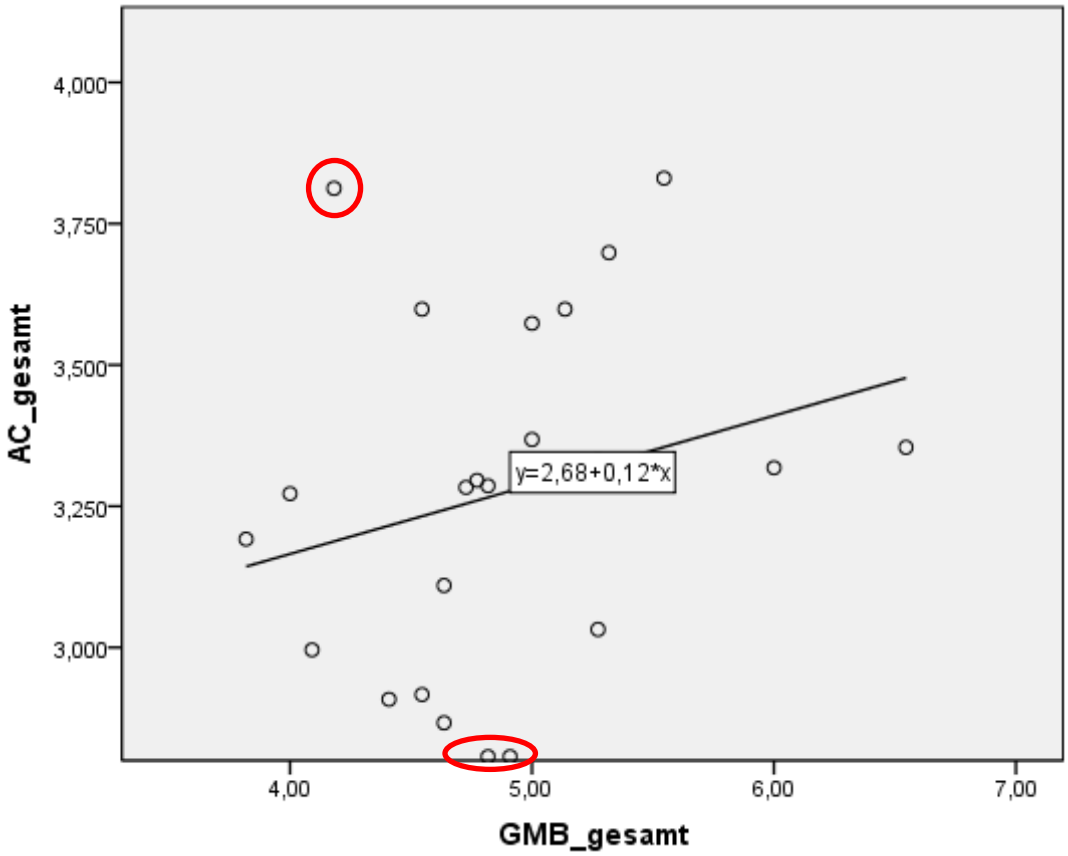


Abbildung 9. Regressionsgerade zur prädiktiven Validität des Stabs-Assessment Centers. Auf der Y-Achse sind die Werte des Assessment Center-Gesamtergebnisses ersichtlich. Auf der X-Achse die Gesamtergebnisse der Vorgesetztenbeurteilung.

Die im Anhang D ersichtlichen, standardisierten Koeffizienten und die Beta-Gewichte der einzelnen Übungen zeigen, dass auch im Stabs-Assessment Center die Gruppendiskussion mit β = .58 den höchsten und mit p = .04 sogar signifikanten Wert aufweist. Der Postkorb erzielt mit β = .28 den zweithöchsten, jedoch nicht signifikanten Wert. Die Selbstpräsentation weist in diesem Assessment Center mit β = -.42 einen deutlich schlechteren Wert auf. Das Kick-off Meeting erzielt mit β = -.16 ebenso einen negativen Wert. Das Rollenspiel und die Präsentationsübung erzielen mit β = .12 und β = .03 ebenso geringe, nicht signifikante Werte. Die Korrelationen nullter Ordnung und die partiellen Korrelationen verhalten sich in diesem Assessment Center etwas anders. Die Selbstpräsentation erzielt in beiden Fällen einen negativen Wert, betrachtet man allerdings die Korrelation nullter Ordnung,

wird deutlich, dass diese mit r = -.17 deutlich geringer negativ wird. Das Rollenspiel erzielt bei der Korrelation nullter Ordnung einen leicht negativen Wert. Im Kick-off Meeting ist besonders auffällig, dass die Korrelation nullter Ordnung positiv wird, die partielle allerdings negativ ist. Der Postkorb senkt die Korrelation nullter Ordnung deutlich. Die Gruppendiskussion und die Präsentationsübung sind im Hinblick auf die einzelnen Korrelationskoeffizienten eher unauffällig.

Betrachtet man die in Tabelle 17 dargestellten Werte im Multimodalen Auswahlverfahren, so liegt das Multiple R = .32 deutlich über den beiden Assessment Centern. Der Determinationskoeffizient R^2 = .10 ist nur geringfügig höher als in den beiden Assessment Centern. Hier wird zumindest zehn Prozent der Variation des Gesamtergebnisses der Vorgesetztenbeurteilung durch den Regressor (Gesamtergebnis im Assessment Center) vorhergesagt. Das angepasste R^2 liegt bei .07. Der Standardfehler der Schätzung liegt bei .66 und bedeutet, dass der geschätzte Wert von Y (Gesamtergebnis der Vorgesetztenbeurteilung) im Mittel um .66 Punkte vom wahren Wert abweicht. Sieht man sich die in Abbildung 10 dargestellte Regressionsgerade an, wird auch hier deutlich, dass es zwei Ausreißer gibt. Auch diese beeinflussen das Ergebnis aufgrund mangelnder Hebelwirkung nur in minimalem Maße.

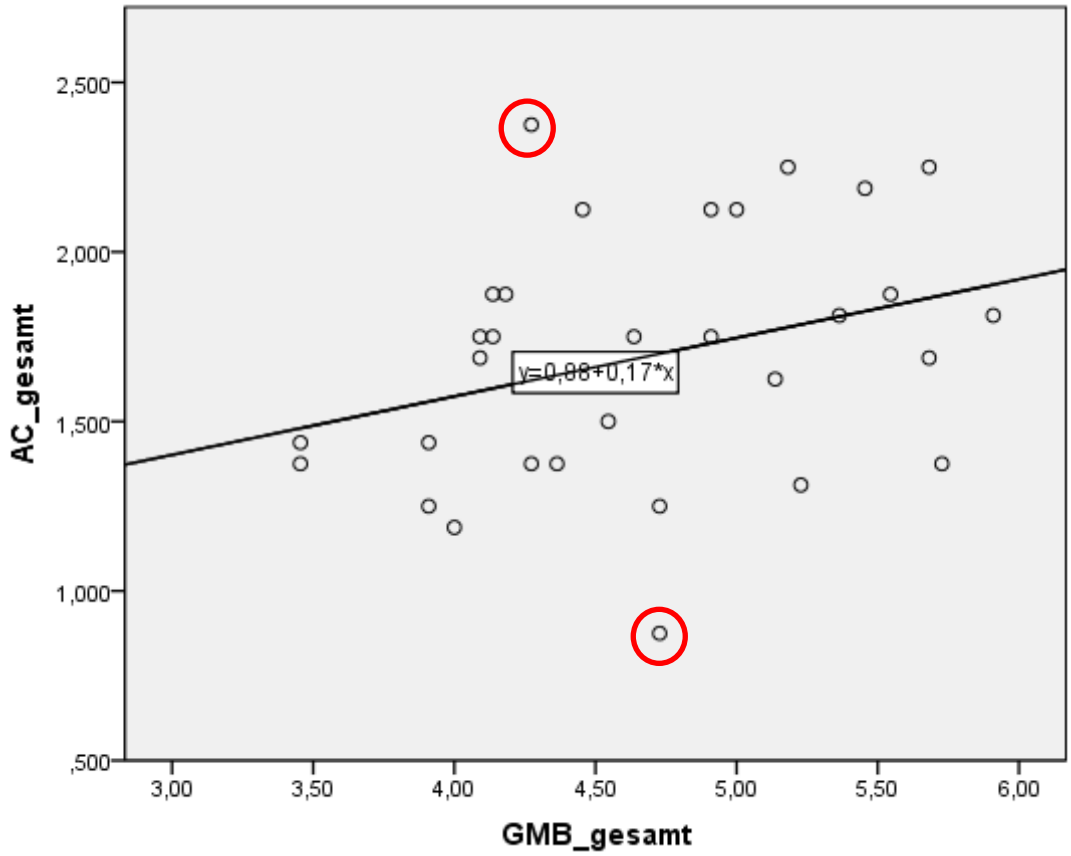


Abbildung 10. Regressionsgerade zur prädiktiven Validität des Multimodalen Auswahlverfahrens. Auf der Y-Achse sind die Werte des Assessment Center-Gesamtergebnisses ersichtlich. Auf der X-Achse die Gesamtergebnisse der Vorgesetztenbeurteilung.

Die im Anhang D ersichtlichen, standardisierten Koeffizienten und Beta-Gewichte der einzelnen Übungen im Multimodalen Auswahlverfahren zeigen deutlich, dass in diesem Assessment Center das Beratungsgespräch mit β = .57 und das Kollegengespräch mit β = .52 die höchsten, jedoch nicht signifikanten Werte erzielen. Diese beiden Gesprächsübungen haben in den beiden Assessment Centern zuvor sehr geringe Werte, nur knapp über β = .10, aufgewiesen. Die Präsentationsübung konnte gegenüber dem Trainee-Assessment Center ebenso eine Verbesserung erzielen und liegt mit β = .33 knapp hinter der Gesprächsübung. Im Anschluss folgen mit einem sehr geringen Anteil die beiden Testbausteine des IST-2000-R, die numerischen und verbalen Fähigkeiten, mit einem Wert von β = .05 und β = .01. Die Bausteine des strukturierten Interviews, Interview 1 und 2, erzie-

len mit β = -.12 und β = -.17 einen negativen Wert. Sieht man sich die Korrelationen nullter Ordnung und die partiellen Korrelationskoeffizienten an, wird deutlich, dass für die beiden Testbausteine die Korrelation nullter Ordnung in beiden Fällen ansteigen. Die Korrelation nullter Ordnung sinkt leicht für Interview 1 und die von Interview 2 erhöht sich leicht. Die Korrelationen nullter Ordnung für die beiden Gesprächsübungen sinken gravierend, insbesondere im Kollegengespräch, hier wird nur noch ein r = .03 erreicht. Für das Beratungsgespräch sinkt der Wert auf r = .21.

Zur Überprüfung der letzten beiden Unterhypothesen wird Bezug auf die in Anhang D ersichtlichen Korrelationskoeffizienten genommen.

Hypothese 5.4 lautete: *Der Zusammenhang des Gesamtdurchschnitts der Vorgesetztenbeurteilung mit dem Gesamtergebnis des Multimodalen Auswahlverfahrens ist signifikant höher ausgeprägt als mit dem Gesamtergebnis des Trainee-Assessment Centers.*

Ein Vergleich der beiden Korrelationskoeffizienten zeigt, dass der Korrelationskoeffizient für das Trainee Assessment Center mit r = .14 (r' = .14, $SE_{r'}$ = .11) niedriger ist als im Multimodalen Auswahlverfahren mit r = .32 (r' = .33, $SE_{r'}$ = .12). Die Prüfgröße z der Fishers-z-transformierten Korrelationskoeffizienten liegt allerdings bei z = 0.63 (p = .26) und somit deutlich unter der kritischen Prüfgröße z = 1.96. Der Unterschied ist folglich nicht signifikant.

Hypothese 5.5 lautete: *Der Zusammenhang des Gesamtdurchschnitts der Vorgesetztenbeurteilung mit dem Gesamtergebnis des Multimodalen Auswahlverfahrens ist signifikant höher ausgeprägt als mit dem Gesamtergebnis des Stabs-Assessment Centers.*

Ein Vergleich der beiden Korrelationskoeffizienten zeigt, dass der Korrelationskoeffizient für das Trainee Assessment Center mit r = .25 (r' = .26, $SE_{r'}$ = .11) niedriger ist als im Multimodalen Auswahlverfahren mit r = .32 (r' = .33, $SE_{r'}$ = .12). Die Prüfgröße z der Fishers-z-transformierten Korrelationskoeffizienten liegt allerdings bei z = 0.26 (p = .40)

und somit deutlich unter der kritischen Prüfgröße $z = 1.96$. Der Unterschied ist folglich auch hier nicht signifikant.

Insgesamt zeigt sich, dass im Multimodalen Auswahlverfahren zwar eine Steigerung von $R = .15$ bzw. $R = .25$ auf $R = .32$ erzielt werden konnte, der Unterschied jedoch nicht signifikant ist. Mögliche Gründe dafür werden in der Gesamtdiskussion genannt.

Gesamtdiskussion

Die im vorangegangenen Kapitel skizzierten Ergebnisse der Hypothesenüberprüfung werden in diesem Kapitel noch einmal zusammengefasst und diskutiert. Zudem werden die Limitationen dieser Untersuchung erläutert und Implikationen für weitere Studien gegeben. Den Abschluss bildet ein Ausblick für weitere Forschungsrichtungen in Bezug auf Instrumente zur Prognose des beruflichen Erfolgs.

Zusammenfassung und Interpretation der Ergebnisse

Im nachfolgenden werden die einzelnen Ergebnisse der im vorherigen aufgeführten Kapitel im Detail betrachtet und interpretiert sowie in einen Gesamtzusammenhang gebracht.

Beginnend mit Hypothese 1, die sich mit der Konstruktvalidität des Trainee-Assessment Centers befasst, wird folgendes Ergebnis erzielt: Insgesamt zeigt sich über alle Übungen hinweg, dass der durchschnittliche Korrelationskoeffizient der verschiedenen Anforderungsdimensionen innerhalb derselben Übung mit $\bar{r}$ = .77 (r' = 1.02) deutlich über dem durchschnittlichen Korrelationskoeffizienten der gleichen Anforderungsdimensionen über verschiedene Übungen hinweg mit $\bar{r}$ = .39 (r' = .41) liegt. Die Beobachter vergeben im Trainee-Assessment Center überwiegend Pauschalurteile innerhalb einer Übung und differenzieren kaum zwischen den einzelnen Anforderungsdimensionen. Dieselben Anforderungsdimensionen über verschiedene Übungen hinweg werden hingegen sehr unterschiedlich und in Abhängigkeit des Gesamturteils einer Übung beurteilt. Dies widerspricht dem ursprünglichen Gedanken dieses Assessment Centers, nämlich kompetenzbezogene Stärken und Schwächen zu erfassen und darauf basierend Einstellungsentscheidungen zu treffen. Darüber hinaus kann das Feedback an die Kandidaten, zu kompetenzbezogenen Stärken und Schwächen, zu falschen Rückmeldungen führen. Sind kompetenzbezogene

Stärken das entscheidende Einstellungskriterium für den jeweiligen Fachbereich, kann dies zu Fehlentscheidungen führen.

Im Folgenden wird noch genauer auf Auffälligkeiten bei den ermittelten Korrelationskoeffizienten eingegangen. Die Postkorbübung erzielt mit $\bar{r}$ = .87 (r' = 1.33) die höchste durchschnittliche Korrelation der Dimensionen innerhalb der Übung. Dieses Ergebnis weist darauf hin, dass die drei innerhalb dieser Übung erfassten Dimensionen Auffassungsgabe, Entscheidungsfähigkeit sowie strategisches Denken und Handeln pauschal ähnlich von den Beobachtern bewertet werden. Dies zeigen die im Anhang A einzeln ersichtlichen Korrelationskoeffizienten. Das Ergebnis ist wahrscheinlich darauf zurückzuführen, dass der Spielraum für die Beobachter für eine unterschiedliche Bewertung in dieser Übung relativ gering ist. Die Lösungsergebnisse in dieser Übung sind für die Beobachter schriftlich zusammengefasst und müssen nur auf ihre Korrektheit überprüft werden. Sind diese überwiegend korrekt, erhält der Kandidat in allen gemessenen Dimensionen einen hohen Wert.

Die Gruppendiskussion erzielt mit $\bar{r}$ = .72 (r' = .91) den niedrigsten Wert. Dies ist dadurch zu erklären, dass in dieser Übung viele verschiedene Aspekte abgeprüft werden. Betrachtet man beispielsweise die Dimension Initiative und Engagement, so wird ein Wortführer in diesem Kriterium aller Wahrscheinlichkeit nach hoch bewertet, was aber im Extremfall zu Lasten der Kooperationsfähigkeit und folglich zu einer etwas niedrigen Bewertung in dieser Dimension führen kann. Dies wird im Anhang A mit einer Korrelation dieser beiden Dimensionen von r = .64 bestätigt.

Die Korrelationskoeffizienten der einzelnen Übungen zeigen deutlich, dass die im Postkorb erfassten Dimensionen Auffassungsgabe, Entscheidungsfähigkeit, strategisches Denken und Handeln die niedrigsten Korrelationswerte mit $\bar{r}$ = .15, $\bar{r}$ = .11. und $\bar{r}$ = .27 erzielen. Dies deutet darauf hin, dass die im Postkorb erfassten Dimensionen von den Be-

obachtern anders bewertet bzw. etwas anders verstanden werden als in den anderen Übungen. Die Dimensionen Kommunikationsfähigkeit, Initiative und Engagement, Stabilität und Selbstvertrauen sowie Überzeugungskraft weisen sehr hohe Werte mit $r > .50$ auf. Auffällig ist, dass diese vier Dimensionen alle in der Selbstpräsentation erfasst werden. Dies deutet daraufhin, dass diese Kompetenzen an sich gut funktionieren und für den Beobachter klarer zu sein scheint, was unter diesen Dimensionen erfasst wird.

Eine Betrachtung der Konstruktvalidität des Stabs-Assessment Centers zeigt ein ähnliches Bild wie im Trainee Assessment Center. Der durchschnittliche Korrelationskoeffizienten über alle Übungen hinweg liegt mit $\bar{r} = .76$ ($r' = 1.02$) auf einem ähnlichen Niveau wie das Trainee Assessment Center mit $\bar{r} = .77$ ($r' = 1.02$). Im Durchschnitt nicht mit eingerechnet ist hier allerdings der deutlich abweichende Wert mit $\bar{r} = .25$ ($r' = .26$). Dieser würde den durchschnittlichen Korrelationskoeffizienten deutlich verringern. Da dieser aber elektronisch ausgewertet und nicht von den Beobachtern beurteilt wird, handelt es sich hierbei um eine nicht vergleichbare Übung und wird deshalb nicht mit eingerechnet.

Die Gruppendiskussion erzielt wie im Trainee-Assessment Center mit $\bar{r} = .72$ ($r' = .91$) den geringsten Wert. Der Effekt scheint der gleiche zu sein. Das Rollenspiel weist mit $\bar{r} = .85$ ($r' = 1.26$) den höchsten Wert auf. Hier scheint es für die Beobachter am schwierigsten zu sein, zwischen den einzelnen Anforderungsdimensionen zu differenzieren.

Im elektronischen Postkorb scheint die Differenzierung der erfassten Anforderungsdimensionen Initiative und Engagement, Entscheidungsfähigkeit und Komplexitätsmanagement gut zu funktionieren, da die im Anhang ersichtlichen Korrelationskoeffizienten zwischen $r = .18$ und $r = .31$ liegen.

Bei Betrachtung der einzelnen Korrelationskoeffizienten der gleichen Anforderungsdimensionen sind die Werte deutlich niedriger als im Trainee-Assessment Center ($\bar{r} = .39$

(r' = .41)). Der Durchschnitt liegt im Stabs-Assessment Center bei $\bar{r}$ = .24 (r' = .25). Die Dimension Entscheidungsfähigkeit weist sogar einen negativen Zusammenhang von $\bar{r}$ = -.02 (r' = -.02) auf. Dies heißt, dass tendenziell ein hoher Wert dieser Dimension in einer Übung zu einem niedrigen Wert dieser Dimension in einer anderen Übung führt. Gemessen wird diese in der Präsentations- und Postkorbübung. Auch die weiteren im Postkorb erfassten Dimensionen Initiative und Engagement sowie Komplexitätsmanagement weisen sehr geringe Werte mit $\bar{r}$ = .15 (r' = .15) bzw. $\bar{r}$ =.16 (r' = .16) auf. Dies legt nahe, dass im Postkorb eine andere Beurteilung erfolgt als in den anderen Übungen. Daneben liegen die Werte für Kommunikationsverhalten mit r ¯=. 18 (r' = .18), Kritikfähigkeit mit r ¯ = .15 (r' = .15), Innovations- und Veränderungsverhalten (IV) mit r ¯ = .17 (r' = .17) auf einem ähnlich niedrigen Niveau. Dies deutet daraufhin, dass diese Kompetenzen an sich schlecht funktionieren und den Beobachtern nicht klarer zu sein scheint, was unter diesen Dimensionen erfasst wird. Die Anforderungen an diese Dimensionen in einer Übung scheinen andere zu sein, als in den anderen Übungsbausteinen. Der höchste Wert wird mit r ¯ = .48 (r' = .41) für die Dimension Auftreten und Wirkung erzielt. Dies scheint damit die einzige Dimension zu sein, die klar formuliert und wohl am besten operationalisiert ist.

Auch im Stabs-Assessment Center gilt, dass der durchschnittliche Korrelationskoeffizient der verschiedenen Anforderungsdimensionen innerhalb der gleichen Übung mit $\bar{r}$ = .76 (r' = 1.00) deutlich über dem durchschnittlichen Korrelationskoeffizient zwischen den gleichen Dimensionen über verschiedene Übungen hinweg mit $\bar{r}$ = .24 (r' = .25) liegt. Es gilt auch hier, dass im Rahmen des Stabs-Assessment Centers sehr stark übungsbezogen und kaum dimensionsbezogen beurteilt wird. Der Unterschied ist insbesondere bei den Korrelationskoeffizienten für die gleichen Anforderungsdimensionen noch gravierender als im Trainee-Assessment Center. Die Grundannahme des Konstruktes, kompetenzbezogene

Stärken und Schwächen der Kandidaten zu ermitteln, ist auch in diesem Assessment Center nicht erfüllt. Die Folgen für ein kompetenzbezogenes Feedback an die Kandidaten können zu noch größeren Fehlrückmeldungen und Fehlentscheidungen in Bezug auf die Einstellung führen.

Fasst man die Ergebnisse der Hypothesenüberprüfung 1-2 zusammen, so ergibt sich ein klares Bild. In den beiden klassischen Assessment Centern wird deutlich, dass die Beobachter weniger dimensionsbezogen, sondern stark übungsbezogen bewerten, was dem ursprünglichen Konstrukt des eigenschaftstheoretischen Ansatzes und der zugrunde liegenden Annahme der beiden Assessment Centern widerspricht. Des Weiteren bergen diese beiden Assessment Center die Gefahr, dass die Kandidaten ein falsches kompetenzbezogenes Feedback erhalten und Fehleinstellungen erfolgen. Zur Vermeidung dieser Fehler wurde den Assessment Center Übungen im Multimodalen Auswahlverfahren ein übungsbezogenes Beurteilungsprinzip zu Grunde gelegt.

Eine weitere Zielsetzung der Überarbeitung der Übungen im Multimodalen Auswahlverfahren aus den beiden Assessment Centern war eine höhere Trennschärfe (Hypothese 3). Sieht man sich den Korrelationskoeffizienten der Präsentationsübung mit dem Kollegengespräch im Trainee-Assessment Center an, wird deutlich, dass die beiden Übungen eine signifikante Korrelation von $r = .34$ aufweisen und somit ein mittelstarker Zusammenhang der beiden Übungen besteht (Cohen, 1988). Dies bedeutet, dass diese beiden Übungen im Trainee-Assessment Center nicht trennscharf formuliert und teilweise redundant sind. Im Multimodalen Auswahlverfahren existiert hingegen mit $r = .03$ kein signifikanter Zusammenhang der beiden Übungen. D.h. durch eine bessere Operationalisierung und Schärfung der Zielsetzungen der beiden Übungen kann eine höhere Trennschärfe erzielt werden. Auch das übungsbezogene Beurteilungsprinzip statt des dimensionsbezogenen mag einen positiven Einfluss auf die Trennschärfe der beiden Übungen haben. Auf-

grund der Signifikanz des Korrelationsunterschiedes liegt nahe, dass ein solcher Effekt auch bei der gleichen Überarbeitungsweise für andere Übungen erzielt werden könnte.

Die Annahme, dass aufgrund der dimensionsbezogenen Beurteilung im Trainee- und Stabs-Assessment Center die Interrater-Korrelation geringer ist als im Multimodalen Auswahlverfahren, konnte nicht bestätigt werden (Hypothese 4). In allen drei eingesetzten Auswahlverfahren ist die Interrater-Korrelation im Durchschnitt hoch und liegt auf einem ähnlichen Niveau wie in den im Theorieteil skizzierten Studien. Die Werte liegen in allen drei Verfahren sehr eng beieinander. Im Stabs-Assessment Center liegt der Durchschnitt mit $\bar{r}$ = .76 (r' = 1.00) leicht höher als im Trainee-Assessment Center und dem Multimodalen Auswahlverfahren mit $\bar{r}$ = .70 (r' = .87). Dies ist vermutlich darauf zurückzuführen, dass die Beobachter in allen drei Verfahren sehr gut und einheitlich geschult wurden und sehr geübt sind. Darüber hinaus handelt es sich um einen sehr kleinen Kreis an Beobachtern, die im Rahmen aller Verfahren teilgenommen haben. Folglich sollte auch das Beobachtungsniveau relativ konstant sein. Auch die Zusammensetzung der Beobachterteams mit jeweils einem Fachbereichsbeobachter und einer Person aus dem Personalbereich, wirkte sich in allen drei Verfahren vermutlich vorteilhaft auf die Interrater-Korrelation aus.

Die Hypothese 5 befasste sich mit der prädiktiven Validität der eingesetzten Auswahlverfahren. Es wurde die Vermutung angestellt, dass aufgrund der Überarbeitungen der Assessment Center-Übungen und des Hinzufügens eines strukturierten Interviews und zwei Bausteinen eines Intelligenztests das multimodal konzipierte Auswahlverfahren eine höhere prädiktive Validität besitzt als die beiden zuvor eingesetzten klassischen Assessment Center. Insgesamt wurde eine Steigerung von R = .15 (Trainee-Assessment Center) bzw. R = .25 (Stabs-Assessment Center) auf R =.32 (Multimodales Auswahlverfahren) erzielt. Diese Werte könnten in allen drei Verfahren über R > .50 gesteigert werden, wenn man eine andere Gewichtung der einzelnen Übungen heranziehen würde (siehe Anhang

D). Um in dieser empirischen Untersuchung jedoch einen fairen Vergleich vorzunehmen, ist für alle drei Verfahren das erzielte Gesamtergebnis als Prädiktor der Vorgesetztenbeurteilung ermittelt worden. Im nachfolgenden werden die im Ergebnisteil skizzierten Regressionskoeffizienten für die einzelnen Verfahren diskutiert und in einen Gesamtzusammenhang gebracht, wobei der Hauptfokus auf dem neuen, Multimodalen Auswahlverfahren liegt.

Unter Betrachtung der einzelnen Bausteine im Trainee-Assessment Center wird deutlich, dass die Selbstpräsentation und die Gruppendiskussion ein hohes Beta-Gewicht aufwiesen, das gleiche gilt für die Korrelation nullter Ordnung. Das Mitarbeitergespräch (Kollegengespräch) trägt einen ebenso positiven, allerdings deutlich geringeren Beitrag bei. Die Präsentationsübung wirkt sich auch unter Betrachtung der Korrelation nullter Ordnung negativ aus. Dies mag daran liegen, dass die Präsentation eine hohe Korrelation mit der Selbstpräsentation in Höhen von r = .37 (ρ =.05) und der Gruppendiskussion in Höhe von r = .34 (ρ =.06) aufweist. D.h. diese Übung hat in diesem Assessment Center keinen Mehrwert geleistet und hätte gestrichen werden können. Interessant ist der Postkorb, dieser erzielt ein leicht positives β-Gewicht, die Korrelation nullter Ordnung ist jedoch leicht negativ. Betrachtet man die Intra-Korrelationskoeffizienten mit den anderen Übungen, sind diese, bis auf das Mitarbeitergespräch in geringem Maße, alle negativ. D.h. diese Übung ist weitestgehend trennscharf zu den anderen Bausteinen, jedoch korreliert der im Postkorb erzielte Wert negativ mit dem Gesamtdurchschnitt der Vorgesetztenbeurteilung. Zusammengefasst waren in diesem Verfahren die Gruppendiskussion, die Selbstpräsentation und das Kollegengespräch einzeln betrachtet die besten Prädiktoren für beruflichen Erfolg (gemessen durch die Vorgesetztenbeurteilung) in dem zu Grunde liegenden Unternehmen.

Die einzelnen Bausteine im Stabs-Assessment Center zeigen ein sehr durchwachsenes Bild auf. Die Gruppendiskussion weist ungeschlagen ein hohes β-Gewicht auf, gefolgt

von der Postkorbübung und einem deutlich geringeren Wert des Rollenspiels (Beratungsgespräch). Betrachtet man jedoch die Korrelation nullter Ordnung, erzielt der Postkorb mit r = .10 einen eher geringeren Wert. Dies mag an der positiven Intra-Korrelation dieser Übung mit der Selbstpräsentation (r = .18), des Kick-off-Meetings (r = .18) und dem Rollenspiel (r = .05) liegen. Die Präsentationsübung erzielt ein sehr geringes β-Gewicht und eine etwas höhere Korrelation nullter Ordnung in Höhe von r = .11. Die Selbstpräsentation und das Kick-off-Meeting erzielen in diesem Assessment Center ein negatives β-Gewicht. Betrachtet man allerdings die Korrelation nullter Ordnung, liegt diese für das Kick-Off-Meeting im deutlich positiven Bereich (r = .20). Dies mag an der positiven Intra-Korrelation dieser Übung mit der Gruppendiskussion (r = .28), dem Rollenspiel (r = .13) und dem Postkorb (r = .18) liegen. Zusammengefasst waren in diesem Verfahren die Gruppendiskussion, das Kick-off Meeting, die Präsentationsübung und der Postkorb einzeln betrachtet die besten Prädiktoren für beruflichen Erfolg (gemessen durch die Vorgesetztenbeurteilung) in dem zu Grunde liegenden Unternehmen. Jedoch gab es deutliche Überschneidungen des Kick-off-Meetings und des Postkorbs mit den anderen Übungsbausteinen, so dass diese teilweise redundant waren. Insgesamt ist das multiple R für dieses Verfahren gesamt zwar höher als für das Trainee-Assessment Center, jedoch waren die einzelnen Übungen überwiegend nicht trennscharf konzipiert.

Betrachtet man die einzelnen Bausteine des Multimodalen Auswahlverfahrens, so konnte das strukturierte Interview, anders als aufgrund des Theorieteils vermutet, mit β = -.12 und β = -.17 keine Steigerung der prädiktiven Validität erzielen. Dies bestätigen auch die Korrelationskoeffizienten nullter Ordnung. Betrachtet man die Intra-Korrelationskoeffizienten im gesamten Verfahren, so wird deutlich, dass Interview 1 mit der Präsentationsübung zu r = .26 (ρ =.08) und mit Interview 2 zu r = .49 (ρ =.003) signifikant korreliert (siehe Anhang D). Interview 2 könnte somit gestrichen werden, da dies bereits in ho-

hem Maße durch Interview 1 abgedeckt wird. Fraglich ist, ob aufgrund der Korrelation von Interview 1 und der Präsentation und des negativen β-Gewichtes auch dieser Baustein gestrichen werden sollte. Weitere Gründe, die gegen den Einsatz des gesamten Interviews sprechen, sind, dass dieser Übungsbaustein von den Beobachtern und Teilnehmern nicht gut angenommen wurde. Den Teilnehmern war in der Durchführung oftmals nicht klar, was mit diesem Baustein bezweckt wurde und eine Beantwortung der Fragen war meistens nur durch erhöhte Hilfestellung seitens der Beobachter möglich. Folglich ist dann im Anschluss den Beobachtern eine Beurteilung der Kandidaten in dieser Übung entsprechend schwer gefallen. Darüber hinaus war den Beobachtern das Instrument zuvor nicht bekannt, was die Beurteilung ebenso erschwert hat. Neben der Bekanntheit ist auch den meisten Beobachtern die Zielsetzung dieser Übung trotz Schulung nicht eindeutig klar geworden. Sollte man diesen Baustein zukünftig nach wie vor einsetzen, sollte dieser besser geschult und die Verhaltensanker zur Beurteilung verbessert werden.

Die Messung der numerischen und verbalen Fähigkeiten der Kandidaten mittels des IST-2000-R konnten mit β = .02 und einem β = .01 keine deutliche Steigerung der prädiktiven Validität des Auswahlverfahrens, wie aufgrund des Theorieteils vermutet, erzielen. Das Gewicht des numerischen Teils liegt, wie in der Literatur beschrieben leicht über dem verbalen Teil. Ein Grund für den im Gegensatz zur Literatur erzielten niedrigen Wert der numerischen Fähigkeiten ist die im Anhang ersichtliche Korrelation dieses Testbausteins mit der Präsentationsübung (r = .17). Sieht man sich die Korrelationen nullter Ordnung an, wird deutlich, dass die Werte auf r = .16 (numerisch) und r = .05(verbal) steigen und somit zumindest ein geringer, wenn auch nicht signifikanter Zusammenhang mit der Vorgesetztenbeurteilung existiert.

Betrachtet man daneben den Gesamtdurchschnitt der beiden Testbauseine, schneiden die Kandidaten im Vergleich zu den anderen Übungen deutlich schlechter ab. Der er-

zielte Mittelwert der beiden einzelnen Testbausteine liegt auf einer ähnlichen Höhe und für beide Bausteine zusammen bei M = 1.24. Dieser Wert ist im Vergleich zu den anderen Übungsbausteinen deutlich niedriger (siehe Tabelle 18).

Tabelle 18

Deskriptive Statistiken im Multimodalen Auswahlverfahren

	Mittelwert	Standardabweichung	Häufigkeiten
Vorgesetztenbeurteilung	4,68	.69	31
Test gesamt	1,24	.69	31
Interview gesamt	1,98	.52	31
Präsentation gesamt	1,79	.85	31
Gespräch gesamt	1,76	.64	31

Folglich stellt sich die Frage, ob das Unternehmen diesen Baustein bei nachfolgend erklärten Rahmenbedingungen zukünftig einsetzten sollte. Aufgrund des hohen Wettbewerbs am Standort des Unternehmens scheint die bereits im Theorieteil erwähnte Basisrate (der Anteil an geeigneten Bewerbern) eher gering zu sein. Die Selektionsrate ist hingegen sehr hoch, was wiederum einen negativen Einfluss auf die Trefferquote hat. Sollte man diese beiden Bausteine künftig noch einsetzten wollen, sollte das Unternehmen darüber nachdenken, ihre Personalmarketingmaßnahmen deutlich zu erhöhen, um den Anteil an geeigneten Bewerbern zu steigern.

Ein weiterer Grund für den eher geringen Anteil an der prädiktiven Validität der beiden Testbausteine könnte die erfassten Dimensionen in der Vorgesetztenbeurteilung sein. Betrachtet man die erfassten Kompetenzen, so wird deutlich, dass die verbalen und numerischen Fähigkeiten, wenn überhaupt, nur einen Einfluss auf die Beurteilung der Dimensi-

onen Kommunikationsverhalten (verbal) und Kenntnisse und Fertigkeiten sowie ggfs. Arbeitsquantität und –qualität (verbal oder numerisch, je nach Tätigkeitsfeld der Kandidaten) hat. Die anderen Dimensionen der Vorgesetztenbeurteilung sind höchstwahrscheinlich nicht oder nur in geringem Maße von den numerischen und verbalen Fähigkeiten der Kandidaten abhängig.

Das Beratungs- und das Kollegengespräch tragen mit einem β = .58 und β = .52 einen hohen Beitrag zur prädiktiven Validität des Multimodalen Auswahlverfahrens bei. Betrachtet man die Werte für das Kollegengespräch im Trainee-Assessment Center mit β = .14 bzw. das Beratungsgespräch im Stabs-Assessment Center mit β = .12, wird deutlich, dass die Überarbeitung dieser beiden Gesprächssituationen einen enormen Einfluss auf die prädiktive Validität dieser beiden Bausteine hat. Eine bessere bzw. trennschärfere Operationalisierung der Verhaltensanker und die Veränderung des Beurteilungsprinzips von einer dimensionsbezogenen Übungsbeurteilung hin zu einer übungsbezogenen sowie einer Reduzierung der Skala von sechs auf vier Stufen haben sich positiv auf die Prognose des beruflichen Erfolgs ausgewirkt. Eine noch deutlichere Verbesserung wird in der Präsentationsübung deutlich. Wenn die zuvor skizzierten statistischen Ergebnisse des Trainee-Assessment Centers vor der Konzeption des Multimodalen Auswahlverfahrens bekannt gewesen wären, hätte man die Präsentationsübung aus dem Trainee-Assessment Center aller Wahrscheinlichkeit nicht ausgewählt. In dem hier vorliegenden Fall waren die Werte allerdings nicht bekannt. Es wurde eine Steigerung des β-Gewichtes von β = -.28 im Trainee-Assessment Center auf β = .33 im Multimodalen Auswahlverfahren erreicht. Diese Übung wurde auch drastisch verändert. Der Umfang des überflüssigen Materials wurde auf ein Minimum reduziert und die Zielsetzung der Übung deutlich geschärft. Die Beurteilung war durch die Veränderung des Beurteilungsprinzips und eine deutlich bessere Operationalisierung für die Beobachter besser möglich. Dies wirkte sich in dieser Übung auch auf die

Interrater-Korrelation aus, es konnte eine Steigerung von r = .64 auf r = .77 erzielt werden (siehe Anhang).

Insgesamt ist ersichtlich, dass das neue Auswahlverfahren die prädiktive Validität verbessern konnte, jedoch ist der Korrelationsunterschied des Multimodalen Auswahlverfahrens gegenüber den beiden Assessment Centern nicht signifikant. Dies liegt insbesondere an der geringen Anzahl der Vorgesetztenbeurteilungen mit N = 21, 23 bzw. 31. Auf weitere Limitationen der vorliegenden empirischen Untersuchung wird im nächsten Kapitel genauer eingegangen.

Limitationen der empirischen Studie und Implikationen für die Praxis

Wie bereits im vorangegangenen Kapitel angedeutet, gibt es Limitationen, die die Aussagekraft der vorliegenden empirischen Untersuchung einschränkt. Eine gravierende Einschränkung ergibt sich dadurch, dass die abgelehnten Bewerber nicht weiterverfolgt wurden, was zu einer bedeutenden Varianzeinschränkung der Stichprobe und somit auch zu geringeren Validitätswerten führt. Diesem Problem konnte u.a. aus Datenschutzgründen seitens des Unternehmens nicht entgegnet werden. Zukünftige Forschungen sollten allerdings auch die Weiterverfolgung der nicht eingestellten Personen mit einbeziehen.

Daneben führt die geringe Anzahl der Personen, von welchen eine Vorgesetztenbeurteilung vorliegt (N = 21-31), zu einer deutlichen Einschränkung der Prognosekraft der eingesetzten Verfahren. Die Aussagekraft von Hypothese 5 ist deshalb sehr begrenzt und schwierig auf andere Verfahren zu übertragen. An dieser Stelle wäre ferner auch eine Weiterverfolgung der Kandidaten in ihrer beruflichen Entwicklung über einen längeren Zeitraum empfehlenswert.

Zudem ist das alleinige Kriterium der Vorgesetztenbeurteilung zur Messung des beruflichen Erfolgs zwar auch in der Literatur weit verbreitet, aber als objektiver und alleiniger Maßstab fraglich, da diese stark subjektiv vom Vorgesetzten beeinflusst ist. Unabhän-

gig von diesem subjektiven Einfluss stellt sich, wie bereits angedeutet die Frage, ob der Vorgesetzte in seinem Urteil durch das Assessment Center Ergebnis im Vorfeld der Vorgesetztenbeurteilung beeinflusst wird. Somit sollten weitere Kriterien, wie der Gehaltszuwachs oder die erreichte Position nach einer bestimmten Anzahl von Jahren, hinzugezogen werden, um zur Messung des beruflichen Erfolgs ein objektiveres Kriterium zu erhalten.

Eine weitere Limitation der empirischen Untersuchung ist, dass die Zusammenhänge einzelner Anforderungsdimensionen aus dem Trainee- und Stabs-Assessment Centers aufgrund des gravierenden Unterschieds in der Definition mit den Dimensionen der Vorgesetztenbeurteilung nicht erfasst wurden. Dieser Aspekt könnte in künftigen Forschungen aufgegriffen werden, um möglicherweise die sechs bzw. sieben Dimensionen von Arthur et al. (2003) zu bestätigen.

Daneben könnten zukünftige Forschungen den bereits aufgegriffenen Zusammenhang mit den Big Five Persönlichkeitsfaktoren fokussieren, um möglicherweise die Prognose des beruflichen Erfolgs zu verbessern. Wie bereits im Theorieteil im Rahmen der prädiktiven Validität von Assessment Centern beschrieben, weisen die in Assessment Center-Übungen gemessenen Beurteilungskriterien einen Zusammenhang mit den kognitiven Fähigkeiten und der Persönlichkeit der Kandidaten auf (Dilchert & Ones, 2009; Meriac et al., 2008). Der stärkste Zusammenhang ergab sich mit dem Persönlichkeitsmerkmal *Extraversion* in Höhe von $R = .19$ (Meriac et al., 2008, S. 1047). Auch Salgado (1998b) belegt, dass es einen signifikanten Einfluss der beiden Dimensionen *Gewissenhaftigkeit* und *Emotionale Stabilität* des Fünf Faktoren Modells (FFM) auf den Berufserfolg gibt. So konnte nachgewiesen werden, dass Gewissenhaftigkeit die Varianz der allgemeinen kognitiven Fähigkeiten um 12 Prozent und die emotionale Stabilität um 7 Prozent in zivilen Berufen erhöht (Salgado, 1998b, S. 282). Ein ähnliches Ergebnis erzielte eine weitere Studie von Salgado (2003a). Es wurde eine operationale Validität von $R = .28$ für Gewissenhaftigkeit

ein Wert von $R = .16$ für emotionale Stabilität nachgewiesen (Salgado, 2003a, S. 329). Die anderen drei Dimensionen des Fünf Faktoren-Modells *Extraversion*, *Offenheit* und *Verträglichkeit* hingegen erzielten wie in bereits vorangegangenen Studien keinen signifikanten Zusammenhang mit Berufserfolg (Salgado, 2003a, S. 331).

Darüber hinaus stellt sich die Frage, wie diese beiden Dimensionen der Big Five Persönlichkeitsfaktoren zu erfassen sind. Eine Metaanalyse Judge, Rodell, Klinger, Simon und Crawford (2013) macht deutlich, dass keine eindeutige Übereinstimmung erzielt werden kann, ob man sich auf bestimmte Unterfacetten der Big Five Persönlichkeitsfaktoren fokussieren soll oder eine übergeordnete Erfassung der Persönlichkeitseigenschaft als solche ein besserer Prädiktor für beruflichen Erfolg darstellt (Judge et al., 2013, S. 92).

Auch Hülsheger und Maier (2008) empfehlen in Ihrer Metaanalyse über Prädiktoren für beruflichen Erfolg aus dem deutschsprachigen Raum, sich nicht auf den alleinigen Einsatz von Intelligenztests zu fokussieren, sondern Auswahlverfahren um Persönlichkeitstests zu erweitern (Hülsheger & Maier, 2008, S. 117).

Insgesamt hat die empirische Untersuchung deutlich gemacht, dass durch den Einsatz des dargestellten Multimodalen Auswahlverfahrens die Prognose des beruflichen Erfolgs der Kandidaten verbessert werden konnte. Somit konnte auch die Zuverlässigkeit der Einstellungsentscheidung verbessert werden.

Das Unternehmen konnte sowohl Kosten durch Fehlentscheidung minimieren als auch Kosten in der Durchführung einsparen. Die Attraktivität des Unternehmens für den Bewerber konnte erhöht werden, da die Durchführungszeit reduziert und das Feedback für den Kandidaten verbessert wurde. Folglich sollte das Unternehmen zur Einstellung neuer Mitarbeiter das Multimodale Auswahlverfahren beibehalten.

Zukünftig könnte überlegt werden, das Interview und den Test, welche einen geringen Anteil an der gesamten Validität des Verfahrens hatten, wegzulassen, um die Effizienz

des Verfahrens nochmals zu steigern. Das Unternehmen sollte auch überlegen, zur weiteren Verbesserung der Prognose des beruflichen Erfolgs Items von Persönlichkeits- oder Integritätstest in das Auswahlverfahren zu integrieren, da die beiden Facetten Gewissenhaftigkeit und emotionale Stabilität bisher nicht ausreichend berücksichtig sind, aber wie gezeigt wurde, gute Prädiktoren für beruflichen Erfolg darstellen.

Literaturverzeichnis

Achouri, C. Einführung in die psychologische Eignungsdiagnostik, 79–85.

Achouri, C. (2011). *Human Resources Management. Eine praxisbasierte Einführung* (Lehrbuch, 1. Aufl.). Wiesbaden: Gabler.

Amelang, M. & Schmidt-Atzert, L. (2006). *Psychologische Diagnostik und Intervention* (Springer-Lehrbuch, 4. Aufl.). Heidelberg: Springer.

Amelang, M. & Schmidt-Atzert, L. (2009). *Psychologische Diagnostik und Intervention* (Springer-Lehrbuch, 5. Aufl.). Heidelberg: Springer.

Arthur, W., Day, E. A., McNelly, T. L. & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology, 56* (1), 125–154.

Becker, N., Höft, S., Holzenkamp, M. & Spinath, F. M. (2011). The predictive validity of assessment centers in German-speaking regions: A meta-analysis. *Journal of Personnel Psychology, 10* (2), 61–69.

Bortz, J. (2004). *Statistik für Human- und Sozialwissenschaftler* (6. Aufl.). Heidelberg: Springer.

Bowler, M. C. & Woehr, D. J. (2006). A meta-analytic evaluation of the impact of dimension and exercise factors on assessment center ratings. *Journal of Applied Psychology, 91* (5), 1114–1124.

Bray, D. W., Campbell, R. J. & Grant, D. L. (1974). *Formative years in business: A long-term AT&T study of managerial lives.* New York: Wiley.

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

Bühner, M. & Ziegler, M. (2009). *Statistik für Psychologen und Sozialwissenschaftler* (ps Psychologie). München [u.a.]: Pearson Studium.

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56* (2), 81–105.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2.ed., 2.print). Hillsdale, NJ: Erlbaum.

De Kock, F., Born, M. & Lievens, F. (2009). *A brief review of accuracy research on assessor judgement in Assessment Centers.* (Presentation at the 29th assessment centre study group conference, Hrsg.).

Dilchert, S. & Ones, D. S. (2009). Assessment center dimensions. Individual differences correlates and meta-analytic incremental validity. *International journal of selection and assessment, 17* (3), 254–270.

Drees, H. B. (1994). *Untersuchung zur Validität eines Assessment-Centers. Hinweise zur empirischen Überprüfung eines Assessment-Centers unter besonderer Berücksichtigung unterschiedlicher Beobachtertrainings*.

Fahrmeir, L., Künstler, R., Pigeot, I. & Tutz, G. (2003). *Arbeitsbuch Statistik* (Dritte, überarbeitete und erweiterte Auflage). Berlin, Heidelberg: Springer Berlin Heidelberg.

Fecker, T. (1989). *Möglichkeiten und Grenzen einer Kontrolle der Beurteilungsqualität im Assessment-Center mit Hilfe statistischer Verfahren* (Betriebswirtschaftliche Schriftenreihe, Bd. 50). Münster: Lit.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics. And sex and drugs and rock'n'roll* (4. ed). Los Angeles, Calif.: SAGE.

Fisseni, H.-J. & Preusser, I. (2007). *Assessment-Center. Eine Einführung in Theorie und Praxis*. Göttingen [u.a.]: Hogrefe.

Fleeson, W. & Gallagher, P. (2009). The implications of Big Five standing for the distribution of trait manifestation in behavior: fifteen experience-sampling studies and a meta-analysis. *Journal of personality and social psychology, 97* (6), 1097–1114.

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C. & Bentson, C. (1987a). Journal of Applied Psychology Monograph - Meta-Analysis of Assessment Center Validity. *Journal of Applied Psychology, 72* (3), 493–511.

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C. & Bentson, C. (1987b). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72* (3), 493–511.

Gaugler, B. B. & Thornton, G. C. (1989). Number of assessment center dimensions as a determinant of assessor accuracy. *Journal of Applied Psychology, 74* (4), 611–618.

Goffin, R. D., Rothstein, M. G. & Johnston, N. G. (1996). Personality testing and the assessment center: Incremental validity for managerial selection. *Journal of Applied Psychology, 81* (6), 746–756.

Howard, A. (1974). An assessment of assessment centers. *Academy of Management journal : AMJ, 17* (1), 115–134.

Hufnagl, H. (2001). *Vom Assessment-Center zum multimodalen Auswahlverfahren. So nehmen Sie jede Hürde!* (Berufswahl und Studium). Würzburg: Lexika-Verlag.

Hülsheger, U. R. & Maier, G. W. (2008). Persönlichkeitseigenschaften, Intelligenz und Erfolg im Beruf: Eine Bestandsaufnahme internationaler und nationaler Forschung. *Psychologische Rundschau, 59* (2), 108–122.

Hülsheger, U. R., Maier, G. W. & Stumpp, T. (2007). Validity of General Mental Ability for the Prediction of Job Performance and Training Success in Germany: A meta-analysis. *International journal of selection and assessment, 15* (1), 3–18.

Hunter, J. E. & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96* (1), 72–98.

Judge, T. A., Rodell, J. B., Klinger, R. L., Simon, L. S. & Crawford, E. R. (2013). Hierarchical representations of the five-factor model of personality in predicting job performance: Integrating three organizing frameworks with two theoretical perspectives. *Journal of Applied Psychology, 98* (6), 875–925.

Kelbetz, G. & Schuler, H. (2002). Verbessert Vorerfahrung die Leistung im Assessment Center? *Zeitschrift für Personalpsychologie, 1* (1), 4–18.

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

Klimoski, R. & Brickner, M. (1987). Why do assessment centers work? The puzzle of assessment center validity. *Personnel Psychology, 40* (2), 243–260.

Kuncel, N. R. & Sackett, P. R. (2014). Resolving the assessment center construct validity problem (as we know it). *Journal of Applied Psychology, 99* (1), 38–47.

Lienert, G. A. & Raatz, U. (1998). *Testaufbau und Testanalyse* (6. Aufl., Studienausg). Weinheim: Beltz Psychologie-Verl.-Union.

Lievens, F. (2002). Trying to understand the different pieces of the construct validity puzzle of assessment centers: An examination of assessor and assessee effects. *Journal of Applied Psychology, 87* (4), 675–686.

Lorenzo, R. V. (1984). Effects of assessorship on managers' proficiency in acquiring, evaluating, and communicating information about people. *Personnel Psychology, 37* (4), 617–634.

Melchers, K. G., Henggeler, C. & Kleinmann, M. (2007). Do within-dimension ratings in assessment centers really lead to improved construct validity? *Zeitschrift für Personalpsychologie, 6* (4), 141–149.

Meriac, J. P., Hoffman, B. J., Woehr, D. J. & Fleisher, M. S. (2008). Further evidence for the validity of assessment center dimensions: A meta-analysis of the incremental criterion-related validity of dimension ratings. *Journal of Applied Psychology, 93* (5), 1042–1052.

Moses, J. L. (1973). The development of an assessment center for the early identification of supervisory potential. *Personnel Psychology, 26* (4), 569–580.

Neidig, R. D. & Neidig, P. J. (1984). Multiple assessment center exercises and job relatedness. *Journal of Applied Psychology, 69* (1), 182–186.

Obermann, C. (2009). *Assessment Center. Entwicklung, Durchführung, Trends ; mit originalen AC-Übungen* (4. Aufl.). Wiesbaden: Gabler.

Obermann, C. (2013). *Assessment Center. Entwicklung, Durchführung, Trends mit originalen AC-Übungen* (5., vollst. überarb. und erw. Aufl). Wiesbaden: Springer Gabler.

Sackett, P. R. & Dreher, G. F. (1982). Constructs and assessment center dimensions: Some troubling empirical findings. *Journal of Applied Psychology, 67* (4), 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* (1), 187–190.

Sackett, P. R. & Harris, M. M. (1988). A further examination of the constructs underlying assessment center ratings. *Journal of Business and Psychology, 3* (2), 214–229.

Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C. & De Fruyt, F. (2003). International Validity Generalization of GMA and cognitive abilities: A european community meta-analysis. *Personnel Psychology, 56* (1), 573–605.

Salgado, J. F. (2003a). Predicting job performance using FFM and non-FFM personality measures. Predicting job performance using FFM and non-FFM personality measures. *Journal of Occupational and Organizational Psychology, 76,* 323–346.

Salgado, J. F. (1998b). Big Five personality dimensions and job performance in Army and civil occupations: A European perspective. *Human Performance, 11* (2-3), 271–288.

Schippmann, J., Hughes, G. & Prien, E. (1987). The use of structured multi-domain job analysis for the construction of assessment center methods and procedures. *Journal of Business and Psychology* (1), 353–366.

Schippmann, J. S., Hughes, G. L. & Prien, E. P. (1987). The use of structured multi-domain job analysis for the construction of assessment center methods and procedures. *Journal of Business and Psychology, 1* (4), 353–366.

Schmidt, F. L. & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, *62* (5), 529–540.

Schmidt, F. L. & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology. Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124* (2), 262–274.

Schmidt, F. L., Hunter, J. E. & Pearlman, K. (1981). Task differences as moderators of aptitude test validity in selection: A red herring. *Journal of Applied Psychology, 66* (2), 166–185.

Schmidt, F. L., Hunter, J. E., Pearlman, K. & Shane, G. S. (1979). Further tests of the Schmidt-Hunter Bayesian validity generalization procedure. *Personnel Psychology, 32* (2), 257–281.

Schmitt, M. (1990). *Konsistenz als Persönlichkeitseigenschaft? Moderatorvariablen in der Persönlichkeits- und Einstellungsforschung* (Lehr- und Forschungstexte Psychologie). Berlin [u.a.]: Springer.

Schuler, H. (Hrsg.). (2007). *Assessment Center zur Potenzialanalyse* (Wirtschaftspsychologie). Göttingen [u.a.]: Hogrefe.

Shore, T. H., Shore, L. M. & Thornton, G. C. (1992). Construct validity of self- and peer evaluations of performance dimensions in an assessment center. *Journal of Applied Psychology, 77* (1), 42–54.

Strenze, T. (2007). Intelligence and socioeconomic success: A meta-analytic review of longitudinal research. *Intelligence, 35* (5), 401–426.

Thornton, G. C. & Byham, W. C. (1982). *Assessment centers and managerial performance* (Organizational and occupational psychology). New York, NY: Acad. Press.

Turnage, J. J. & Muchinsky, P. M. (1984). A comparison of the predictive validity of assessment center evaluations versus traditional measures in forecasting supervisory job performance: Interpretive implications of criterion distortion for the assessment paradigm. *Journal of Applied Psychology, 69* (4), 595–602.

Wernimont, P. & Campbell, J. (1968). Signs, Samples and Criteria. *Journal of Applied Psychology, 52* (5), 372–376.

Wirtz, M. A. (2002). *Deskriptive Statistik* (Statistische Methoden für Psychologen, 2., überarb. und erw. Aufl). Weinheim: Juventa-Verl.

Woehr, D. J. (2003). The Construct-Related Validity of Assessment Center Ratings: A Review and Meta-Analysis of the Role of Methodological Factors. *Journal of Management, 29* (2), 231–258.

Zenglein, C. (2010). *Evaluation und Überprüfung des praktischen Nutzens eines Development Centers für Führungskräfte und Experten* (Berichte aus der Psychologie). Aachen: Shaker.

Anhang

Anhang A: Korrelationskoeffizienten zur Konstruktvalidität des Trainee-Assessment Centers

Tabelle I: Korrelationskoeffizienten innerhalb der Selbstpräsentation (N=70)

Kompetenzen	Kommunikations-fähigkeit	Konflikt- und Kritikfähigkeit	Initiative und Engagement	Stabilität & Selbstvertrauen	Überzeugungs-kraft	Subjektiver Gesamt-eindruck
Kommunikationsfähigkeit	1.00	.75**	.77**	.82**	.79**	.85**
Konflikt- und Kritikfähigkeit	.75**	1.00	.72**	.68**	.74**	.81**
Initiative und Engagement	.77**	.72**	1.00	.80**	.82**	.84**
Stabilität &Selbstvertrauen	.82**	.68**	.80**	1.00	.80**	.87**
Überzeugungskraft	.79**	.74**	.82**	.80**	1.00	.91**
Subjektiver Gesamteindruck	.85**	.81**	.84**	.87**	.91**	1.00

Anmerkung. r = zweiseitiger Korrelationskoeffizient nach Pearson, ** Die Korrelation ist auf dem Niveau .01 signifikant.

Tabelle II: Korrelationskoeffizienten innerhalb der Präsentation (N=70)

Kompetenzen	Konflikt-& Kritikfähigkeit	Überzeugungs-kraft	Entscheidungs-fähigkeit	Strategisches Den-ken und Handeln	Auffassungs-gabe	Subjektiver Gesamt-eindruck
Konflikt- und Kritikfähigkeit	1.00	.79**	.73**	.66**	.65**	.80**
Überzeugungskraft	.79**	1.00	.76**	.76**	.65**	.88**
Entscheidungsfähigkeit	.73**	.76**	1.00	.76**	.72**	.84**
Strategisches Denken und Handeln	.66**	.76**	.76**	1.00	.80**	.86**
Auffassungsgabe	.65**	.65**	.72**	.80**	1.00	.81**
Subjektiver Gesamteindruck	.80**	.88**	.84**	.86**	.81**	1.00

Anmerkung. r = zweiseitiger Korrelationskoeffizient nach Pearson, ** Die Korrelation ist auf dem Niveau .01 signifikant.

Tabelle III: Korrelationskoeffizienten innerhalb des Mitarbeitergesprächs (N=70)

Kompetenzen	Kommunikati-onsfähigkeit	Konflikt- und Kritikfähigkeit	Kooperationsfä-higkeit	Strategisches Den-ken und Handeln	Subjektiver Gesamt-eindruck
Kommunikationsfähigkeit	1.00	.77**	.81**	.68**	.84**
Konflikt- und Kritikfähigkeit	.77**	1.00	.80**	.69**	.83**
Kooperationsfähigkeit	.81**	.80**	1.00	.66**	.90**
Strategisches Denken und Handeln	.68**	.69**	.66**	1.00	.83**
Subjektiver Gesamteindruck	.84**	.83**	.90**	.82**	1.00

Anmerkung. r = zweiseitiger Korrelationskoeffizient nach Pearson, ** Die Korrelation ist auf dem Niveau .01 signifikant.

Tabelle IV: Korrelationskoeffizienten innerhalb der Gruppendiskussion (N=70)

Kompetenzen	Kommunikati-onsfähigkeit	Kooperationsfä-higkeit	Initiative und Engagement	Überzeugungskraft	Subjektiver Gesamt-eindruck
Kommunikationsfähigkeit	1.00	.83**	.65**	.66**	.83**
Kooperationsfähigkeit	.82**	1.00	.64**	.70**	.80**
Kooperationsfähigkeit	.65**	.64**	1.00	.88**	.85**
Strategisches Denken und Handeln	.66**	.70**	.88**	1.00	.91**
Subjektiver Gesamteindruck	.83**	.80**	.85**	.91**	1.00

Anmerkung. r = zweiseitiger Korrelationskoeffizient nach Pearson, ** Die Korrelation ist auf dem Niveau .01 signifikant.

Tabelle V: Korrelationskoeffizienten innerhalb des Postkorbs (N=70)

Kompetenzen	Entscheidungs-fähigkeit	Strategisches Denken und Handeln	Auffassungsgabe	Subjektiver Gesamteindruck
Entscheidungsfähigkeit	1.00	.89**	.84**	.92**
Strategisches Denken und Handeln	.89**	1.00	.82**	. 92**
Auffassungsgabe	.84**	.82**	1.00	. 92**
Subjektiver Gesamteindruck	. 92**	. 92**	. 92**	1.00

Anmerkung. r = zweiseitiger Korrelationskoeffizient nach Pearson, ** Die Korrelation ist auf dem Niveau .01 signifikant.

Tabelle VI: Korrelationskoeffizienten Kommunikationsfähigkeit (N=70) (mehr als zwei Übungen)

Übungen	Gruppendiskussion	Mitarbeitergespräch	Selbstpräsentation
Gruppendiskussion	1.00	.44**	.47**
Mitarbeitergespräch	.44**	1.00	.55**
Selbstpräsentation	.47**	.55**	1.00

Anmerkung. r = zweiseitiger Korrelationskoeffizient nach Pearson, ** Die Korrelation ist auf dem Niveau .01 signifikant.

Tabelle VII: Korrelationskoeffizienten Konflikt- und Kritikfähigkeit (N=70) (mehr als zwei Übungen)

Übungen	Präsentation	Mitarbeitergespräch	Selbstpräsentation
Präsentation	1.00	.19	.36**
Mitarbeitergespräch	.19	1.00	.50**
Selbstpräsentation	.36**	.50**	1.00

Anmerkung. r = zweiseitiger Korrelationskoeffizient nach Pearson, ** Die Korrelation ist auf dem Niveau .01 signifikant.

Tabelle VII: Korrelationskoeffizienten Überzeugungskraft (N=70) (mehr als zwei Übungen)

Übungen	Gruppendiskussion	Präsentation	Selbstpräsentation
Gruppendiskussion	1.00	.55**	.62**
Präsentation	.55**	1.00	.50**
Selbstpräsentation	.62**	.50**	1.00

Anmerkung. r = zweiseitiger Korrelationskoeffizient nach Pearson, ** Die Korrelation ist auf dem Niveau .01 signifikant.

Tabelle VIII: Korrelationskoeffizienten Subjektiver Gesamteindruck (N=70) (mehr als zwei Übungen)

Übungen	Gruppendiskussion	Präsentation	Mitarbeitergespräch	Selbstpräsentation	Postkorb
Gruppendiskussion	1.00	.51**	.46**	.67**	.31**
Präsentation	.51**	1.00	.41**	.51**	.09
Mitarbeitergespräch	.46**	.41**	1.00	.54**	.23
Selbstpräsentation	.67**	.51**	.54**	1.00	.19
Postkorb	.31*	.09	.23	.19	1.00

Anmerkung. r = zweiseitiger Korrelationskoeffizient nach Pearson, ** Die Korrelation ist auf dem Niveau .01 signifikant,

* Die Korrelation ist auf dem Niveau .05 signifikant.

Anhang B: Korrelationskoeffizienten zur Konstruktvalidität des Stabs-Assessment Centers

Tabelle I: Korrelationskoeffizienten innerhalb der Selbstpräsentation (N=84)

Kompetenzen	Innovations- und Veränderungsmanagement	Auftreten und Wirkung	Subjektiver Gesamteindruck
Innovations- und Veränderungsmanagement	1.00	.69**	.75**
Auftreten und Wirkung	.69**	1.00	.83**
Subjektiver Gesamteindruck	.75**	.83**	1.00

Anmerkung. r = zweiseitiger Korrelationskoeffizient nach Pearson, ** Die Korrelation ist auf dem Niveau .01 signifikant.

Tabelle II: Korrelationskoeffizienten innerhalb der Präsentation (N=84)

Kompetenzen	Kommunikati-onsverhalten	Kritikfähigkeit	Entscheidungs-fähigkeit	Komplexitätsma-nagement	Unternehmeri-sches Denken und Handeln	Subjektiver Gesamt-eindruck
Kommunikationsverhalten	1.00	.67**	.65**	.52**	.62**	.80**
Kritikfähigkeit	.67**	1.00	.72**	.66**	.75**	.81**
Entscheidungsfähigkeit	.65**	.72**	1.00	.63**	.73**	.78**
Komplexitätsmanagement	.52**	.66**	.63**	1.00	.75**	.76**
Unternehmerisches Denken und Handeln	.62**	.75**	.73**	.75**	1.00	.80**
Subjektiver Gesamteindruck	.80**	.81**	.78**	.76**	.80**	1.00

Anmerkung. r = zweiseitiger Korrelationskoeffizient nach Pearson, ** Die Korrelation ist auf dem Niveau .01 signifikant.

Tabelle III: Korrelationskoeffizienten innerhalb der Gruppendiskussion (N=84)

Kompetenzen	Kooperationsverhalten	Initiative und Engagement	Unternehmerisches Denken und Handeln	Subjektiver Gesamteindruck
Kooperationsverhalten	1.00	.69**	.62**	.80**
Initiative und Engagement	.69**	1.00	.77**	.87**
Unternehmerisches Denken und Handeln	.62**	.77**	1.00	.82**
Subjektiver Gesamteindruck	.80**	.87**	.82**	1.00

Anmerkung. r = zweiseitiger Korrelationskoeffizient nach Pearson, ** Die Korrelation ist auf dem Niveau .01 signifikant.

Tabelle IV: Korrelationskoeffizienten innerhalb des Rollenspiels (N=84)

Kompetenzen	Kommunikati-onsverhalten	Konfliktma-nagement	Kritikfähigkeit	Ziel- und Ergeb-nisorientierung	Subjektiver Gesamt-eindruck
Kommunikationsverhalten	1.00	.85**	.85**	.82**	.80**
Konfliktmanagement	.85**	1.00	.85**	.87**	.92**
Kritikfähigkeit	.85**	.85**	1.00	.77**	.84**
Ziel- und Ergebnis-orientierung	.82**	.87**	.77**	1.00	.86**
Subjektiver Gesamteindruck	.89**	.92**	.84**	.86**	1.00

Anmerkung. r = zweiseitiger Korrelationskoeffizient nach Pearson, ** Die Korrelation ist auf dem Niveau .01 signifikant.

Tabelle V: Korrelationskoeffizienten innerhalb des Kick-off-Meetings (N=84)

Kompetenzen	Konfliktma-nagement	Kooperations-verhalten	Innovations-und Ver-änderungsverhalten	Auftreten und Wirkung	Ziel- und Ergeb-nisorientierung	Subjektiver Gesamt-eindruck
Konfliktmanagement	1.00	.86**	.76**	.81**	.71**	.86**
Kooperationsverhalten	.86**	1.00	.72**	.75**	.58**	.85**
Innovations-und Verände-rungsverhalten	.76**	.72**	1.00	.83**	.79**	.84**
Auftreten und Wirkung	.81**	.75**	.83**	1.00	.81**	.90**
Ziel- und Ergebnisorientie-rung	.71**	.58**	.79**	.81**	1.00	.80**
Subjektiver Gesamteindruck	.89**	.85**	.85**	.90**	.80**	1.00

Anmerkung. r = zweiseitiger Korrelationskoeffizient nach Pearson, ** Die Korrelation ist auf dem Niveau .01 signifikant.

Tabelle VI: Korrelationskoeffizienten innerhalb des Postkorbs (N=84)

Kompetenzen	Initiative und Engagement	Entscheidungsfähigkeit	Komplexitätsmanagement
Initiative und Engagement	1.00	.31**	.18
Entscheidungsfähigkeit	.31**	1.00	.25*
Komplexitätsmanagement	.18	.25*	1.00

Anmerkung. r = zweiseitiger Korrelationskoeffizient nach Pearson, ** Die Korrelation ist auf dem Niveau .01 signifikant,

* Die Korrelation ist auf dem Niveau .05 signifikant.

Anhang C: Interrater-Korrelationskoeffizienten in den eingesetzten Auswahlverfahren

Tabelle I: Interrater-Korrelationen im Trainee-Assessment Center (N=70)

Kompetenzen	**Selbst-präsentation**	**Präsentation**	**Gruppendiskussion**	**Mitarbeitergespräch**	**Postkorb**
r (Kommunikationsfähigkeit (R1+R2))	.84**		.55**	.64**	
r (Kooperationsfähigkeit (R1+R2))			.59**	.81**	
r (Konflikt- und Kritikfähigkeit (R1+R2))	.63**	.57**		.65**	
r (Initiative und Engagement (R1+R2))	.73**		.65**		
r (Stabilität &Selbstvertrauen (R1+R2))	.75**			.67**	
r (Überzeugungskraft (R1+R2))	.77**	.67**	.69**		
r (Auffassungsgabe (R1+R2))		.58**			.80**
r (Entscheidungsfähigkeit (R1+R2))		.67**			.85**
r (Strategisches Denken & Handeln (R1+R2))		.65**			.85**
r (Subjektiver Gesamteindruck (R1+R2))	.80**	.67**	.34**	.78**	.81**
Durchschnittlicher Korrelationskoeffizient $\bar{r}$	.76	.64	.56	.71	.83

Anmerkung. r = zweiseitiger Korrelationskoeffizient nach Pearson, R1 = Rater 1, R2 = Rater 2, ** Die Korrelation ist auf dem Niveau .01 signifikant.

Tabelle II*:* Interrater-Korrelationen im Stabs-Assessment Center (N=84)

Kompetenzen	**Selbst-präsentation**	**Präsentation**	**Gruppen-diskussion**	**Rollenspiel**	**Kick-off-Meeting**
r (Kommunikationsverhalten (R1+R2))		.74**		.83**	
r (Kooperationsverhalten (R1+R2))			.72**		.88**
r (Konfliktmanagement (R1+R2))				.84**	.88**
r (Kritikfähigkeit (R1+R2))		.64**		.75**	
r (Innovations-& Veränderungsverhalten (R1+R2))	.73**				.78**
r (Auftreten und Wirkung (R1+R2))	.80**				.79**
r (Initiative und Engagement (R1+R2))			.79**		
r (Entscheidungsfähigkeit (R1+R2))		.76**			
r (Ziel- & Ergebnisorientierung (R1+R2))				.87**	.79**
r (Komplexitätsmanagement (R1+R2))		.85**			
r (Unternehmerisches Denken & Handeln (R1+R2))		.77**	.64**		
r (Subjektiver Gesamteindruck (R1+R2))	.50**	.72**	.80**	.82**	.78**
Durchschnittlicher Korrelationskoeffizient r̄	.68	.75	.74	.82	.81

Anmerkung. r = zweiseitiger Korrelationskoeffizient nach Pearson, R1 = Rater 1, R2 = Rater 2, ** Die Korrelation ist auf dem Niveau .01 signifikant.

Tabelle III*:* Interrater-Korrelationen im Multimodalen Auswahlverfahren (N=90)

Statisti-sches Maß	Interview (R1+R2)	Interview 2 (R1+R2)	Beratungsgespräch (R1+R2)	Kollegengespräch (R1+R2)	Präsentation (R1+R2)	Durchschnittlicher Korrelationskoeffizient r̄
r	.62**	.73**	.70**	.69**	.77**	.70

Anmerkung. r = zweiseitiger Korrelationskoeffizient nach Pearson, R1 = Rater 1, R2 = Rater 2, ** Die Korrelation ist auf dem Niveau .01 signifikant.

Anhang D: Tabellen zur prädiktiven Validität der eingesetzten Auswahlverfahren

Tabelle I: Korrelationskoeffizient zwischen dem Gesamtergebnis des Trainee-Assessment Centers und der Vorgesetztenbeurteilung

Korrelationen

		AC_gesamt	GMB_gesamt
AC_gesamt	Pearson-Korrelation	1	,143
	Sig. (2-seitig)		,538
	N	70	21
GMB_gesamt	Pearson-Korrelation	,143	1
	Sig. (2-seitig)	,538	
	N	21	21

Anmerkung. N = Gesamtstichprobenanzahl.

Tabelle II-IV: Modellübersichten und Koeffizienten der Regressionsgleichung im Trainee-Assessment Center.

Modellübersicht

Modell	R	R-Quadrat	Angepasstes R-Quadrat	Standardfehler der Schätzung
1	,143[a]	,020	-,031	,73375

a. Prädiktoren: (Konstante), AC_gesamt

Modellübersicht

Modell	R	R-Quadrat	Angepasstes R-Quadrat	Standardfehler der Schätzung
1	,516[a]	,266	,021	,72399

a. Prädiktoren: (Konstante), PK_gesamt, PP_gesamt, MA_gesamt, GD_gesamt, SP_gesamt

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	t	Sig.	Korrelationen		
		B	Standardfehler	Beta			Nullte Ordnung	Partiell	Teil
1	(Konstante)	2,131	2,051		1,039	,315			
	SP_gesamt	,313	,381	,245	,822	,424	,302	,208	,189
	PP_gesamt	-,284	,262	-,279	-1,082	,296	-,124	-,269	-,249
	GD_gesamt	,469	,525	,268	,893	,386	,279	,225	,206
	MA_gesamt	,145	,263	,135	,549	,591	,164	,140	,127
	PK_gesamt	,019	,143	,032	,134	,895	-,048	,035	,031

a. Abhängige Variable: GMB_gesamt

Tabelle V: Korrelationskoeffizienten der Übungen des Trainee-Assessment Centers

Korrelationen

		GMB_gesamt	SP_gesamt	PP_gesamt	GD_gesamt	MA_gesamt	PK_gesamt
Pearson-Korrelation	GMB_gesamt	1,000	,302	-,124	,279	,164	-,048
	SP_gesamt	,302	1,000	,370	,566	,086	-,108
	PP_gesamt	-,124	,370	1,000	,342	-,195	-,017
	GD_gesamt	,279	,566	,342	1,000	-,181	-,251
	MA_gesamt	,164	,086	-,195	-,181	1,000	,069
	PK_gesamt	-,048	-,108	-,017	-,251	,069	1,000
Sig. (1-seitig)	GMB_gesamt	.	,091	,297	,110	,238	,418
	SP_gesamt	,091	.	,050	,004	,355	,321
	PP_gesamt	,297	,050	.	,065	,199	,471
	GD_gesamt	,110	,004	,065	.	,216	,136
	MA_gesamt	,238	,355	,199	,216	.	,383
	PK_gesamt	,418	,321	,471	,136	,383	.
H	GMB_gesamt	21	21	21	21	21	21
	SP_gesamt	21	21	21	21	21	21
	PP_gesamt	21	21	21	21	21	21
	GD_gesamt	21	21	21	21	21	21
	MA_gesamt	21	21	21	21	21	21
	PK_gesamt	21	21	21	21	21	21

Tabelle VI: Korrelationskoeffizient zwischen dem Gesamtergebnis des Stabs-Assessment Centers und der Vorgesetztenbeurteilung

Korrelationen

		GMB_gesamt	AC_gesamt
GMB_gesamt	Pearson-Korrelation	1	,247
	Sig. (2-seitig)		,256
	N	23	23
AC_gesamt	Pearson-Korrelation	,247	1
	Sig. (2-seitig)	,256	
	N	23	84

Anmerkung. N = Gesamtstichprobenanzahl.

Tabelle VII-IX*:* Modellübersichten und Koeffizienten der Regressionsgleichung im Stabs-Assessment Center.

Modellübersicht[b]

Modell	R	R-Quadrat	Angepasstes R-Quadrat	Standardfehler der Schätzung
1	,247[a]	,061	,016	,62313

a. Prädiktoren: (Konstante), AC_gesamt

b. Abhängige Variable: GMB_gesamt

Modellübersicht

Modell	R	R-Quadrat	Angepasstes R-Quadrat	Standardfehler der Schätzung
1	,620[a]	,384	,153	,68879

a. Prädiktoren: (Konstante), PK_gesamt, PP_gesamt, SP_gesamt, GD_gesamt, RS_gesamt, KO_gesamt

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	t	Sig.	Korrelationen		
		B	Standardfehler	Beta			Nullte Ordnung	Partiell	Teil
1	(Konstante)	4,084	1,641		2,489	,024			
	SP_gesamt	-,381	,263	-,418	-1,447	,167	-,169	-,340	-,302
	PP_gesamt	,030	,213	,033	,143	,888	,106	,036	,030
	GD_gesamt	,479	,209	,576	2,286	,036	,410	,496	,477
	RS_gesamt	,085	,163	,121	,519	,611	-,012	,129	,108
	KO_gesamt	-,144	,258	-,158	-,558	,584	,204	-,138	-,116
	PK_gesamt	,257	,212	,283	1,211	,244	,096	,290	,253

a. Abhängige Variable: GMB_gesamt

Tabelle X: Korrelationskoeffizienten der Übungen des Stabs-Assessment Centers

Korrelationen

		GMB_gesamt	SP_gesamt	PP_gesamt	GD_gesamt	RS_gesamt	KO_gesamt	PK_gesamt
Pearson-Korrelation	GMB_gesamt	1,000	-,169	,106	,410	-,012	,204	,096
	SP_gesamt	-,169	1,000	-,229	,200	,311	-,336	,181
	PP_gesamt	,106	-,229	1,000	,037	-,241	-,180	-,154
	GD_gesamt	,410	,200	,037	1,000	,017	,279	-,148
	RS_gesamt	-,012	,311	-,241	,017	1,000	,127	,053
	KO_gesamt	,204	-,336	-,180	,279	,127	1,000	,179
	PK_gesamt	,096	,181	-,154	-,148	,053	,179	1,000
Sig. (1-seitig)	GMB_gesamt	.	,221	,316	,026	,478	,176	,332
	SP_gesamt	,221	.	,147	,180	,075	,058	,205
	PP_gesamt	,316	,147	.	,433	,134	,206	,242
	GD_gesamt	,026	,180	,433	.	,469	,099	,250
	RS_gesamt	,478	,075	,134	,469	.	,282	,406
	KO_gesamt	,176	,058	,206	,099	,282	.	,207
	PK_gesamt	,332	,205	,242	,250	,406	,207	.
H	GMB_gesamt	23	23	23	23	23	23	23
	SP_gesamt	23	23	23	23	23	23	23
	PP_gesamt	23	23	23	23	23	23	23
	GD_gesamt	23	23	23	23	23	23	23
	RS_gesamt	23	23	23	23	23	23	23
	KO_gesamt	23	23	23	23	23	23	23
	PK_gesamt	23	23	23	23	23	23	23

Tabelle XI: Korrelationskoeffizient zwischen dem Gesamtergebnis des Multimodalen Auswahlverfahrens und der Vorgesetztenbeurteilung

Korrelationen

		GMB_gesamt	AC_gesamt
GMB_gesamt	Pearson-Korrelation	1	,321
	Sig. (2-seitig)		,078
	N	31	31
AC_gesamt	Pearson-Korrelation	,321	1
	Sig. (2-seitig)	,078	
	N	31	90

Anmerkung. N = Gesamtstichprobenanzahl.

Tabelle XII-XIV*:* Modellübersichten und Koeffizienten der Regressionsgleichung im Multimodalen Auswahlverfahren

Modellübersicht

Modell	R	R-Quadrat	Angepasstes R-Quadrat	Standardfehler der Schätzung
1	,321[a]	,103	,072	,66019

a. Prädiktoren: (Konstante), AC_gesamt

Modellübersicht[b]

Modell	R	R-Quadrat	Angepasstes R-Quadrat	Standardfehler der Schätzung
1	,542[a]	,294	,079	,65763

a. Prädiktoren: (Konstante), Präsentation, Interview2, NW_VERB, K_Gespräch, NW_NUM, Interview1, B_Gespräch

b. Abhängige Variable: GMB_gesamt

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	t	Sig.	Korrelationen		
		B	Standardfehler	Beta			Nullte Ordnung	Partiell	Teil
1	(Konstante)	3,715	1,923		1,932	,066			
	NW_NUM	,004	,015	,047	,244	,810	,157	,051	,043
	NW_VERB	,001	,015	,010	,054	,957	,048	,011	,009
	Interview1	-,149	,265	-,122	-,563	,579	-,059	-,117	-,099
	Interview2	-,180	,221	-,169	-,814	,424	-,233	-,167	-,143
	B_Gespräch	,416	,217	,568	1,923	,067	,205	,372	,337
	K_Gespräch	,352	,193	,524	1,825	,081	,029	,356	,320
	Präsentation	,267	,149	,333	1,796	,086	,324	,351	,315

a. Abhängige Variable: GMB_gesamt

Tabelle XV: Korrelationskoeffizienten der Bausteine des Multimodalen Auswahlverfahrens

		GMB_gesamt	K_Gespräch	B_Gespräch	NW_NUM	NW_VERB	Interview1	Interview2	Präsentation
Pearson-Korrelation	GMB_gesamt	1,000	,029	,205	,157	,048	-,059	-,233	,324
	K_Gespräch	,029	1,000	-,786	-,169	-,046	-,011	,119	-,062
	B_Gespräch	,205	-,786	1,000	,190	,032	,123	-,135	,096
	NW_NUM	,157	-,169	,190	1,000	,345	-,083	-,126	,168
	NW_VERB	,048	-,046	,032	,345	1,000	-,069	-,017	,050
	Interview1	-,059	-,011	,123	-,083	-,069	1,000	,486	,256
	Interview2	-,233	,119	-,135	-,126	-,017	,486	1,000	,047
	Präsentation	,324	-,062	,096	,168	,050	,256	,047	1,000
Sig. (1-seitig)	GMB_gesamt	.	,437	,134	,199	,398	,376	,104	,038
	K_Gespräch	,437	.	,000	,182	,402	,476	,262	,370
	B_Gespräch	,134	,000	.	,153	,432	,255	,234	,304
	NW_NUM	,199	,182	,153	.	,029	,329	,249	,184
	NW_VERB	,398	,402	,432	,029	.	,355	,465	,395
	Interview1	,376	,476	,255	,329	,355	.	,003	,082
	Interview2	,104	,262	,234	,249	,465	,003	.	,402
	Präsentation	,038	,370	,304	,184	,395	,082	,402	.
H	GMB_gesamt	31	31	31	31	31	31	31	31
	K_Gespräch	31	31	31	31	31	31	31	31
	B_Gespräch	31	31	31	31	31	31	31	31
	NW_NUM	31	31	31	31	31	31	31	31
	NW_VERB	31	31	31	31	31	31	31	31
	Interview1	31	31	31	31	31	31	31	31
	Interview2	31	31	31	31	31	31	31	31
	Präsentation	31	31	31	31	31	31	31	31