Die psychometrische Güte des Motivational Value Systems Questionnaire

Untersuchungen zu Objektivität, Reliabilität und Validität.

Inaugural-Dissertation zur Erlangung der Doktorwürde
der Philosophischen Fakultät II
(Psychologie, Pädagogik und Sportwissenschaft)
der Universität Regensburg

vorgelegt von
Josef Merk
aus Tegernsee

Die Arbeit entstand in gemeinsamer Betreuung durch die Fakultät für Psychologie, Pädagogik und Sportwissenschaft der Universität Regensburg und die Fakultät Betriebswirtschaft der Ostbayerischen Technischen Hochschule Regensburg.

Regensburg 2016

Erstgutachter: Prof. Dr. Peter Fischer
Zweitgutachter: Prof. Dr. Thomas Falter

Zusammenfassung

Diese Arbeit widmet sich der Untersuchung der psychometrische Güte des Motivational Value Systems Questionnaires (MVSQ), eines Fragebogens zur Messung von persönlichen Wertesystemen im Arbeitskontext. Wertesysteme sind psychologische Bezugsrahmen, die Menschen dabei helfen, Wahrnehmungen zu beurteilen, Handlungsalternativen zu vergleichen und dadurch Verhalten zu koordinieren. Im Arbeitskontext beeinflussen sie durch diese Funktionen, welche Aufgaben, Arten der Zusammenarbeit und Führungsstile jemand subjektiv bevorzugt und folglich auch, welche Arbeit jemand gerne erledigt und welche nicht. Die valide Messung von Wertesystemen stellt dabei eine Grundvoraussetzung sowohl für die Anwendung der Wertesystem-Theorie in der Praxis, als auch für dessen weitere Erforschung dar. Das Ziel dieser Arbeit ist die Untersuchung der Güte des MVSQ anhand der Hauptgütekriterien Objektivität, Reliabilität und Validität, um damit eben diese Grundvoraussetzung zu überprüfen. Neben den Tatsachen, dass das Instrument noch nicht auf dessen psychometrische Güte untersucht wurde und ferner keine validierten Fragebögen zur Erhebung von Wertesystemen auf Basis derselben Theorie existieren, behandelt diese Arbeit auch aus methodischer Sicht eine Forschungslücke. Denn beim MVSQ handelt es sich um einen Fragebogen im multidimensionalen Forced-Choice-Format, was einerseits praktische und konzeptuelle Vorteile gegenüber dem klassischen Rating-Format besitzt (Resistenz gegenüber Faking und konzeptuelle Passung zum Wertesystem-Konstrukt). Andererseits werden im Forced-Choice-Format ipsative Daten generiert, auf die viele klassische Methoden (z.B. klassische Testtheorie, Faktorenanalyse, etc.) nicht sinnvoll angewendet werden können. Um diese Restriktionen aufzulösen und eine plausible Bestimmung der Reliabilität und Validität vornehmen zu können, wird den Daten ein Thurstonian Item-Reponse-Theorie-Modell angepasst, dass sowohl bezogen auf den Umfang des Modells, als auch bzgl. des Schätzverfahrens neuartig ist. Außerdem beinhaltet die Arbeit eine Itemanalyse, auf dessen Basis Empfehlungen zur Überarbeitung der Items abgegeben wurden und den Vergleich der ursprünglichen Fragebogenversion mit einer überarbeiteten Version. Die Objektivität wurde logisch analytisch und die Reliabilität in Form von Test-Retest- und empirischer Reliabilität untersucht. Insbesondere die Reliabilitätskoeffizienten zeigen den Bedarf einer weiteren Überarbeitung einiger Itemformulierungen an, wozu die Kennwerte der Thurstonian IRT-Modelle als Indikatoren der Qualität der Items verwendet werden können. In den Kapiteln zur Validität wurden zudem die Konstrukt- und Kriteriumsvalidität eingehend

untersucht. Hinweise auf die Konstruktvalidität wurden dabei im Rahmen von Untersuchungen zur faktoriellen, konvergenten und divergenten Validität gefunden. Ferner haben drei Untersuchungen zur konkurrenten, prädiktiven und inkrementellen Validität gezeigt, dass die Wertesystem-Messungen kriteriumsbezogene Validität besitzen. Zum einen wurde festgestellt, dass es je nach Zugehörigkeit zu einem Studiengang oder Studiengangsschwerpunkt bzw. Aufgabenbereich oder einer Hierarchieebene charakteristische Wertesystempräferenzen gibt. Daraus können Empfehlungen zur Studiengangs- bzw. Berufswahl abgeleitet werden. Zum anderen konnten Wertesysteme in Abhängigkeit der Kongruenz mit Aufgaben in einer Untersuchung Anteile der empfundenen Motivationsintensität erklären. In einer Feldstudie haben einige Wertesysteme zudem Anteile der erbrachten Leistung vorhergesagt. Beide Befunde sind von hoher Relevanz für die Berufspraxis, denn Sie zeigen nicht nur, dass Menschen auf unterschiedliche Art motivierbar sind, sondern erklären auch wer eher auf welche Art motivierbar ist. In Summe sprechen die Ergebnisse dieser Arbeit somit sowohl für die Güte des Instruments in den drei Hauptgütekriterien, wie auch für die Gültigkeit der zugrunde liegenden Theorie der Wertesysteme.

„A map is not the territory it represents,
but, if correct, it has a similar structure to
the territory, which accounts for its usefulness.“

– Alfred Korzybski, 1933

Danksagung

An erster Stelle gilt mein Dank Prof. Dr. Thomas Falter, der für mich in vielerlei Hinsicht den Stein ins Rollen gebracht hat. Vor allem danke ich ihm für die Wegbereitung und Betreuung dieser Arbeit. Nicht minder dankbar bin ich für die vielen inspirierenden und lehrreichen Gespräche und Diskussionen. Großer Dank geht auch an Prof. Dr. Peter Fischer, für die Möglichkeit, diese kooperative Promotion überhaupt durchzuführen und ebenso für fachlichen Rat und die Betreuung dieser Arbeit.

In Dankbarkeit bin ich auch der OTH Regensburg, ganz besonders den Kollegen und Mitarbeitern an der Fakultät Betriebswirtschaft verbunden. Die Zusammenarbeit empfand ich immer als sehr freundlich und kooperativ. Ebenso danke ich den Kollegen und Mitarbeitern am Lehrstuhl, insbesondere für die unkomplizierte Hilfsbereitschaft in organisatorischen Angelegenheiten.

Außerdem danke ich Michael Pfaller für viele spannende Diskussionen über statistische und nicht-statistische Fragestellungen. Ich habe viel daraus gelernt. Vielen Dank auch für die Heranführung an R und LaTeX. Zu großem Dank fühle ich mich auch Dr. Wolff Schlotz für den immer sehr aufschlussreichen Rat in methodischen Fragen verpflichtet. Dave Zes danke ich für die Unterstützung bei der Anwendung seines R-Pakets.

Herzlich bedanken möchte ich mich des Weiteren bei Gerald Singer für die Nutzung des MVSQ und Randy Rückner vom HPCC der Universität Regensburg, für die stets schnelle und nützliche Unterstützung bei der Nutzung des HPCC. Außerdem möchte ich allen danken, die mich auf die eine oder andere Art und Weise auf diesem Weg begleitet haben, als Freunde, Korrekturleser, Inspiratoren oder Diskussionspartner. Danke Sally Oey, Corinna Käser, Irmgard Hausmann, Manuela Kronseder, Johanna Prasch, Volker Viereck, Sepp Plank, David Elsweiler, Bernd Männel, Florian Wemhoff, Felix Wemhoff, Tobias Tzschaschel, Timo Bongartz, Moritz Karpf, Mira Wimmer und Julia Pfleging. Zu guter Letzt danke ich meinen Eltern aus tiefem Herzen für ihr Vertrauen, ihre Zuversicht und Unterstützung dabei, meinen Weg zu gehen.

Inhaltsverzeichnis

Kapitel 1

Einleitung

Jeden Tag gibt es zahlreiche Situationen, in denen Menschen psychologische Bewertungen vornehmen müssen, um darauf basierend Präferenzen zu formulieren und Entscheidungen zu treffen. Was frühstücke ich? Welchen Artikel in der Zeitung lese ich? Fahre ich mit dem Auto oder dem Fahrrad ins Büro? Was sind meine Prioritäten am heutigen Arbeitstag? Welche E-Mail beantworte ich zuerst? Das sind einige Beispiele, denen gemeinsam ist, dass Alternativen verglichen und eine Priorität festgelegt werden muss, damit eine Entscheidung getroffen werden kann. Wertesysteme stellen dabei die psychologischen Bezugsrahmen dar, die Menschen helfen, ihre Wahrnehmungen zu beurteilen, bevorzugte Handlungsalternativen zu bestimmen und dadurch Verhalten zu koordinieren (Kluckhohn, 1951; Latham & Pinder, 2005; Rokeach, 1973; Schwartz & Bilsky, 1987). Sie sind auch ein zentraler Faktor der Arbeitsmotivation (Locke & Latham, 2004), denn als Bewertungskriterien beeinflussen sie, welche Aufgaben, Arten der Zusammenarbeit und Führungsstile jemand subjektiv bevorzugt. Manche Menschen lieben es, wenn ihnen ihr Vorgesetzter freie Hand lässt und sie sich eigenverantwortlich komplexen Fragestellungen widmen können. Andere bevorzugen es, klare Zielvorgaben zu haben, die zudem erlauben, sich mit anderen Kollegen im Wettbewerb zu messen. Für wieder andere steht im Vordergrund, an Entscheidungen zu partizipieren, sich im Team abzustimmen und für ein harmonisches Miteinander einzustehen. Bei allen drei Beispielen handelt es sich um Vorstellungen, die, wenn sie einträten, bei den jeweiligen Personen zu positiven Bewertungsergebnissen führen würden. Sie können auch als Beschreibungen von Wertesystemen verstanden werden. Um zu verstehen, *warum* Menschen so handeln, wie sie es tun – und folgerichtig auch, *was* sie bei der Arbeit motiviert und demotiviert – ist es unverzichtbar, ihre Wertesysteme zu kennen (Allport, 1961; Locke, 1991; Rokeach, 1973).

Ein Instrument, das entworfen wurden, um Wertesysteme im beruflichen Kontext zu messen, ist der Motivational Value Systems Questionnaire (MVSQ). Es wurde von Thomas Falter und Gerald Singer entwickelt und erfasst sieben Wertesysteme in den zwei Dimensionen der Annäherung und Vermeidung. Die Konzeptualisierungen der Wertesysteme beruhen dabei auf einer Theorie von Clare W. Graves (1966, 1970, 1974). Der Fragebogen wird zwar bereits eingesetzt,

eine Untersuchung seiner psychometrische Güte steht jedoch noch aus und stellt das Kernziel dieser Arbeit dar. Die zentrale Forschungsfrage dieser Arbeit ist demnach die Frage nach der psychometrischen Güte des MVSQ. Bei der psychometrischen Güte handelt es sich um ein vielschichtiges Konzept, dessen Überprüfung eine Vielzahl von Untersuchungen unterschiedlicher Gütekriterien erfordert. Diese können dabei in die drei Hauptgütekriterien Objektivität, Reliabilität und Validität zusammengefasst werden, die anhand weiterer Unterkriterien in dieser Arbeit analysiert werden.

1.1 Relevanz

Die Arbeit besitzt in mehrerlei Hinsicht Relevanz. Zunächst gehört es zur guten psychologischen Praxis, dass Fragebögen auf ihre Güte hin nach wissenschaftlichen Standards untersucht werden (Moosbrugger, 2012). Die drei erstgenannten Kriterien sind dabei stets Objektivität, Reliabilität und Validität. Diese sollen als Fundament der wissenschaftlichen Validierung des MVSQ untersucht werden.

Auch aus testtheoretischer Sicht sind die Ergebnisse dieser Arbeit relevant, da in ihr ein Thurstonian Item-Reponse-Theorie-Modell geschätzt wird, das speziell den Anforderungen von Forced-Choice-Fragebögen wie dem MVSQ gerecht wird. Ein Novum stellt dabei einerseits der Umfang des TIRT-Modells (speziell die Anzahl von sieben Items pro Block) und andererseits der verwendete Schätzalgorithmus dar. Zu beiden Punkten existieren bisher keine Veröffentlichungen.

Darüber hinaus ist die dem Instrument zugrunde liegende Theorie der Wertesysteme relativ unerforscht, obgleich sie Eingang in einige populärwissenschaftliche Literatur gefunden hat (z.B. Bär-Sieber et al., 2014; Beck et al. 2007; Versnel & Koppenol, 2005). Dabei fehlt jedoch eine wissenschaftliche Fundierung sowohl der Inhalte der Theorie, als auch der Messung der darin enthaltenen Wertesysteme. Der Beitrag dieser Arbeit wird sein, neue Impulse für die Erforschung der Wertesysteme und ihrer Auswirkungen auf Drittvariablen zu geben. Außerdem wird die Frage beleuchtet, ob bzw. inwiefern es sinnvoll ist, die beiden Dimensionen der Wertesysteme im MVSQ (Annäherung und Vermeidung) als orthogonale Konstrukte zu konzeptualisieren.

Des Weiteren wird in dieser Arbeit ein integrierter Satz von Begriffen der Motivationspsychologie im Zusammenhang mit Wertesystemen entwickelt. Dabei wird einerseits die theoretische Einbettung des Konstrukts Wertesystem in die motivationspsychologische Theorienlandschaft vollzogen, indem Gemeinsamkeiten, Zusammenhänge und Unterschiede zu relevanten Konstrukten und Konzepten aufgezeigt werden. Andererseits werden die dargestellten Zusammenhänge, insbesondere von Wertesystemen zu den Konzepten von Persönlichkeit und Motivation, verwendet, um daraus die Untersuchungen zur Validität konzeptuell herzuleiten.

Zu guter Letzt ist die gesellschaftliche Relevanz der in dieser Arbeit vorgenommenen Untersuchungen zu nennen. Diese liegt darin begründet, dass Wertesystemen eine hohe Bedeutung für motivationspsychologische Fragen zugeschrieben werden kann, da sie ausdrücken *was* Menschen motiviert. Das Wissen über die eigenen Wertesysteme kann Menschen dabei ein hilfreiches Mittel sein, um Frustration zu vermeiden und ein selbstbestimmtes Leben zu führen. Denn eine Person, die weiß, was sie motiviert, tut sich leichter, einen passenden und potenziell erfüllenden Beruf zu finden, in dem zudem die Wahrscheinlichkeit für eine gesteigerte Produktivität erhöht ist (Kleinbeck, 2010). Diese wiederum kann Unternehmen dabei helfen, wirtschaftlich nachhaltig erfolgreich zu sein, was einen ökonomischen Anreiz dafür darstellt, den eigenen Mitarbeitern ein werteorientiertes Arbeitsumfeld zu bieten (Fischer et al., 2013a). Mit Werteorientierung ist hier gemeint, dass bei der Personalauswahl, der Team- und Personalentwicklung, sowie bei der Mitarbeiterführung die Wertesysteme der involvierten Personen berücksichtigt werden. Auf gesellschaftlicher Ebene kann der Wert des Wissens über die eigenen Wertesysteme vor allem in dem Wissen über die Existenz unterschiedlicher Wertesysteme gesehen werden. Denn wer sich bewusst ist, das Menschen unterschiedliche Wertesysteme bevorzugen können, dem fällt es auch leichter, den Wert eines jeden Wertesystems zu erkennen und schließlich auch anzuerkennen. Letzteres führt zu mehr Toleranz und Akzeptanz innerhalb der Gesellschaft, was in Zeiten erhöhter Mobilität und sich verändernder Bevölkerungsstrukturen von hoher Bedeutung ist. Um jedoch all diese Punkte zu erreichen, bedarf es des Wissens über die Wertesysteme. Eine der einfachsten Methoden, sich seiner Wertesysteme bewusst zu werden, ist sie in einem validierten Fragebogen zu messen. Somit kann der Bogen der diagnostischen Fragestellungen dieser Arbeit hin zur inhaltlichen Aufhängung in der humanistischen Psychologie gespannt werden.

1.2 Aufbau der Arbeit

Diese Arbeit besteht insgesamt aus 11 Kapiteln. Davon behandeln die ersten beiden Kapitel theoretische und methodische Grundlagen. In Kapitel 2 wird dabei die Theorie der Wertesysteme eingeführt und eine Begriffsbestimmung vorgenommen. Zudem wird das Konstrukt Wertesystem in einer theoretischen Auseinandersetzung von relevanten Konstrukten abgegrenzt und ein Literaturüberblick zur Werteforschung gegeben. Im Anschluss werden in Kapitel 3 die messtheoretischen und methodischen Grundlagen der Arbeit gelegt. Danach folgt als einführende Untersuchung in Kapitel 4 die Begutachtung der Objektivität anhand der drei Unterarten Durchführungs-, Auswertungs- und Interpretationsobjektivität.

Als erstes die Messgenauigkeit betreffende Kapitel, widmet sich Kapitel 5 der deskriptivstatistische Evaluation der Items des MVSQ. Eine solche Analyse ist stets relativ am Anfang der Fragebogenentwicklung erforderlich (Jonkisz et al., 2012). Das Folgekapitel widmet sich der Spezifizierung und Schätzung eines geeigneten Testmodells, was die Basis für die in Kapitel

7 durchgeführte Untersuchung der empirischen und der Test-Retest-Reliabilität darstellt. Als abschließende Untersuchung zur Reliabilität des Instruments werden in Kapitel 8 die Fragebogenversionen miteinander verglichen. In dieser wird zudem der Bezug zur Itemanalyse in Kapitel 5 hergestellt. Diese Analyse erfolgt *nach* der Reliabilitätsberechnung, da ebenso die Reliabilitäten der beiden Fragebogenversionen miteinander verglichen werden, was erst nach der Anpassung eines geeigneten Testmodells möglich ist.

Die Kapitel 9 und 10 behandeln in mehreren Untersuchungen die Validität des Instruments. In Kapitel 9 wird dabei anhand der faktoriellen, der konvergenten und der divergenten Validität die Konstruktvalidität des MVSQ untersucht. Kapitel 10 widmet sich der Kriteriumsvalidität mit Untersuchungen zur konkurrenten, prädiktiven und inkrementellen Validität.

Abschließend wird die Arbeit durch eine Diskussion der Ergebnisse kritisch bewertet. Die Bedeutung der einzelnen Analysen im Hinblick auf die Güte des Instruments wird herausgestellt, sowohl die Grenzen als auch der theoretische und praktische Nutzen der Arbeit werden aufgezeigt, aus der Arbeit resultierende Forschungsfragen zusammengefasst und dadurch ein Ausblick auf Forschungs- und Anwendungsmöglichkeiten gegeben.

Kapitel 2

Inhaltstheoretischer Hintergrund

Dieses Kapitel setzt sich mit den inhaltstheoretischen Hintergründen des MVSQ auseinander. Dazu werden zunächst die der Arbeit zentralen Konstrukte *Wert* und *Wertesystem* definiert und ein entsprechender Überblick des Forschungsstands, inklusive kurzem geschichtlichen Abriss, gegeben. Des Weiteren wird der Bezug zu anderen in der Motivationspsychologie relevanten Konstrukten, wie Bedürfnissen und Zielen, hergestellt, sowie die Zusammenhänge mit Persönlichkeit und Motivation gezeigt. Der übergeordnete Zweck dieses Kapitels liegt in der integrativen Weiterentwicklung der motivationspsychologischen Theorie, sowie der Klarstellung, was im MVSQ gemessen werden kann.

2.1 Das Wertekonstrukt

Das Wertekonstrukt existiert bereits seit den Anfängen psychologischer Forschung. Als eine erste Wertetheorie kann eine Typologisierung von sechs Lebensformen durch Eduard Spranger (1921) gesehen werden. Darin werden sechs Lebensformen durch jeweils unterschiedliche Werte beschrieben, die von Spranger als z.B. „der theoretische Mensch“ (S. 109), „der ökonomische Mensch“ (S. 130) oder „der soziale Mensch“ (S. 171) bezeichnet wurden. Auf dieser Theorie aufbauend entwickelten Vernon und Allport (1931) ein erstes Instrument zur Messung von Werten, das später als *Allport-Vernon-Lindzey Study of Values* (SOV) bekannt wurde (Allport et al., 1960). Es erhob die dominanten Werte von Menschen anhand der Klassifizierung von Spranger. Ebenfalls Mitte des 20. Jahrhunderts setzten sich immer mehr und auch bekannte Psychologen mit dem Wertekonstrukt auseinander. Zum Beispiel beschrieben Kurt Lewin (1952), Fritz Heider (1958) und Gordon Allport (1961) einstimmig Werte als eine bedeutende Determinante für menschliches Verhalten. Aus dieser Zeit stammt auch die sehr häufig zitierte Definition von Kluckhohn (1951, S. 395), die wie folgt lautet:

> *„A value is a conception, explicit or implicit, distinctive of an individual or characteristic of a group, of the desirable that influences the selection from available modes, means, and ends of actions.“*

In den 70er Jahren erschien das Buch *The Nature of Human Values* von Milton Rokeach (1973), das als *das* wichtigste Grundlagenbuch zum Thema Werte gesehen werden kann. Darin werden einerseits grundlegende und bis heute häufig zitierte Definitionen beschrieben, sowie der Rokeach Value Survey vorgestellt, der Grundlage für den heute international sehr häufig eingesetzten Werte-Fragebogen, den Schwartz Value Survey ist (Schwartz, 1992).

In der jüngeren, die letzten 30 Jahre umfassenden Werteforschung, ist vor allem Shalom Schwartz als bedeutender Werteforscher zu nennen. Er hat in zahlreichen Artikel die Entwicklung der Wertetheorie vorangetrieben (Schwartz, 1992, 1994; Schwartz & Bilsky, 1987, 1990), sowie seine *Theorie der universellen menschlichen Werte* in unzähligen empirischen Studien mit dem Schwartz Value Survey untersucht (für einen Überblick vgl. Schwartz, 2012). Dabei sind insbesondere die Hinweise auf die interkulturelle Gültigkeit der Theorie hervorzuheben (Fischer & Schwartz, 2011; Schwartz, 1994; Schwartz & Rubel, 2005).

Nach Schwartz und Bilsky (1987) stellen Werte die „kognitiven Repräsentationen“ (S. 550) von (a) biologischen Bedürfnissen, (b) Anforderungen für die zwischenmenschlichen Koordination und (c) Kriterien des Gemeinwohls dar. Sie fungieren dabei als Bewertungsmaßstäbe dessen, was jeweils subjektiv gut und wünschenswert ist (Kluckhohn, 1951; Rokeach, 1973). Vereinfacht gesagt – und darüber herrscht im Großen und Ganzen Konsens zwischen Werteforschern – drücken Werte das aus, was für Menschen subjektiv *wertvoll* und allgemein wünschenswert ist. In ihrer Funktion als psychologische Bewertungskriterien leiten sie die Auswahl von Verhalten (Hitlin & Piliavin, 2004; Rohan, 2000). Dabei sind sie anwendbar auf jegliche Stimuli, wie Objekte, Situationen und Ereignisse (Locke, 1969; Rohan, 2000) oder andere Personen und das Selbst (Hitlin, 2003; Lord & Brown, 2001; Schwartz, 1992). Da Werte stets positiv konnotiert sind, treffen Menschen bevorzugt Entscheidungen und führen Handlungen aus, die mit ihren Werten kongruent sind (Bardi & Schwartz, 2003; Rokeach, 1973; Verplanken & Holland, 2002). Auch interagieren sie lieber mit anderen Menschen, die die eigenen Werte teilen (Schwartz & Bilsky, 1987).

Des Weiteren ist zu sagen, dass Menschen mehrere Werte präferieren können und diese in einer hierarchischen Struktur angeordnet sind (Rokeach, 1973). Dies ist wichtig, da das Wertekonstrukt nur insofern sinnvoll als Bewertungsstandard verstanden werden kann, wenn es ermöglicht, unterschiedliche Valenzen von Alternativen zu bestimmen (Locke, 1991).

Neben den bereits aufgeführten Inhaltstheorien von Werten ist auch noch das häufig zitierte Modell der Kulturdimensionen von Hofstede (1980, 1984) zu nennen. In diesem werden vier Typen von Werten unterschieden, die von Hofstede (1984) als „Standards für gut und schlecht“ (S. 389) definiert werden. Wenig Beachtung hingegen erfuhr das Wertemodell von Graves (1966, 1970, 1974). Das mag zum einen daran liegen, dass Graves wenig Bezug zu der damals

vorherrschenden Forschung nahm[1] und zum anderen hat Graves nur eine Veröffentlichung bei einer psychologischen Zeitschrift.

Der Begriff *Wertesystem* kommt weit seltener vor als der Ausdruck Wert. Nach Rokeach (1973) ist ein Wertesystem eine fortdauernde Organisation von Werten, die Menschen hilft, zwischen Alternativen zu wählen und Entscheidungen zu treffen. Jeder Mensch besitzt dabei *ein* individuelles System von Werten. Schwartz (1996) geht in seiner theoretischen Auseinandersetzung tiefer und beleuchtet die Begriffe Werte-Typ, Wertepräferenz und Wertesystem. Im Unterschied zu Rokeach hält Schwartz die Arbeit mit Einzelwerten für wenig praktikabel, da Einzelwerte zum einen wenig reliabel seien und es zweites unzählige Einzelwerte gäbe. Die Frage danach, welche aller möglichen Einzelwerte in einer Untersuchung verwendet werden, kann somit nur schwer ex ante beantwortet werden, was wiederum wenig wissenschaftlich ist. Schwartz bündelt deshalb in seiner Theorie der grundlegenden Werte mehrere ähnliche Einzelwerte in sogenannten Werte-Typen, wobei er den Begriff zwar bevorzugt, aber austauschbar mit Wertesystem verwendet. Wertesysteme bzw. Werte-Typen können nach Schwartz (1996) als „integrated wholes“ (S. 126) gesehen werden, die als Ganzes mit anderen Variablen interagieren können und auch als Ganzes gemessen werden müssen. Laut Schwartz gibt es zehn Wertesysteme, die universell gültig sind und die jeder Mensch in unterschiedlichem Ausmaß präferieren kann. Wertepräferenz beschreibt dabei nach Schwartz (1996) einfach die Bevorzugung eines Wertesystems über ein anderes.

2.2 Die Theorie von Clare W. Graves

In den 60er und 70er Jahren des 20. Jahrhunderts entwickelte Clare W. Graves (1966, 1969, 1970, 1971a,b, 1974, 2005)[2] eine Theorie von Wertesystemen. Die Theorie postuliert acht Wertesysteme, die hierarchisch angeordnet sind und unterschiedliche motivationale Systeme beschreiben (Graves, 1970). Beispielsweise kann ein solches System durch das Streben nach persönlichem Erfolg im Wettbewerb mit Anderen gekennzeichnet sein, oder aber durch den Wunsch nach Gemeinschaft in Harmonie und Konsens (Graves, 1970, 1974). Hinsichtlich des Konstrukts *Wertesystem* hat Graves selbst vergleichsweise wenig definitorische Arbeit geleistet. Er war vielmehr auf die inhaltliche Erforschung seiner Konzeptualisierung der Wertesysteme fokussiert. Eine der wenigen Beschreibungen des Konstrukts stammt aus dem Artikel von 1966, indem er Wertesysteme als psychologische Gleichgewichtszustände beschreibt, die bestimmen, wie eine Person denkt, fühlt, beurteilt, handelt und motiviert ist.

Graves konzeptualisiert sie des Weiteren als dynamische kognitive Strukturen, die nach eigenen Regeln und Prinzipien funktionieren und bestimmen, was eine Person als wünschens-

[1] In seinem Artikel von 1970 referenzierte er z.B. nur eine Quelle von Maslow und einen seiner früheren Artikel.

[2] Das Buch von 2005 erschien weit nach Graves Tod. Es basiert zu großen Teilen auf Originalunterlagen und wurde von C. Cowan und N. Todorovic editiert.

wert, positiv und richtig empfindet (Graves, 1970). Zwar definiert er Werte nicht direkt als Ausdruck des Wünschenswerten, wie z.B. Kluckhohn, Rokeach oder Schwartz das getan haben, doch in den Beschreibungen der Wertesysteme tauchen immer wieder die Begriffe *Wunsch* und *wünschen* auf. Zudem ist auch bei Graves die Funktion von Wertesystemen das Leiten von Entscheidungen und Handlungen. Es ist deshalb legitim zu sagen, dass die anfangs dargestellten Definitionen von Wert und Wertesystem auch für die Theorie von Graves gelten.

Ausgangspunkt der Entwicklung der Theorie war folgende Frage (Graves, 1971c, S. 11):

> *„What will be the nature and character of conceptions of psychological health of biologically mature human beings who are intelligent but relatively unsophisticated in psychological knowledge in general, and theory of personality, in particular?"*

Im Laufe seiner Forschung ließ er zahlreiche[3] Versuchspersonen (VP) Konzepte dessen erarbeiten, was diese unter „psychologisch gesund" verstehen (Graves, 1971c). Aus diesen Konzepten bildete Graves als Kern seiner Theorie ein Modell von acht Wertesystemen. Bei seinen Testpersonen handelte es sich um erwachsene Personen im Alter zwischen 18 und 61 Jahren, die relativ wenig Wissen über Psychologie bzw. Persönlichkeitspsychologie besaßen (Graves, 1971c). Seine Vorgehensweise war dabei so, dass er die VP zuerst die Konzepte und danach ein Klassifizierungssystem der Konzepte erarbeiten ließ, aus denen er letztlich sechs der acht Wertesysteme herleitete (Graves, 1971c). Diese von ihm auch als „Existenzebenen" oder „Persönlichkeitssysteme" bezeichneten Kategorien beschreiben jeweils ein kohärentes System von Werten. Als Bezeichnung für die Wertesysteme verwendete Graves Kombinationen von Buchstaben, wobei die Buchstaben A bis H die psychologischen Systeme und die Buchstaben N bis U die Umweltsysteme repräsentierten. Diese Methode verdeutlicht, dass Graves Wertesysteme als untrennbar von der Umwelt konzeptualisierte. Er hypothetisierte, dass sie sich stets in Wechselwirkungen zwischen Person und Umwelt entwickeln (Graves, 1970).

Nachdem er diese sechs Wertesysteme (C-P bis H-U) tiefgehend untersucht hatte, fügte er geleitet von Beobachtungen weiterer Personen seines Umfeldes, die nicht an seiner Forschung teilnahmen, ein weiteres Wertesystem zum Modell hinzu, das er als B-O bezeichnete (Graves, 1971c). Das achte Wertesystem (A-N) konstruierte er auf Basis logischer Überlegungen und im Gegensatz zu den übrigen Wertesystemen nicht auf empirischer Basis. Es beinhaltet die „Werte" *Überleben* und *Fortpflanzung* und ist im Gegensatz zu den übrigen Wertesystemen nicht bewusstseinsfähig. Unter definitorischer Stringenz handelt es sich bei diesen Begriffen auch nicht um Werte, sondern um Grundbedürfnisse. Die insgesamt sieben dem Bewusstsein zugänglichen Wertesysteme unterteilte Graves des Weiteren in zwei Kategorien, die er als *express-self* und *sacrifice-self* bezeichnete (Graves, 1971c). Bei Wertesystemen, die der ersten Gruppe angehören (C-P, E-R und G-T) geht es darum, das eigene Selbst auszudrücken. Bei der

[3] Anzahl unbekannt, vermutlich jedoch mehrere Hundert, da er die Untersuchung über viele Jahre hinweg in jedem Semester wiederholte.

zweiten Gruppe von Wertesystemen (B-O, D-Q, F-S und H-U) steht das Streben im Mittelpunkt, sich seinem sozialen Umfeld unterzuordnen.

Darüber hinaus sei erwähnt, dass Graves das Modell der Wertesysteme als offene Hierarchie von Systemen konzipierte (Graves, 1970). „Offen" bedeutet dabei zweierlei. Zum einen drückt es auf das gesamte Modell bezogen aus, dass es offen für die Entstehung neuer Wertesysteme ist. Auf der Ebene der Wertesysteme bedeutet offen hingegen, dass Menschen *offen* dafür sein können, ihre bevorzugten Wertesysteme zu verändern (Graves, 1971c). Im Laufe des Lebens können sich demnach die Präferenzen innerhalb der eigenen Wertesystemhierarchie verändern.

Wie schon beschrieben, ist das Modell in der psychologischen Forschung wenig bekannt, hat jedoch in jüngerer Vergangenheit anscheinend größeren Anklang in der Management-Praxis gefunden, da in den letzten gut zehn Jahren mehrere populärwissenschaftliche Bücher veröffentlicht wurden, die das Wertemodell in Zusammenhang mit Personal- und Organisationsentwicklung (Bär-Sieber et al., 2014; Beck et al. 2007; Keijser & Vat, 2009; Krumm, 2012; Versnel & Koppenol, 2003, 2005), Recruiting (Köbler, 2009) und Mediation (Ponschab et al., 2009) setzen. Teilweise werden dort auch Erhebungsverfahren erwähnt, allerdings keine Angaben zu deren psychometrischer Güte gemacht und auch in einschlägigen Verzeichnissen (ZPID, PSYCNDEX, Web of Science / Social Sciences Citation Index und Google Scholar) lassen sich dazu keine wissenschaftlichen Artikel finden.

Falter und Singer haben das Modell aufgegriffen und mit den ihrer Ansicht nach zentralen Werten jedes Wertesystems zusammengefasst, um es leichter erklär- und erinnerbar zu machen (T. Falter, persönliche Kommunikation, 06.03.2013). Tabelle 1 zeigt die Wertesysteme mit den jeweils bezeichnenden Werten pro Wertesystem.[4] Im Folgenden werden nun die Wertesysteme dargestellt und anstatt der von Graves verwendeten Buchstabenkombination, werden die Wertesysteme mit den von den Fragebogenentwicklern festgelegten Hauptwerten bezeichnet und durch eine entsprechende Buchstabenkombinationen abgekürzt.

Die Wertesysteme

Jedes Wertesystem lässt sich durch mehrere Werte beschreiben und repräsentiert einen wünschenswerten Zustand oder Lebensstil. Dadurch fungiert es als Maßstab dafür, was als wünschenswert gesehen wird und was nicht. Die folgenden Formulierungen leiten sich aus mehreren Originalartikeln ab (Graves, 1966, 1969, 1970, 1971c, 1974, 2005) und verwenden die Bezeichnungen des MVSQ.

Im weiteren Verlauf der Arbeit werden die Wertesysteme zum besseren Verständnis stets fett und kursiv geschrieben. Wenn also von ***Erfolg*** die Rede ist, dann ist damit das Wertesystem „Erfolg" gemeint, das zudem weitere Werte enthält. Es sei zudem darauf hingewiesen, dass es irreführend sein kann, Wertesysteme mit nur einem Wort zu beschreiben, da jedes Wertesystem

[4] Aus dem MVSQ-Ergebnisbericht entnommen.

Tabelle 1. Beschreibung der Wertesysteme.

Denominator	weitere Werte
Geborgenheit (GB)	Verbundenheit, Beständigkeit, Tradition
Macht (MA)	Entscheidungsfreude, Durchsetzungswille, Konfliktfreude
Gewissheit (GW)	Regeltreue, Struktur, Disziplin
Erfolg (ER)	Ergebnisorientierung, Status, Wettbewerb
Gleichheit (GL)	Harmonie, Netzwerke, Konsens
Verstehen (VE)	Freiheit, Wissen, Innovation
Nachhaltigkeit (NA)	Gesellschaftl. Verantwortung, Gesellschaftl. Relevanz, Ganzheitlichkeit

aus vielen, kohärenten Werten besteht. Der MVSQ ähnelt dabei der Umsetzung von Schwartz, der in der Theorie der grundlegenden Werte ebenso Wertesysteme beschreibt und diese auch mit *einem* Wort betitelt.

Geborgenheit (GB)

Bezeichnende Werte des Wertesystems ***Geborgenheit*** sind Verbundenheit, Beständigkeit und Tradition. Menschen, die dieses Wertesystem hoch ausgeprägt haben, weisen eine starke Orientierung hin zur engen Gruppe auf, die gleichbleibend und einander sehr loyal sein muss. Die Gruppe steht über dem Individuum, Bräuche, Rituale und Traditionen stellen ein bestimmendes Merkmal des täglichen Lebens dar.

Macht (MA)

Personen, die hohe Ausprägungen auf dem ***Macht***-Wertesystem haben, streben danach, Entscheidungen zu treffen und die Welt nach ihren Vorstellungen zu gestalten. Sie empfinden es als wünschenswert, Stärke zu zeigen, sich durchzusetzen und Autorität zu haben. Macht und der eigene Wille stehen im Zentrum der Aufmerksamkeit und leiten die täglichen Entscheidungen.

Gewissheit (GW)

Menschen mit einer starken Präferenz des ***Gewissheit***-Wertesystems streben nach Ordnung und Vorhersehbarkeit. Sie schätzen klare Vorgaben, Regeleinhaltung und Ausführung nach Vorschrift. Eine Organisation funktioniert dann, wenn jeder seinen Platz in der Hierarchie hat und die entsprechende Rolle pflichtbewusst und diszipliniert ausfüllt.

Erfolg (ER)

Personen, die das Wertesystem ***Erfolg*** präferieren, schätzen Wettbewerb, unternehmerisches Denken, Gewinnmaximierung und die Anerkennung für die erbrachte Leistung. Pragmatismus und Ergebnisorientierung spielen eine wichtige Rolle, um dem persönlichen Erfolg als übergeordnetem Ziel näher zu kommen. Status und Prestige sind dabei häufig geeignete Ausdrucksformen des persönlichen Erfolgs.

Gleichheit (GL)

Für Menschen mit einer hohen Ausprägung auf dem ***Gleichheit***-Wertesystem steht der Mensch im Vordergrund. Angestrebte Werte sind Harmonie und Konsens mit den Mitmenschen, der Austausch von Gefühlen und gegenseitiges Kümmern. Soziale Bindung spielt eine wichtige Rolle, weswegen Netzwerke einen hohen Stellenwert einnehmen.

Verstehen (VE)

Personen mit einem hohen Anteil von ***Verstehen*** zeichnen sich dadurch aus, dass sie verstehen wollen. Dazu gehört, stets kritisch zu hinterfragen und theoretische Modelle über Zusammenhänge und Wirkungsweisen zu formulieren. Sie schätzen konzeptuelle Arbeit, das Analysieren von Zusammenhängen und den Aufbau von Wissen als Selbstzweck.

Nachhaltigkeit (NA)

Für Menschen mit hohen Ausprägungen auf ***Nachhaltigkeit*** steht gesellschaftliche Verantwortung und das im Fokus, was in ihren Augen wirklich für die Welt von Bedeutung ist. Es ist wünschenswert, globale ökologische Probleme zu lösen und die Welt zu einem besseren Ort zu machen. Dazu sind eine weltumfassende Sicht und die Antizipation langfristiger Trends vonnöten, um innerhalb des eigenen räumlichen und zeitlichen Kontextes die Weichen für eine nachhaltige Entwicklung der Welt als Ganzes zu stellen.

2.3 Werte-Prozesstheorien

Neben den genannten Inhaltstheorien existieren auch eine Reihe von Prozesstheorien, die das Wertekonstrukt behandeln. Allen voran ist hier die Gruppe der Erwartungs-Wert-Theorien zu nennen (Beckmann & Heckhausen, 2010) und als deren wichtigste Vertreter die Valenz-Instrumentalitäts-Erwartungs-Theorie (Vroom, 1964), das Risikowahl-Modell (Atkinson, 1957) sowie die Prospect Theory (Kahneman & Tversky, 1979; Tversky & Kahneman, 1981). Diese Theorien haben gemeinsam, dass sie Valenz und Erwartung miteinander verknüpfen und dadurch erklären, wie groß die Motivation ist, eine Handlung zu tätigen. Das Prinzip ist

dabei das der Nutzenmaximierung und lautet übertragen auf den Zusammenhang: Je größer die Valenz, also der Wert, der einer Handlung und deren Folgen beigemessen wird, und je positiver die Erwartung, umso größer ist auch die Motivation, eine entsprechende Handlung auszuführen. Diese Theorien stellen auf der Prozessseite dar, wie Werte bzw. Wertbeimessungen Entscheidungen und Handlungen steuern. Nimmt man nun die Inhaltsperspektive hinzu, dann bestimmt die Übereinstimmung der Handlung und der erwarteten Handlungsfolgen mit den eigenen Wertepräferenzen den Wert der Handlung. Je größer die Kongruenz, desto größer auch die Valenz und umso attraktiver ist die Entscheidung bzw. Handlung.

Eine weitere Prozesstheorie, in der Valenzen eine Rolle spielen, ist die Equity-Theorie von Adams (1963). Sie besagt, dass Menschen nach Fairness und gerechter Beurteilung ihrer Leistung streben. Der Wert, der dabei der Leistung beigemessen wird, hängt von den Wertepräferenzen der beurteilenden Person ab. Haben nun Vorgesetzte oder Kollegen andere Wertepräferenzen, messen sie der Leistung unter Umständen einen niedrigeren Wert bei als die handelnde Person, was diese dann als ungerecht empfinden kann. Häufig wird die Leistung dabei monetär ausgedrückt (Lawler, 1968).

2.4 Abgrenzung zu verwandten Konstrukten

Nachdem das Konstrukt Wertesystem nun definiert und die der Arbeit zugrunde liegende Wertetheorie vorgestellt wurde, werden sie in diesem Abschnitt zu ähnlichen Konstrukten in Bezug gesetzt und davon abgegrenzt. Dabei werden zudem aktuell relevante Theorien erwähnt, sodass dieser Abschnitt zugleich einen Überblick des aktuellen Stands der Theorie darstellt.

2.4.1 Bedürfnisse

Das Bedürfnis-Konstrukt zählt zu den ältesten und bekanntesten psychologischen Konstrukten. Bedürfnisse gelten als grundlegende Triebfedern menschlichen Verhaltens und stellen damit äußerst wichtige Faktoren der Motivation dar. Sie können dabei als unterbewusste Erscheinungen von „Ist-Sollwert-Diskrepanzen“ (Kuhl, 2010, S. 339) verstanden werden und es gibt eine Vielzahl an Theorien bezüglich möglicher Kategorisierungen (z.B. Deci & Ryan, 1985; Maslow, 1943; McClelland, 1987; Murray, 1938). Entscheidend in Bezug auf das Wertekonstrukt ist, dass Bedürfnisse im Prinzip für alle Menschen gleichermaßen gelten. Jeder Mensch hat dieselben Bedürfnisse (z.B. Hunger, Durst, Anschluss und Autonomie), die er ultimativ zum Überleben braucht (Ryan et al., 1996) und interindividuelle Unterschiede liegen nicht auf der Inhaltsebene, sondern nur auf der Prozessebene vor (Locke, 1991). Anhand von Bedürfnissen kann man Menschen also nur daran unterscheiden, wie oft bzw. wann jemand Hunger hat und in welcher Form er dieses Bedürfnis nach Essen stillt, nicht aber daran, *dass* jemand Hunger hat. Werte hingegen können sich sehr wohl interindividuell unterscheiden und Menschen

dementsprechend auch daran unterschieden werden (Locke & Henne, 1986). Damit einher geht auch, dass Bedürfnisse als angeboren und Werte als durch Kognition und Erfahrung erlernt gelten (Latham & Pinder, 2005; Locke, 1991; Rokeach, 1973).

Bedürfnisse werden als eher tief liegende psychologische Konstrukte verstanden, die relativ weit entfernt von der Handlung sind. Der Zusammenhang zwischen Bedürfnis und Handlung kann als sehr komplex angesehen werden, da es eine Vielzahl möglicher kognitiver Prozesse zwischen dem Auftreten eines Bedürfnisses und einer Handlung geben kann (Locke, 1991). Dazu gehören z.B. das Identifizieren des Bedürfnisses, die Bewertung der Bedeutung (Attribution), die Formulierung und Beurteilung von Handlungsalternativen oder das Treffen von Entscheidungen. Zudem können Wissen, Werte, Ziele, Interessen und Erwartungen als Einflussfaktoren auftauchen. Außerdem ist nicht jede Handlung eine Reaktion auf die Mangelerscheinung, sondern viele Entscheidungen sind vorhersehender Natur, werden also getätigt, bevor ein Bedürfnis auftritt und um dieses zu vermeiden. Im Vergleich dazu sind also Werte „näher“ an der Handlung.

Damit im Einklang steht auch, dass Werte leichter ins Bewusstsein gerufen werden können als Bedürfnisse und damit auch leichter verbalisiert werden können (Latham & Pinder, 2005; McClelland, 1985). Sie gehören demnach eher den expliziten Motiven an, Bedürfnisse dagegen der impliziten Motivklasse (Heckhausen & Heckhausen, 2010) und sind auch durch Introspektion zugänglich (Locke, 1991).

Nach Rokeach (1973) können Werte als kognitive Repräsentationen von Bedürfnissen verstanden werden. Das bedeutet, das Werte dafür verwendet werden können, um zu erklären, wie Menschen dasselbe Bedürfnis auf unterschiedliche Art befriedigen. Auch Locke (1991) versteht Werte als Werkzeuge der Bedürfnisbefriedigung, die durch ihre Funktion der Auswahl von Handlungen benötigt werden, damit Handlungen getätigt werden, die im Endeffekt der Bedürfnisbefriedigung dienen. Wenn man Bedürfnisse als implizit und Werte als explizit ansieht, dann stimmen auch McClelland et al. (1989) damit überein, dass Werte Bedürfnisse kanalisieren können (Kehr, 2004b). Allerdings muss es nicht so sein, dass Werte immer direkt der Bedürfnisbefriedigung dienen, weil Bedürfnisse erstens eben implizit und dadurch unbekannt sein können und zweitens menschliches Verhalten oft auch antizipativ ist, also dem Bedürfnisauftreten vorbeugt (Locke, 1991).

In Summe kann festgehalten werden, dass Bedürfnisse wenig geeignet sind, um direkte Verhaltensursachen zwischen Menschen zu differenzieren (Locke, 1997), sondern viel mehr dafür taugen, allgemeingültig zu erfüllende Defizite zu erklären. Hierin liegen auch die Hauptunterschiede zu Werten, denn Werte haben erstens keinen Defizit-Charakter, sondern sind motivationale Dispositionen, die konstant gelten (Rokeach, 1973; Schwartz, 1992). Zweitens können sich Werte zwischen Personen unterscheiden und demnach individuelle Unterschiede im Verhalten erklären.

2.4.2 Interessen

Die Interessenforschung kann in zwei Perspektiven aufgeteilt werden: Interesse als situationsbedingtes Emotionserleben und Interesse als dispositionale Persönlichkeitseigenschaft (Low et al., 2005; Silvia, 2006). Als Ersteres ist Interesse stark kontextabhängig und beschreibt die momentan erlebte Bereitschaft zur Aufmerksamkeit (Krapp, 1999). In diesem Sinne ist Interesse als psychologischer Zustand zu verstehen. Aus der zweiten Sicht handelt es sich bei Interessen um auf konkrete Objekte oder Domänen ausgerichtete „evaluative Orientierungen" (Eccles & Wigfield, 2002, S. 114). Zum Beispiel interessieren sich manche Menschen für Literatur, andere für Statistik und wieder andere für Natur. Je nach persönlichem Interesse würde dann zum Beispiel die Bewertung der Schulfächer Deutsch, Mathematik oder Biologie ausfallen. Interessen können also zur Evaluation von Stimuli herangezogen werden, wodurch sie zu motivationalen Dispositionen werden (Larson et al., 2002). In diesem dispositionalen Sinne können sie leicht mit Werten verglichen werden und sind diesen relativ ähnlich. Im Unterschied zu Werten sind Interessen jedoch weniger breit, sondern stattdessen spezifischer anwendbar (Roe & Ester, 1999; Rokeach, 1973). Sie beziehen sich auf konkrete Objekte, Konzepte oder Aktivitäten, sind weniger zur Bewertung von Personen oder Situationen im allgemeinen geeignet und zahlenmäßig kann eine Person wesentlich mehr Interessen als Werte haben (Rokeach, 1973). Werte sind demnach grundlegender und tiefer in der Psyche verankert als Interessen. Als Folge ist ein kausaler Einfluss von Werten auf Interessen plausibler als die Annahme, dass Interessen Werte beeinflussen. Empirische Befunde zu dieser Fragestellung liegen jedoch nicht vor und könnten ein interessantes Forschungsgebiet darstellen, denn denkbar ist auch ein wechselseitiger Einfluss. Zum Abschluss sei gesagt, dass es zum Interesse-Konstrukt auf inhaltlicher Ebene die Theorie der sogenannten Big Six Interessen gibt, die häufig in Studien im beruflichen Umfeld und zur Kategorisierung von Berufen eingesetzt werden (Larson et al., 2002; Mount et al., 2005). Die Big Six beinhalten als Kategorien z.B. *künstlerisch*, *sozial* und *unternehmerisch* (Holland, 1997), die auch als Persönlichkeitstypen verstanden werden können (Larson et al., 2002). Je nachdem, wie spezifisch oder generisch nun eine Klassifizierung von Interessen formuliert wird, um so ferner oder näher ist das Interessen-Konstrukt dem Werte-Konstrukt. Eine Person mit einer hohen Ausprägung der sozialen Interessen hat vermutlich auch hohe Ausprägungen von sozialen Werten, also dem Wertesystem ***Gleichheit*** und die Ausprägung der Klasse der unternehmerischen Interessen korreliert vermutlich hoch mit dem Wertesystem ***Erfolg***. Bei diesen Beispielen ist die Ähnlichkeit zwischen Interesse und Wertesystem hoch. Bei spezifischen Interessen wie Interesse an Mathematik oder Literatur ist die Ähnlichkeit jedoch zu keinem Wertesystem offensichtlich und der Zusammenhang zwischen Interesse und Wertesystem nicht eindeutig.

2.4.3 Ziele

Auch bei Zielen können theoretisch die Prozess- und die Inhaltsperspektive eingenommen werden, allerdings handelt es sich bei den verbreiteten Theorien zum Zielkonstrukt um Prozesstheorien. Es gibt umfassende empirische Forschungsergebnisse zu Zielen und insbesondere den damit einhergehenden Prozessen der Zielsetzung und Zielerreichung (Gollwitzer & Sheeran, 2006; Locke & Latham, 2002, vgl.). Dabei ist zu sagen, dass beim Zielsetzungsprozess (Auswahl von Zielen) inhaltliche Aspekte eine wesentlich größere Rolle spielen als bei der Zielerreichung, diese aber nicht in Zieltheorien behandelt werden. Dafür wird in der Regel auf andere Konstrukte wie Bedürfnisse, Werte, Interessen, etc. zurückgegriffen. Diese stellen die Haupteinflussfaktoren darauf dar, welches Ziel oder welche Ziele sich jemand setzt (Bagozzi et al., 2003; Gollwitzer et al., 2011; Perugini & Bagozzi, 2001). Man kann sogar soweit gehen, zu sagen, dass Ziele die konkreten Ausformulierungen von Motiven (Bedürfnissen wie Werte) sind (Kehr, 2004a; Locke, 1968) bzw. der „Mechanismus“ (Latham & Pinder, 2005, S. 491) sind, der zwischen Werten und Handlung wirkt.

Definiert werden kann ein Ziel als die interne Repräsentation eines erwünschten, zukünftigen Zustands bzw. Erreichens eines Handlungsergebnisses (Kleinbeck, 2010; Locke, 1997; Ryan et al., 1996). Allein der Aspekt des Wünschenswerten rückt das Zielkonstrukt sehr nahe an das Wertekonstrukt heran und macht deutlich, dass Ziele eng mit Werten verwandt sind. Im Unterschied zu Werten sind Ziele jedoch näher an der Handlung (Locke & Henne, 1986) und ebenso wie Interessen weniger abstrakt, sondern spezifischer auf einzelne Zusammenhänge anwendbar (Locke, 1991). Anders herum können Werte auch als abstrakte, „transsituationale Ziele“ (Schwartz, 1994, S. 21) oder „supergoals“(Rokeach, 1973, S. 14) verstanden werden. Um diesen Zusammenhang zu verdeutlichen, sei an dieser Stelle ein Beispiel angeführt. Eine Person kann zum Beispiel das Ziel haben, bei einem Sportwettkampf zu gewinnen. Dabei handelt es sich eindeutig um ein Ziel, denn der einzelne Sportwettkampf ist ein konkretes Ereignis und das Ziel ist ein erwünschter und zukünftiger Zustand. Wertesysteme liegen hier „dahinter“ und steuern die Wertigkeit des Ziels, wobei das Prinzip der Kongruenz greift (Feather, 1995; Locke, 1991). Es erscheint deshalb wahrscheinlich, dass diese Person mit dem Ziel des Gewinnens eine hohe Ausprägung auf dem Wertesystem ***Erfolg*** hat, da dieses Wertesystem bei der Auswahl des Ziels selbigem einen hohen Wert verliehen hat. Für eine andere Person, die z.B. das Wertesystem ***Verstehen*** hoch ausgeprägt hat, läge ein möglicher Wert im Ziel des Gewinnens darin, dass die Person durch die Teilnahme am Wettbewerb verstehen kann, wie das Gewinnen funktioniert (wie viel Training ist erforderlich, welche ist die richtige Technik, wie fühlt es sich an, Erster zu sein, einen Pokal zu erhalten?). Das Ziel des Gewinnens ist dabei dasselbe wie bei der ersten Person, die Grundlage für die Auswahl des Ziels jedoch eine andere. Dieses Beispiel soll verdeutlichen, wie der Zusammenhang zwischen Wertesystemen und Zielen konzeptualisiert werden kann, jedoch nicht darüber hinwegtäuschen, dass Menschen mit den Zielen des Gewinnens von Wettbewerben höchstwahrscheinlich häufiger das Wertesystem ***Erfolg*** statt

Verstehen hoch ausgeprägt haben, da für zweitere das Ziel nur dann einen psychologischen Wert hat, wenn es noch neu und unerforscht ist. Um den Bogen zu spannen, sei gesagt, dass das „supergoal“ für Personen mit ***Erfolg***-Orientierung das Gewinnen allgemein ist, gleich in welchem Wettbewerb oder welcher Situation. Für Menschen mit ***Verstehen***-Orientierung ist das transsituationale Ziel das Verstehen von Zusammenhängen und Funktionsprinzipien. Bezogen auf den Wettbewerb ist das Gewinnen als konkretes Ziel eben nur dann attraktiv, wenn es kongruent zum Wertesystem ist.

Des Weiteren ist das Konstrukt des Vorsatzes (Implementation Intention) anzuführen, das unter der Rubrik Ziele eingeordnet werden kann. Denn es handelt sich dabei um eine konkrete Ausformulierung eines Ziels, die beschreibt, wie bzw. wann ein gesetztes Ziel erreicht werden kann (Gollwitzer, 1999; Gollwitzer & Sheeran, 2006). Bezogen auf das vorherige Beispiel könnte ein solcher Vorsatz lauten: „Ab sofort trainiere ich fünf mal die Woche“ oder „Wenn ich Feierabend habe, gehe ich ins Training“. Auch Vorsätze sind demnach dem Wertekonstrukt insofern untergeordnet, dass sie weniger abstrakt und häufig zeitlich weniger stabil sind. Im Vergleich zum Zielkonstrukt sind Vorsätze der Zielerreichung unterzuordnen, wobei dennoch auch Einflüsse der Wertepräferenzen auf die Auswahl der Vorsätze plausibel sind.

Abschließend kann gesagt werden, dass das Zielkonstrukt ein breit erforschtes Konstrukt ist, das mit vielen weiteren Bereichen, wie z.B. Motivation (Locke, 1996), Selbstregulierung (Latham & Locke, 1991), Selbstwirksamkeit (Bandura & Cervone, 1983; Locke et al., 1984; Zimmerman et al., 1992) und Leistung (Locke & Latham, 1990) in Bezug gesetzt wurde und dass sich Zieltheorien vor allem prozessualen Fragen widmen. Für die inhaltliche Klassifizierung von Zielen werden andere Konstruktklassen, wie Wertesysteme, Bedürfnisse und Interessen, herangezogen.

2.4.4 Einstellungen

Einstellungen haben drei zentrale Eigenschaften (Fischer et al., 2013b). Sie sind innere Zustände, die von variierender zeitlicher Dauer sein können, haben evaluativen Charakter und beziehen sich auf ein Objekt. Ihre zentrale Funktion besteht darin, die Bewertung gegenüber einem Einstellungsobjekt darzustellen bzw. auszudrücken (Eagly & Chaiken, 1993). In dieser Funktion sind auch Einstellungen dem Wertekonstrukt sehr ähnlich, denn auch Werte leiten Bewertungen und steuern damit Denken und Handeln. Der Hauptunterschied zwischen diesen beiden Konstruktarten ist, dass Einstellungen nicht zeitlich überdauernd sein müssen und folglich weniger bzw. nicht dispositional sind (Graumann & Willig, 1983). Auch unterschiedlich ist, dass sich Einstellungen sowohl auf konkrete wie auch auf abstrakte Objekte beziehen können. Werte im Vergleich gelten als eher abstrakt (Roe & Ester, 1999). Man könnte sich Wertesysteme somit auch als diejenigen Einstellungen vorstellen, die tief verankert, d.h. zeitlich überdauernd und eher abstrakterer Natur sind (Hitlin, 2003). Anders herum kann man Einstellungen auch als

die Ergebnisse einer Bewertung konzeptualisieren, die auf Basis von Wertesystemen gemacht wurden (Rohan, 2000).

Im Unterschied zu Zielen, die ebenfalls der Evaluation dienen und als nicht dispositional gelten, können Einstellungen auch auf abstrakte Objekte, wie Ideen oder Konzepte gerichtet sein (Fischer et al., 2013b). Eine weit verbreitete Theorie zum Einstellungskonstrukt ist die Theorie des geplanten Verhaltens, die Einstellungen vor allem bezogen auf das Verhalten berücksichtigt (Ajzen, 1991; Fishbein & Ajzen, 2010). Sie beschreibt, welche Rolle Einstellungen bei der Planung und Ausführung von Verhalten spielen und wie sie mit anderen Verhaltensdeterminanten, wie der wahrgenommenen Kontrolle und sozialen Normen zusammenwirken. Die Theorie deckt also weitere Bereiche von Motivation ab und kann als eine Art Meta-Theorie gesehen werden.

2.4.5 Motive

Motive werden traditionell als „handlungsleitende Wirkgrößen“ (Kleinbeck, 2010, S. 300) beschrieben, die „individuelle Präferenzen für bestimmte Anreizklassen“(Brandstätter et al., 2013, S. 5) zum Ausdruck bringen. Sie sind also verhaltensbestimmende Größen, die in der Person liegen. Nach dieser Definition sind die bisher genannten Konstrukte gleichzeitig auch Motive, denn sie liegen alle in der Person und können Verhalten determinieren. Motive werden zudem danach unterteilt, ob sie dem Bewusstsein zugänglich, also explizit oder nicht direkt zugänglich, also implizit sind (Brunstein, 2010a). Eine der bekanntesten Inhaltstheorien zum Motivkonstrukt ist McClelland’s (1987) Theorie der Motivation, in der die drei Hauptmotive Leistung, Macht und Anschluss sowohl als explizite wie auch als implizite Motive konzeptualisiert werden.

Wertesysteme (wie auch Interessen, Einstellungen und Ziele) können relativ eindeutig den expliziten Motiven zugeordnet werden, da sie dem Bewusstsein in der Regel relativ einfach zugänglich sind (Pinder, 2008), Bedürfnisse können beides sein. Sogenannte „auto-motives“ (Bargh, 1990; Bargh & Ferguson, 2000) wären ein Beispiel für vollständig implizite Motive bzw. die prä-bewusste Wirkung von expliziten Motiven. Das Motivkonstrukt kann als Sammelbecken oder Überbegriff für unterschiedliche handlungsleitende Determinanten verstanden werden.

2.5 Persönlichkeit

Beim Begriff Persönlichkeit denken viele Forscher (und Laien gleichermaßen) zuerst an stabile Persönlichkeitsdimensionen, sogenannte *Traits* oder Dispositionen (Ajzen, 2005). Die Definition betreffend stimmen Forscher allgemein überein, dass Persönlichkeit ein Satz von Charakteristika einer Person ist, der auf einzigartige Weise deren Denken, Fühlen und Handeln beeinflusst (Pervin et al., 2005; Ryckman, 2008). Es ist ein psychologisches Konstrukt, in dem im Grunde alle personeninternen Eigenschaften subsumiert werden können, gleich ob diese genetischen Ursprungs sind oder im Laufe der Ontogenese angeeignet wurden (Ryckman,

2008). Persönlichkeitsforscher widmen sich nun häufig entweder der Kategorisierung dieser Persönlichkeitseigenschaften und/oder den Unterschieden, zu denen diese Charakteristika im Verhalten führen.

Ungeachtet der Frage danach, wie viele und welche Persönlichkeitseigenschaften es gibt, ist als quasi Meta-Theorie der Persönlichkeitseigenschaften die Eigenschaftstheorie von Cattell (1946) zu nennen. Sie gilt als weithin akzeptierte und empirisch fundierte Klassifizierung von Persönlichkeitseigenschaften in drei Kategorien (Scheffer & Heckhausen, 2010): Kognitive Dispositionen (Fähigkeiten), Temperamentsdispositionen und motivationale Dispositionen. Unter kognitiven Dispositionen versteht Cattell Fähigkeiten oder die „Art der Reaktion auf die Komplexität einer Situation" (Cattell, 1973, S. 31) bei klaren Zielvorstellungen. Temperamentseigenschaften beschreiben stilistische Merkmale des Verhaltens (typische Verhaltensstile) und sagen etwas über das *wie* von Verhalten aus, ungeachtet der motivationalen Lage. Letztere wird durch die motivationalen Wesenszüge bestimmt und beschreibt, *was* eine Person anstrebt (Scheffer & Heckhausen, 2010; Winter et al., 1998).

Die oben genannten Konstrukte (Wertesysteme, Bedürfnisse, Interessen, Ziele, Einstellungen und Motive) haben gemein, dass sie der motivationalen Kategorie zugeordnet werden können, da sie aus unterschiedlichen Beschaffenheiten heraus Einfluss auf Motivation haben und dazu verwendet werden, die Frage nach den Gründen von Verhalten zu beantworten. Dies gilt ungeachtet dessen, ob sie eher Zustände beschreiben, die nur kurze Zeit bestehen (z.B. eine Einstellung, die sich durch eine Diskussion mit einem Kollegen in kurzer Zeit verändern kann), oder langfristig stabile Traits (wie z.B. Wertesysteme) sind, wobei streng genommen nur zeitlich stabile Konstrukte als Dispositionen bezeichnet werden sollten. Deshalb sind Werte (neben Bedürfnissen) diejenigen der genannten Konstrukte, die eindeutig in die Klasse der motivationalen Dispositionen eingeordnet werden können (Graumann & Willig, 1983). Hier sei angemerkt, dass Werte von manchen Forschern auch als „außerhalb" der Persönlichkeit gesehen werden (z.B. Latham & Pinder, 2005; Locke & Latham, 2004). Dies ist so zu erklären, dass das Persönlichkeitskonstrukt auch enger gesehen werden kann, d.h. nur diejenigen Aspekte abdeckt, die hier als Temperamentsdispositionen beschrieben wurden. Eben diese Dimension wird häufig synonym mit dem Begriff Persönlichkeit verwendet und durch die Big Five (Digman, 1990; Goldberg, 1990, 1993) beschrieben. Dabei handelt es sich um eine der am weitesten verbreiteten, anerkannten und replizierten Theorien der Persönlichkeit (Barrick & Mount, 1991; Roccas et al., 2002; Viswesvaran & Ones, 2000). Sie kann zwar, aufgrund der lexikalischen Herangehensweise, nicht eindeutig einer der drei Dispositionsklassen zugeordnet werden (Eysenck, 1991; Scheffer & Heckhausen, 2010), beschreibt im Großen und Ganzen aber eher typische Verhaltensmerkmale, also Temperamentsdispositionen (Scheffer & Heckhausen, 2010). Als Vertreter der Klasse der kognitiven Dispositionen können z.B. das Intelligenz-Konzept (Cattell, 1987; Maltby et al., 2011) oder das Fähigkeits-Modell von Bartram (2005) und Kurz und Bartram (2002) aufgeführt werden.

Eng verwandt mit dem Konstrukt Persönlichkeit ist das Selbstkonzept (Herzberg & Roth, 2014). Dieses „besteht aus universellem und idealtypischem Wissen über die eigene Person“ (Asendorpf, 2015, S. 109). Es speist sich also aus Erfahrungen und generiert sich aus den Wissensbeständen und Schemata, die man charakteristisch für die eigene Person hält (Asendorpf, 2015; Markus, 1977). Zusammen mit der sozialen Identität (Haslam, 2009; Tajfel, 1982) bildet das Selbstkonzept die Identität und kann somit auch als persönliche Identität bezeichnet werden (Hitlin, 2003). Werte sind ein Teil dessen, was diese persönliche Identität ausmacht und konstituieren dadurch das Selbst (Feather, 1995; Hitlin, 2003; Rokeach, 1973).

2.6 Motivation

Motivation als theoretisches Konzept kann in drei Dimensionen unterteilt werten: Richtung, Intensität und Ausdauer. Darüber sind sich viele Motivationsforscher einig (z.B. Brandstätter et al., 2013; Heckhausen & Heckhausen, 2010; Locke, 1991; Pinder, 2008; Rothermund & Eder, 2011). Anhand dieser Einteilung wird im Folgenden erläutert, auf welche Aspekte der Motivation Wertesysteme einen Einfluss haben.

Die Richtung der Motivation drückt aus, *was* jemanden motiviert bzw. was jemand mit seinem Verhalten erreichen möchte. Sie ist also stark inhaltsgebunden, weshalb logischerweise Inhaltstheorien in Frage kommen, um diese Dimension der Motivation zu beschreiben. Dafür wurden bereits einige Theorien genannt, wie die Bedürfnistheorie von Maslow (1943), die Selbstbestimmungstheorie von Deci und Ryan (1985), und Theorien der Macht- (Schmalt & Heckhausen, 2010), Leistungs- (Brunstein, 2010b) und Anschlussmotivation (Sokolowski & Heckhausen, 2010). Auch Wertetheorien betreffen diese Dimension der Motivation.

Auch auf der Prozessseite gibt es einige Theorien, die sich dieser Dimension der Motivation widmen, wobei viele dieser Theorien auch die anderen beiden Facetten beinhalten, da diese teilweise nur schwer voneinander zu trennen sind. Genannt seien hier exemplarisch die Theorie des geplanten Verhaltens (Ajzen, 2005), Zielsetzungstheorien (Kleinbeck, 2010; Locke & Latham, 2002) und das Rubikon-Modell der Handlungsphasen (Achtziger & Gollwitzer, 2010). Auch die bereits genannten Erwartungs-Wert-Theorien sind Motivationstheorien, die sich mit der Richtung der Motivation auseinander setzen.

Die Intensität von Motivation betrifft nun das Erleben von Verhalten. Es geht darum, wie intensiv eine Handlung erlebt wird und wie viel Anstrengung und Konzentration sie erfordert (Heckhausen & Heckhausen, 2010). Hierbei handelt es sich eher um die Prozessebene des Verhaltens. Theorien, die sich damit auseinandersetzen sind z.B. die Flow-Theorie (Csikszentmihalyi, 1990, 2002; Rheinberg, 2010), die Theorie der Selbstbestimmung mit ihrem Ausführungen zur intrinsischen Motivation (Deci & Ryan, 1985), aber auch Forschung zum Erleben von Emotionen, wie positiver und negativer Affekt (Watson & Tellegen, 1985; Watson et al., 1988) oder das Circumplex-Modell der Emotion (Russell, 2003; Russell et al., 1989).

Der dritte Punkt, die Ausdauer, ist wohl der theoretisch einfachste Aspekt, denn im einfachen Sinne betrifft er auch lediglich die zeitliche Dauer, die eine Person aufwendet, um eine Handlung durchzuführen (Brandstätter et al., 2013). Allerdings kann hierunter auch fallen, wie häufig jemand bereit ist, eine Handlung erneut, auch bei Misserfolg, zu tätigen, oder wie groß das Durchhaltevermögen bei Tätigkeiten ist, die einem nicht entsprechen. Es geht hier also auch um Fragen der Volition (Achtziger & Gollwitzer, 2010; Gollwitzer, 1996; Locke & Kristof, 1996), der Selbstwirksamkeitserwartung (Bandura, 1977, 1982) oder Selbstregulation allgemein (Bandura, 1991; Kuhl, 2010), die einen Einfluss auf die Ausdauer der Motivation haben. Auch Erwartungen und zugeschriebene Valenzen (siehe Erwartungs-Wert-Theorien) haben sicherlich Auswirkungen auf die Persistenz einer Handlung. Bei den genannten Theorien handelt es sich wieder eher um Prozesstheorien.

Wie wirken die drei Dimensionen der Motivation zusammen? Dazu gibt es viele Theorien und Ansätze, die mehr oder weniger umfassend sind. Eine simple und rein logische Überlegung dazu ist, dass Menschen dann eine höhere Intensität empfinden und bereit sind, eine größere Ausdauer an den Tag zu legen, wenn die Richtung der Motivation bei einer Handlung bzw. in einer gegebenen Situation verwirklicht werden kann. Folgt man dieser Logik, dann haben Werte, die ja primär Ausdruck dessen sind, *was* bzw. wohin jemand möchte, auch einen Einfluss auf die empfundene Intensität bei der Handlung und die Bereitschaft, eine Handlung lange durchzuführen oder wiederholt anzugehen. Dieser Zusammenhang ist stark vereinfacht, denn es gibt noch zahlreiche weitere Einflussgrößen, die bisher nicht genannt wurden. Zum Beispiel steuert Attribution, wie etwas wahrgenommen (Stiensmeier-Pelster & Heckhausen, 2010), Gedächtnis, wie es gespeichert (Gruber, 2011) und kognitive Dissonanz, wie es verarbeitet wird (Elliot & Devine, 1994; Festinger, 1962). Jedoch bringt die Ausführung zum Ausdruck, welche Bedeutung Werten bei der Erklärung von Verhalten zukommen kann.

2.6.1 Annäherungs- und Vermeidungsmotivation

Ein interessantes Teilgebiet der Motivationsforschung stellt die Aufteilung in Annäherungs- und Vermeidungsmotivation dar. Die Kernaussage dieses Feldes ist, dass es zwei unabhängig voneinander existierende motivationale Systeme gibt (Carver, 2006; Elliot, 2008). Die Annäherungsmotivation stellt dabei diejenige Tendenz dar, sich *hin zu*, die Vermeidungsmotivation die Tendenz, sich *weg von* etwas zu bewegen. Diese gehen jeweils einher mit dem Wunsch nach Vermehrung von Freude bzw. Verringerung von Schmerz (Higgins, 1997) und erinnern an das Hedonismus-Prinzip.

Bezogen auf das Wertekonstrukt ist anzumerken, dass Werte und Wertesysteme, so, wie sie in der Literatur beschrieben werden und auch in dieser Arbeit bisher konzeptualisiert wurden, die Seite der Annäherungsmotivation betreffen. Das heißt, dass es sich dabei um stets positiv konnotierte Konstrukte handelt, die das ausdrücken, was jemand haben will und nicht, was

jemand vermeiden will. So stellen zum Beispiel für Menschen mit hohen Ausprägungen auf dem Wertesystem ***Gleichheit*** Harmonie und Konsens wertvolle Zustände dar, die als Folge im sozialen Umfeld angestrebt werden bzw. denen sich diese Personen annähern.

Doch auch bei Werten existiert die Auffassung, dass diese negative Valenz haben und folglich etwas Unerwünschtes repräsentieren können (Graf et al., 2011; Quaquebeke et al., 2010). Zwar gab es diese Ideen vereinzelt auch schon früher – Lewin (1952) schreibt z.B., dass Werte dazu dienen können, die negative Valenz einer Sache zu bestimmen und auch Schwartz (1994) geht davon aus, dass es so etwas wie negative Werte gibt. Doch eine wirkliche Einführung eines Konstrukt von negativen Werten wurde erst von Quaquebeke et al. (2010) vorgenommen. Auch die gängigen Werte-Fragebögen von Schwartz, Hofstede oder Rokeach erheben die Vermeidungsdimension von Werten bzw. Wertesystemen nicht als eigenständige Dimension. Ob positive Wertesysteme auch gleichzeitig zu Vermeidungsverhalten führen, also z.B. eine ***Gleichheit***-orientierte Person automatisch z.B. ***Macht*** vermeidet, kann an dieser Stelle nicht beantwortet werden, stellt jedoch eine interessante Forschungsfrage dar. Diese soll im Laufe der Arbeit als Aspekt der Konstruktvalidität behandelt werden.

2.6.2 Arbeitsmotivation

Arbeitsmotivation betrifft die Bereitschaft, die eigenen Ressourcen und Fähigkeiten dafür einzusetzen, eine mit den Organisationszielen übereinstimmende Arbeitsleistung zu erbringen (Sturm et al., 2011). Sie betrifft somit denjenigen Teil der allgemeinen Motivationsforschung, der in Zusammenhang mit Arbeit steht oder sich darauf beziehen lässt (Pinder, 2008). Interessanterweise führt diese vermeintliche Einschränkung motivationspsychologischer Fragen auf den Bereich Arbeit dazu, dass sich neue Forschungsfragen ergeben, die im privaten oder studentischen Umfeld eine untergeordnete Rolle spielen. Dazu zählen z.B. Fragen nach den Ursachen und Abläufen von Arbeitsproduktivität und Leistung, dem organisationalem Commitment (Bindungsverhalten) sowie der Arbeitszufriedenheit (Kleinbeck, 2010; Sturm et al., 2011). Als weitere Aspekte können Mitarbeiterführung, Job Involvement (Arbeitsengagement), Stress und Gruppenprozesse, sowie die Passung zwischen Mitarbeiter und Organisation (Person-Organisation fit) bzw. Beruf (Person-Job fit) genannt werden (Haslam, 2009; Latham, 2012; Locke & Latham, 2004; Pinder, 2008; Sturm et al., 2011). Auf all diese Bereiche können stets die Zusammenhänge mit der Richtung, Intensität und Persistenz der Motivation untersucht werden.

Besonders hervorzuheben ist an dieser Stelle der Einfluss von Wertesystemen auf Arbeitsmotivation in Abhängigkeit der Passung von Wertesystemen und Umgebung. Dazu sei zunächst gesagt, dass Werte und damit auch Wertesysteme auch im Arbeitskontext das Wünschenswerte darstellen (Dose, 1997; Roe & Ester, 1999). Meistens werden sogenannte Arbeits-Werte (*work values*) dabei als spezifischer Ausdruck von allgemeineren „Lebens-Werten“ verstanden (Roe &

Ester, 1999, S. 2). Der Einfluss von Wertesystemen auf Unterschiede im vokationalen Erleben und Verhalten lässt sich nun besonders gut durch das Konzept der Kongruenz bzw. Passung zwischen Wertesystemen und Arbeitskontext erklären. Kleinbeck (2010) beschreibt dies als „Wechselwirkung zwischen persönlichen Motiven und Motivierungspotenzialen“ (S. 44), wobei mit Motivierungspotenzialen Attribute gemeint sind, die der Arbeit inhärent sind und Wertesysteme, wie zuvor dargelegt, als persönliche Motive verstanden werden können. Wenn also bei einer Arbeitsaufgabe die „Möglichkeit zur Befriedigung persönlicher Motive“ (Kleinbeck, 2010, S. 41) besteht, dann besitzt sie das Potenzial, die Bereitschaft und damit auch die Intensität und Persistenz der Motivation zu erhöhen. Koppelt man nun an dieser Stelle die Forschung zu Werten im Arbeitskontext an, dann kann festgestellt werden, dass eine Arbeitsaufgabe dann Motivierungspotenzial besitzt, wenn die Arbeit kongruent zu den eigenen Wertesystemen ist (Kristof-Brown et al., 2005; Meglino et al., 1989). Zu dieser Thematik gibt es eine Reihe von Untersuchungen, in denen der Einfluss von Werten – im Falle von Kongruenz zur Umgebung – auf verschiedene Konstrukte, wie z.B. Arbeitszufriedenheit (Boxx et al., 1991; Meglino et al., 1989), Entscheidungen (Feather, 2002; Feather, 1995; Verplanken & Holland, 2002), Commitment (Abbott et al., 2005; Boxx et al., 1991; O’Reilly et al., 1991), Führungsverhalten (Carter & Greer, 2013; Ping et al., 2010), Managementerfolg (England & Lee, 1974), Personalauswahl (Adkins et al., 1994) und Berufswahl (Judge & Bretz, 1992) untersucht wurde.

Zusammenfassend kann gesagt werden, dass der Zusammenhang zwischen Wertesystemen und vokationalem Empfinden und Verhalten sowohl aus der theoretischen, wie auch der empirischen Perspektive als gesichert gesehen werden kann. Der Zusammenhang gestaltet sich dabei allerdings weniger so, dass Wertesysteme als direkte Einflussfaktoren gelten, sondern mehr die indirekte Ursache – je nach Passung mit dem Kontext – für Verhalten und Empfindungen darstellen (Roe & Ester, 1999).

2.7 Zusammenfassung

In diesem Kapitel wurden die dieser Arbeit zentralen Konstrukte Wert und Wertesystem definiert und das dem MVSQ zugrunde liegende Modell der Wertesysteme vorgestellt. Außerdem wurden verwandte motivationale Konstrukte wie Bedürfnisse, Ziele und Einstellungen präsentiert und die Gemeinsamkeiten und Unterschiede zum Wertekonstrukt herausgestellt. Zur Abrundung wurden Konzepte von Persönlichkeit und Motivation herangezogen und das Wertekonstrukt darin integriert.

Wertesysteme haben Berührungspunkte oder Schnittmengen mit erstaunlich vielen in der Motivationspsychologie relevanten Konstrukten und sind vor allem auf der Prozessseite fester Bestandteil dessen, wie Motivation erklärt wird. Wertesystemen wird dabei die Funktion der Wertbeimessung zugeschrieben. Sie sind also die Maßstäbe, die herangezogen werden, um die Wertigkeit von zeitlich weniger stabilen oder weniger abstrakten Konstrukten zu bestimmen.

Inhaltliche Kategorisierungen von Werten finden jedoch wenig Eingang in die motivationspsychologische Forschungslandschaft. Hier dominieren Theorien zu Bedürfnissen oder den drei „großen“ Motiven Macht, Leistung und Anschluss. Zwar existieren Inhaltstheorien von Werten (z.B. Schwartz und Hofstede), die auch erforscht, jedoch wenig integriert in die klassische Motivationsforschung sind. Dies lässt sich auch daran erkennen, dass nur wenige Lehrbücher Werte-Inhaltstheorien beinhalten. Die Gravessche Theorie wird sogar nahezu vollständig vernachlässigt.

Kapitel 3

Messtheoretische und methodische Grundlagen

Im vorherigen Kapitel wurden die Grundlagen zum Wertekonstrukt gelegt. In diesem Kapitel werden nun die theoretischen Grundlagen zur Messung von Wertesystemen, insbesondere durch den Motivational Value Systems Questionnaire, bereit gestellt. Dafür wird zur Einführung in dieses Kapitel zunächst der MVSQ vorgestellt. Danach folgt ein Überblick der häufig verwendeten Fragebogenformate zur Messung von Werten bzw. Wertesystemen und im Anschluss die Darstellung und Analyse der Formateigenschaften des MVSQ inklusive der testtheoretischen Auswirkungen auf die Auswertung des MVSQ.

3.1 Der Motivational Value Systems Questionnaire

Der *Motivational Value Systems Questionnaire* (*MVSQ*) ist ein multidimensionaler forced-choice-Fragebogen, der entwickelt wurde, um persönliche Wertesysteme im beruflichen Kontext zu messen. Da Wertesysteme keine leistungsbezogenen Konstrukte sind und als zeitlich relativ stabile Konstrukte gelten (Jin & Rounds, 2012), kann der vorliegenden Fragebogen als Persönlichkeitstest verstanden werden (Jonkisz et al., 2012).

3.1.1 Aufbau des Instruments

Der MVSQ setzt sich aus zwei Subfragebögen und einem demographischen Teil zusammen. Beide Subskalen bestehen aus je zehn Fragen (die im FC-Design Blöcke genannt werden) mit jeweils sieben Antwortmöglichkeiten (Items) pro Block. Sie messen je sieben Annäherungs- und Vermeidungswertesysteme.

Im demographischen Teil, der zu Beginn präsentiert wird, werden Nationalität, Deutsch als Muttersprache, Geschlecht, Alter, Studiengang/Ausbildung, Dauer der Berufserfahrung sowie die Zustimmung zu den Datenschutzrichtlinien und der Verwendung der (anonymisierten)

Daten für wissenschaftliche Zwecke erhoben. Falls zutreffend, werden des Weiteren Branche, Funktion und Hierarchieebene abgefragt. Am Ende werden zudem noch Feedback-Fragen zum Verständnis und der Bedienung des Online-Interfaces gestellt.

Vor der Bearbeitung der 20 Blöcke erscheint eine schriftliche Anleitung, die den Bearbeitenden die Ranking-Prozedur erklärt. Seit der Revision des Instruments steht den Bearbeitenden zudem ein 1:45 Minuten dauerndes Video zur Verfügung, das die Prozedur anhand eines einfachen Beispiels im Screencast-Stil erläutert. Bei der Antwortprozedur handelt es sich um ein „Drag & Drop" Verfahren, in dem die Items per Maus oder Wisch-Geste von links nach rechts geschoben werden. Bei jeder Frage sind die Items zunächst auf der linken Seite des Bildschirms in Listenform angeordnet und müssen auf der rechten Seite in eine neue Reihenfolge gebracht werden, die den eigenen Wertepräferenzen entspricht.

Insgesamt misst der MVSQ 14 Merkmale, sieben Annäherungswertesysteme und sieben Vermeidungswertesysteme mit in Summe 140 Items, die sich gleichmäßig auf 20 Blöcke verteilen. Die beiden Subskalen bestehen jeweils aus zehn Blöcken und jeder Block enthält genau ein Item pro Wertesystem. Die Items bilden jeweils unterschiedliche Aspekte des Arbeitslebens ab. Die Wertesysteme wurden dabei als aggregierte Werturteile einer Person in verschiedenen Arbeitssituationen konzipiert, wobei die Itemstämme und Items der Annäherungswertesysteme als positiv formulierte Aussagen und die Itemstämme und Items der Vermeidungswertesysteme als negative Formulierungen operationalisiert wurden.

Sowohl Blöcke, wie auch die Items innerhalb der Blöcke werden in zufälliger Weise angeordnet. Diese Anordnung ist jedoch für alle Testbearbeitenden gleich. Die Blöcke decken folgende Arbeitssituationen ab: Aufgaben, Zusammenarbeit, Zielerreichung, Entscheidungen, Leistung, Konflikt, Anerkennung, Druck, Arbeitsbelastung und Umfeld (Team und Organisationskultur). Die Items der Annäherungswertesysteme repräsentieren erwünschte Zustände oder bevorzugtes Verhalten und die Items der Vermeidungswertesysteme unerwünschte Zustände und missfallendes Verhalten.

Im Folgenden ein Beispiel-Block (Itemstamm und sieben Items der Annäherungssubskala): „Ich bevorzuge ein Umfeld, in dem..."

- „ich mich wohl behütet fühle" (***Geborgenheit***)
- „man schnell Entscheidungen trifft und sofort in Aktion tritt" (***Macht***)
- „ordentlich und diszipliniert gearbeitet wird" (***Gewissheit***)
- „die Erfolgreichen unter sich sind" (***Erfolg***)
- „jeder die Möglichkeit hat, gehört zu werden" (***Gleichheit***)
- „Innovation und persönliche Weiterentwicklung im Fokus stehen" (***Verstehen***)
- „die gesellschaftlich relevanten Fragestellungen mein Handeln bestimmen" (***Nachhaltigkeit***)

3.1.2 Zur Entwicklung des Instruments

Ausgangspunkt für die Entwicklung des Fragebogens durch Falter und Singer war das Ziel, ein Instrument zu erstellen, mit dem die Wertesysteme von Mitarbeitern gemessen werden können, um dadurch Ansatzpunkte für die Erhöhung der Arbeitsmotivation abzuleiten (T. Falter, persönliche Kommunikation, 28.06.2012). Nach eigener Recherche gäbe es dafür keine Instrumente auf dem deutschen Markt, die gängige wissenschaftliche Gütekriterien erfüllen würden. Auf Basis ihrer berufsbedingten Expertise haben sie deshalb einen eigenen Fragebogen entwickelt. Die Entwicklungsstrategie hatte Anteile aus intuitiver und rationaler Itemkonstruktion (Jonkisz et al., 2012), die sich zum einen aus der beruflichen Praxis der Testkonstrukteure und zum anderen aus der Forschungsliteratur zum Gravesschen Wertemodell (Graves, 1966, 1970, 1971c, 1974) speiste. Außerdem war die Konstruktion tendenziell kriteriumsorientiert, da die Items mit dem Ziel formuliert wurden, dass sie möglichst eindeutig zwischen den Wertesystemen differenzieren.

Im ersten Schritt der Testkonstruktion wurden 26 Itemstämme und zwischen zwölf und 15 Items pro Wertesystem formuliert. Während der folgenden zwei Monate haben mehrere Experten die Items begutachtet und Feedback bereitgestellt. Darauf basierend wurden elf Blöcke ausgewählt, die dann von Kollegen, Verwandten und Freunden durch die beiden Formen *Lautes Denken* und *retrospektive Befragungen* getestet wurden. Nach weiteren Feedbackschleifen wurde die erste Version des Fragebogens mit zehn Blöcken pro Subskala fertiggestellt und die Online-Oberfläche programmiert. Der Fragebogen liegt ausschließlich in computerbasierter Form vor. Später wurde auf Basis einer Itemanalyse (Kapitel 5), die im Rahmen dieser Arbeit durchgeführt wurde, die erste Version überarbeitet.

Um die Bedeutung von Wertesystemen für die Motivation hervorzuheben, wurde der Fragebogen *Motivational Value Systems Questionnaire* im wissenschaftlichen Kontext und *my_motivation* für die Beratungspraxis getauft.

Annäherungs- und Vermeidungsdimension

Im MVSQ werden die sieben Wertesysteme in zwei Dimensionen, d.h. als Annäherungs- und Vermeidungswertesysteme gemessen. Die Annäherungsdimension drückt dabei aus, welche Wertesysteme wünschenswert sind, die Vermeidungsdimension, welche Wertesysteme nicht wünschenswert sind. In Übereinstimmung mit der Forschung zur Annäherungs- und Vermeidungsmotivation, werden die beiden Dimensionen unabhängig voneinander gemessen. Das hat zur Folge, dass eine Testperson auf beiden Dimensionen desselben Wertesystems gleichzeitig hohe bzw. niedrige Ausprägungen vorweisen kann. Dies kann so interpretiert werden, dass jedem Wertesystem auch negative Aspekte zugeschrieben werden können bzw. Wertesysteme negativ interpretiert werden können (T. Falter, persönliche Kommunikation, 06.03.2013). In der Praxis führt dies dazu, dass zum Beispiel Wettbewerb als positiv motivational erlebt werden

kann, wenn dadurch die Annäherungsdimension des Wertesystems ***Erfolg*** aktiviert wird, und als frustrierend und demotivierend, wenn stattdessen die Vermeidungsdimension von ***Erfolg*** aktiviert wird. Wie eine Wettbewerbssituation wahrgenommen wird, hängt wiederum vom Individuum, seinen Erfahrungen und Erwartungen ab.

Für die Forschung ergibt sich daraus eine weitere Fragestellung, die auch im Laufe dieser Arbeit beleuchtet werden soll. Diese Fragestellung lautet, ob die Hypothese der Unabhängigkeit der beiden Wertesystemdimensionen auch empirisch haltbar ist. Wären die Dimensionen unabhängig voneinander, dann bedeutet dies aus messtheoretischer Sicht, dass sie orthogonal zueinander stehen (Orthogonalitätshypothese). Das Gegenteil davon stellt die Bipolaritätshypothese dar. Danach handelt es sich bei entsprechenden Annäherungs- und Vermeidungswertesystemen um *ein* Wertesystem, das an beiden Enden des Merkmalskontinuums gemessen wird. Dieselbe Fragestellung wurde schon bei anderen Konstrukten untersucht. Zum Beispiel wurde positiver und negativer Affekt als orthogonal zueinander konzipiert (Watson & Tellegen, 1985), was an anderer Stelle allerdings angezweifelt wurde (Russell & Carroll, 1999). In dieser Arbeit soll deshalb die untergeordnete Forschungsfrage untersucht werden, ob die Orthogonalitätshypothese bzgl. der Annäherungs- und Vermeidungsdimension der Wertesysteme plausibel ist?

3.2 Wertemessung aus konzeptueller und praktischer Sicht

Wie in Kapitel 2 erörtert, handelt es sich bei Wertesystemen um Konstrukte, die hierarchisch angeordnet sind und für Menschen als psychologischen Bezugsrahmen beim Treffen von Werturteilen fungieren. Sie bestimmen, welcher psychologische Wert einer Wahrnehmung (z.B. Situation, Person, Objekt oder Ereignis) beigemessen wird. Der kognitive Prozess bei der Bildung eines Werturteils, d.h. eines Urteils, indem ein Wert als Vergleichskriterium verwendet wird, kann vereinfacht wie folgt dargestellt werden (Brosch, 2013; Moors, 2013; Smith & Ellsworth, 1985): Eine Wahrnehmung wird mit den persönlichen Wertesystemen verglichen. Je höher dabei das mit der Wahrnehmung korrespondierende Wertesystem in der eigenen Hierarchie steht, bzw. je stärker es relativ zu den anderen Wertesystemen ausgeprägt ist, desto höher ist der der Wahrnehmung beigemessene psychologische Wert. Als Beispiel sei ein Absolvent genannt, der als am höchsten ausgeprägtes Wertesystem ***Erfolg*** hat. Nach Abschluss seines Studiums erhält er zwei Jobangebote. Das erste verspricht viele Aufstiegsmöglichkeiten in einem wettbewerbsorientierten Umfeld. Das zweite bietet vor allem ein stabiles Umfeld mit vorhersehbaren Gehaltsentwicklungen und im Voraus geplanten Karrierestufen. Gemäß seiner Wertesystempräferenzen wird der Absolvent dem ersten Jobangebot einen höheren Wert beimessen als dem zweiten. Je höher dabei die Ausprägung des ***Erfolg***-Wertesystems ist, desto höher der beigemessene Wert und desto leichter wird dem Absolventen die Entscheidung fallen. Liegen mehrere Wertesysteme in der persönlichen Hierarchie gleichauf, konkurrieren diese miteinander und erschweren die Formulierung einer eindeutigen Präferenz. Dass zwei

Wertesysteme gleich stark ausgeprägt sind ist zwar denkbar, jedoch macht dieses Beispiel auch deutlich, dass Wertesysteme tendenziell hierarchisch angeordnet sein müssen. Denn ohne Präferenz wären Menschen in vielen Situationen nicht in der Lage, ein Werturteil zu fällen (Locke, 1991).

Im Folgenden werden unter Berücksichtigung dieser Konzeptualisierung nun verschiedene Fragebogenformate auf ihre Eignung zur Messung von Werten begutachtet. Der Fokus liegt dabei auf den gängigen Fragebogenformaten, insbesondere dem Rating- und dem Ranking-Format, die beide häufig zur Wertemessung eingesetzt werden. Zudem wird kurz auf weitere Formate wie z.B. Kurzaufsatz- und Ergänzungsaufgaben eingegangen.

3.2.1 Rating und Ranking

Beim Rating-Format werden Items unabhängig voneinander präsentiert und vom Testbearbeitenden bzgl. Zustimmung oder Ablehnung bewertet (Jonkisz et al., 2012). Häufig kommen dabei numerische oder verbale Skalen zum Einsatz, die dem Antwortenden mehrere Stufen der Bewertung anbieten. Bei der Auswertung gilt die Annahme, dass die Skalenstufen äquidistante Abstände zwischen einander aufweisen, was in der Praxis jedoch nicht überprüfbar ist und deshalb angezweifelt werden darf, inwiefern diese Annahme erfüllt ist (Jonkisz et al., 2012).

Bezogen auf die Wertemessung erscheint das Rating-Format auf Grund folgender zwei Eigenschaften weniger geeignet: (1) Wertesysteme gelten als hierarchisch angeordnet und stehen folglich in Bezug zueinander. Bei der Bewertung eines Rating-Items ist deshalb davon auszugehen, dass mehrere Wertesysteme zum Bewertungsergebnis eines Items beitragen und ein Rating-Item folglich nie nur *ein* Wertesystem misst, sondern das Zusammenspiel mehrerer Wertesysteme bei einer Beurteilung. (2) Da es sich bei Wertesystemen um Konstrukte handelt, die per Definition positiv (Annäherung) bzw. negativ (Vermeidung) sind, besteht die Gefahr, dass sich für die Annäherungswertesysteme Deckeneffekte und für die Vermeidungswertesysteme Bodeneffekte ergeben. Eine Differenzierung innerhalb der Dimensionen der Annäherung bzw. Vermeidung wird dadurch erschwert und die Tatsache, dass Wertesysteme innerhalb einer Dimension in Konkurrenz zueinander stehen, nicht berücksichtigt.

Forced-Choice-Formate – im Speziellen das Ranking-Format, indem mehrere Items gebunden werden, die dann von der Testperson in eine Reihenfolge gebracht werden müssen (Jonkisz et al., 2012) – bilden die vergleichende Natur des Wertekonstrukts (Kamakura & Mazzon, 1991; Meade, 2004) hingegen deutlich besser ab als das Rating-Format. Saville und Willson (1991, S. 222) formulieren es so: „Life is about choices“. Es (das Leben) erfordert ein ständiges Abwägen zwischen Alternativen, es gestaltet sich also als viele Aneinanderreihungen von Ranking-Aufgaben. Demzufolge spiegelt das FC-Format das reale Leben und die real vorkommenden kognitiven Prozesse des Abwägens und Bestimmens von Präferenzen besser wider als das Rating-Format. Zudem kann argumentiert werden, dass Rating-Aufgaben auch deshalb

realitätsferner als Rankingaufgaben sind, weil bei ihnen die individuelle Präferenz durch eine zusätzliche kognitive Leistung in eine Zahl oder Aussage (wie z.B. „stimme voll zu“) transformiert werden muss. Im Gegensatz dazu ist das Anordnen mehrerer Items in einer Reihenfolge intuitiver. Insbesondere unter Anbetracht der Konzeptualisierung des Wertemodells als Organisation mehrerer hierarchisch angeordneter Wertesysteme, erscheint die Operationalisierung von Wertemessungen durch eine Ranking-Prozedur deutlich plausibler.

In der Praxis gibt es Befürworter beider Arten. Zu den Fürsprecher von Rating-Verfahren zählen z.B. Braithwaite und Law (1985), Schwartz (1994) und Maio et al. (1996). Sie argumentieren, dass Rating sowohl methodologisch als auch konzeptuell besser passen würde. Zum einen haben Rating-Daten nützlichere statistische Eigenschaften. Z.B. können gängige statistische Verfahren aus der klassischen Testtheorie, der IRT oder Faktorenanalysen ohne Einschränkungen darauf angewendet werden. Ranking-Verfahren hingegen produzieren ipsative Daten, die einigen statistischen Restriktionen unterliegen. Darauf wird im weiteren Verlauf detailliert eingegangen. Zum anderen seien sich Menschen ihren Wertekonflikten nicht so bewusst, wie es bei Ranking-Prozeduren erforderlich wäre, um eine Präferenz zwischen zwei Werten machen zu können. Schwartz (1994) konstatierte deshalb, dass die Psychologie der Auswahl dadurch gut nachgebildet wird, wenn den Versuchspersonen eine Liste von Werten präsentiert wird, die sie zuerst durchlesen und dann jeden Wert einzeln auf einer Rating-Skala bewerten. Ball-Rokeach und Loges (1994), Feather (1975), Harzing et al. (2009) und Rokeach (1973) vertreten eine gegenteilige Ansicht. Für sie spiegelt eine Testsituation die Konzeption von Werten wider, in der verschiedene Optionen miteinander verglichen werden müssen und eine Entscheidung getroffen werden muss, welche Option am meisten bevorzugt wird. Sie sind folglich Befürworter der Messung von Werten und Wertesystemen durch Ranking. Kritisch ist dabei jedoch zu sehen, dass die Ranking-Befürworter keine Stellung dazu beziehen, dass ipsative Daten aus Ranking-Fragebögen nicht ohne Weiteres mit den gewöhnlichen Verfahren analysiert werden können. Zum Schluss sei gesagt, dass der Schwartz Value Survey (SVS), der am weitesten verbreitete Werte-Fragebogen zwar im Rating-Format gehalten wurde, jedoch in der Anweisung im Testmanual eine Quasi-Umformung in Ranking vorgenommen wird (Glöckner-Rist, 2012). Denn dort heißt es, dass zuerst alle (die Wertesysteme repräsentierenden) Items durchgelesen werden sollen, bevor sie beurteilt werden. Dies kann so gedeutet werden, dass auch Schwartz (1994) davon ausgeht, dass Wertesysteme untereinander konkurrieren und nur relativ zueinander adäquat beurteilt werden können. Warum der SVS dann nicht konsequenterweise im Ranking-Format gestaltet wurde, kann hier nicht beantwortet werden. Es bleibt jedoch die übergeordnete Schlussfolgerung, dass das Ranking-Format besser zur Messung von Wertesystemen geeignet ist als das Rating-Format.

3.2.2 Weitere Itemformate

Offene Formate, wie Kurzaufsatz- und Ergänzungsaufgaben (projektive Verfahren) mögen prinzipiell für die Messung von Wertesystemen geeignet sein. Aufgrund des hohen Auswertungsaufwands und der typischerweise „eingeschränkten Auswertungsobjektivität“ (Jonkisz et al., 2012, S. 41) gelten sie aber als weniger geeignet für psychometrische und experimentalpsychologische Untersuchungen. Außerdem werden projektive Verfahren vor allem dann eingesetzt, wenn entweder unbewusste Konstrukte oder sozial unerwünschte Konstrukte diagnostiziert werden sollen (Eid et al., 2015). Da Wertesysteme nicht nur das subjektiv Erwünschte, sondern auch das für eine Person allgemein – also sozial – Erwünschte ausdrücken (Rohan, 2000) und es sich bei Wertesystemen um bewusstseinsfähige Konstrukte handelt (Latham & Pinder, 2005; Locke & Henne, 1986; Schwartz & Bilsky, 1987), greifen beide Gründe, die für den Einsatz projektiver Verfahren sprechen würden, nicht.

3.3 Formateigenschaften des MVSQ und ihre messtheoretischen Auswirkungen

Beim MVSQ handelt es sich um ein FC-Format, da Items in Gruppen von sieben Items präsentiert werden und diese strikt in eine Reihenfolge gebracht werden müssen. Die Tatsache, dass der MVSQ im Ranking-Format gehalten wurde kann auf Basis der zuvor dargelegten Überlegungen als adäquat eingestuft werden.

Wertet man FC-Fragebögen jedoch nach „klassischen“ Prinzipien aus, d.h. summiert man die Ränge der Items desselben latenten Konstrukts über alle Blöcke auf, erhält man ipsative Testwerte, die einer Reihe von statistischen Restriktionen unterliegen. Im folgenden wird detaillierter auf den Begriff der Ipsativität eingegangen und die dadurch bedingten Eigenschaften des Instruments dargestellt.

3.3.1 Vor- und Nachteile des Forced-Choice-Formats

Forced-Choice-Fragebögen weisen einige Vorzüge gegenüber klassischen, Rating-basierten Fragebögen auf: Sie sind deutlich weniger anfällig für verschiedene Arten antwortverzerrendem Verhaltens. Eine Reihe von Untersuchungen legen diesen Schluss nahe. FC-Fragebögen sind z.B. weniger anfällig für Akquieszenz (Cheung & Chan, 2002), Halo-Effekte (Bartram, 2007) und Impression-Management bzw. soziale Erwünschtheit (Christiansen et al., 2005; Martin et al., 2002). Die Vermeidung von inhaltsunabhängigen Zustimmungs- oder Ablehnungstendenzen und der Tendenz zur Mitte ist dem FC-Format inhärent, da die Beantwortenden gezwungen werden, Präferenzangaben zu machen. Ebenso verhält es sich bei Halo-Effekten, da es im FC-Format unmöglich ist, Items gleich zu bewerten. Impression-Management und Effekte sozialer

Erwünschtheit werden im FC-Format vor allem dann reduziert, wenn sich Testpersonen in Situationen befinden, in denen sie sozial erwünscht antworten wollen. Anders formuliert zeigt sich die Resistenz gegenüber Faking in solchen Situationen deshalb deutlicher, da die Ergebnisse von Rating-Skalen darin stärker manipuliert werden (Jackson et al., 2000; Vasilopoulos et al., 2006). Dieser Unterschied dürfte dann noch stärker auftreten, wenn Items innerhalb einer Itemgruppe ausschließlich in dieselbe Richtung (unidirektional) kodiert sind, da dieselben Items im Rating-Format, je nach Kodierung, unweigerlich zu Decken- bzw. Bodeneffekten führen würden.

All diese positiven Eigenschaften gehen im Grundsatz darauf zurück, dass die Items in FC-Fragebögen nicht unabhängig voneinander, sondern in Gruppen präsentiert werden. Dies bedeutet allerdings auch, dass Items in FC-Instrumenten nicht unabhängig voneinander sind und dass dadurch eine der Grundannahmen der klassischen Testtheorie, nämlich das Axiom der Unabhängigkeit der Items bzw. der Fehler zwischen den Items, verletzt ist (Brown & Maydeu-Olivares, 2011). Die mit FC-Instrumenten erhobenen Daten sind also dadurch gekennzeichnet, dass sie untereinander abhängig sind. Man bezeichnet solche Daten als *ipsativ*. Ipsative Daten unterliegen einigen Einschränkungen bzgl. ihrer statistischen Verwertbarkeit (Brown & Maydeu-Olivares, 2011, 2013), was somit den Hauptnachteil von FC-Fragebögen darstellt.

3.3.2 Ipsativität und ihre Folgen

Der Begriff „ipsativ“ geht zurück auf Cattell (1944) und stammt vom Lateinischen *ipse* (dt. selbst) ab. Ipsative Daten sind also auf sich selbst bezogen und bringen eine Reihe problematischer Attribute mit, die vor allem die statistische Analyse und Auswertung betreffen. Ipsative Daten entstehen in Fragebogenformaten, in denen mehrere Items miteinander verglichen werden. Dazu zählen reine Ranking-Formate ebenso wie „Most/Least like me“-Formate oder Mischformate aus Ranking und Rating. All diese Formate haben gemeinsam, dass Items in Gruppen präsentiert werden und die Testpersonen Präferenzangaben unter Berücksichtigung aller präsentierten Items machen müssen, also *gezwungen* werden, eine Wahl vorzunehmen (*forced*-choice).

Der Vollständigkeit halber sei erwähnt, dass es verschiedene Abstufungen von Ipsativität gibt (Cattell & Brennan, 1994; Chan, 2003; Hicks, 1970). Zum einen gibt es *voll* oder *rein* ipsative Daten, die von Fragebögen generiert werden, in denen alle Items, die gleichzeitig präsentiert werden, in eine Präferenzreihenfolge gebracht werden müssen. Sie sind leicht daran erkennbar, dass die Summe – und folglich auch der Mittelwert – aller Merkmalsausprägungen bei allen Testpersonen stets gleich ist (Meade, 2004). *Partiell* ipsative Daten können z.B. in „Most/Least like me“- oder Mischformaten erzeugt werden.[1] Bei ihnen können sich die Mittelwerte aller Merkmalsausprägungen zwischen den Testpersonen unterscheiden (Hicks, 1970).

[1] In „Most/Least like me“-Formaten gilt dies dann, wenn mehr als vier Items gleichzeitig präsentiert werden, da bei drei Auswahlmöglichkeiten die Reihenfolge aller Items durch zwei Präferenzangaben bereits festgelegt ist und solche Daten ebenfalls *voll* ipsativ wären.

Darüber hinaus ist eine Eigenschaft, die die Abhängigkeit zwischen ipsativen Testwerten verdeutlicht, die erzwungenerweise negative Interkorrelation der Testwerte (Brown & Maydeu-Olivares, 2013; Meade, 2004). Vollipsative Testwerte korrelieren also immer negativ miteinander, wobei die Höhe der Korrelation direkt mit der Anzahl gemessener Konstrukte zusammenhängt. Bei zwei gemessenen Konstrukten (in Itemgruppen mit je zwei Items) ist die Korrelation perfekt negativ ($r = -1$), da die Präferenz eines Items automatisch auch die Position des zweiten Items bestimmt. Mit steigender Anzahl sinkt die mittlere negative Korrelation zwischen den Konstrukten. Sie kann über die Formel $\frac{-1}{k-1}$ mit k Konstrukten berechnet werden und beträgt beim MVSQ $r = -.17$. Eine weitere Eigenschaft vollipsativer Testwerte ist, dass es nicht möglich ist, dass eine Testperson ausschließlich hohe (oder ausschließlich niedrige) Ausprägungen auf allen latenten Konstrukten gleichzeitig hat (Brown & Maydeu-Olivares, 2013). Denn die hohe Bewertung eines Items geht automatisch mit einer niedrigen Bewertung eines anderen Items einher. Diese Eigenschaft fällt in der Praxis allerdings mit steigender Anzahl gemessener Konstrukte immer weniger ins Gewicht, da es mit jedem zusätzlichen Konstrukt weniger wahrscheinlich wird, eine Person in der Stichprobe vorzufinden, die auf allen Konstrukten ausschließlich über- oder unterdurchschnittliche Ausprägungen hat (Brown & Maydeu-Olivares, 2013). Untersuchungen von Baron (1996) und Saville und Willson (1991) haben gezeigt, dass die Ipsativität bei 30 gemessenen latenten Konstrukten nur sehr geringe Auswirkungen auf die Merkmalsausprägungen und deren Eigenschaften hat. Je weniger Konstrukte eine Skala jedoch misst, desto stärker wirkt sich die Ipsativität auf die Eigenschaften der Merkmalsausprägungen aus. Eine Studie von Meade (2004), die ein Instrument untersucht, das acht Konstrukte erhebt, berichtet von erheblichen Verzerrungen der Merkmalsausprägungen, die sich durch das ipsative Format erklären lassen. In einer anderen Studie (Cornwell et al., 1991) haben sich auch drastische Auswirkungen von Ipsativität auf die Merkmalsausprägungen bei vier Merkmalen gezeigt. Für die Skalen des MVSQ mit je sieben gemessenen Wertesystemen pro Subskala bedeutet dies folglich, dass von erheblichen Auswirkungen der Ipsativität auf die psychometrischen Eigenschaften des Instruments ausgegangen werden kann.

Beim MVSQ können die Auswirkungen der Ipsativität konkret in zwei Bereiche gegliedert werden. Der eine betrifft die Durchführung verschiedener psychometrischer Analysen, für die die Daten des MVSQ nicht geeignet sein dürften. Zum Beispiel wurde in mehreren Studien (Cornwell et al., 1991; Dunlap & Cornwell, 1994; Johnson et al., 1988; Meade, 2004) gezeigt, dass ipsative Daten (vergleichbar umfangreicher Instrumente) sowohl ungeeignet sind, um damit Faktorenanalysen durchzuführen, als auch wenig brauchbar für die Berechnungen von Reliabilitäten sind (Brown & Maydeu-Olivares, 2013; Meade, 2004; Tenopyr, 1988). Laut Bartram (1996) liefern klassische Reliabilitätsberechnungen von Instrumenten, die weniger als zehn Konstrukte messen oder Skaleninterkorrelationen größer als .30 haben, keine verlässlichen Ergebnisse. Auch die Gefahr von Fehlinterpretationen ipsativer Testwerte ist hoch (Meade, 2004; Tenopyr, 1988) und sowohl Konstrukt- als auch Kriteriumsvaliditäten unterliegen bei

vergleichbaren Instrumenten Verzerrungen, deren Ausmaß nicht bestimmbar ist (Brown & Maydeu-Olivares, 2013; Johnson et al., 1988). Zwar kann es auch sein, dass ipsative Daten und die daraus berechneten Reliabilitäts- und Validitätskoeffizienten wenig verzerrt sind (Merritt & Marshall, 1984; Tamir & Lunetta, 1977). Um dies zu überprüfen, sind jedoch parallele Messungen mit nicht-ipsativen Instrumenten erforderlich. Da es derzeit keine parallelen Instrumente gibt, kann die Format-bedingte Verzerrung im MVSQ bei klassischer Auswertung nicht bestimmt werden.

Der zweite Bereich betrifft die Frage nach der Verwendung der ipsativen Testwerte. Dazu sei zunächst der Begriff *normativ* erklärt. Testwerte sind dann normativ, wenn sie verwendet werden können, um Merkmale einer Person *relativ* zu einer Population zu interpretieren (Cattell, 1944). Testwerte, die mit klassischem Rating erhoben werden, erlauben diesen Vergleich, wenn sie unabhängig voneinander sind. Das ist in der Regel der Fall, weshalb diese als normativ bezeichnet werden können (Hicks, 1970). Ipsative Testwerte hingegen sind aufgrund ihrer „Selbstbezogenheit“ nicht brauchbar, um einen sogenannten interindividuellen Vergleich der Merkmalsausprägung durchzuführen (Clemans, 1966; Johnson et al., 1988). Sie können nur sinnvoll für den intraindividuellen Vergleich von Merkmalen, also den Vergleich der relativen Wichtigkeit mehrerer Merkmale zueinander innerhalb einer Person verwendet werden (Closs, 1996; Hicks, 1970).

Abschließend sei noch hinzugefügt, dass Eigenschaften ipsativer Formate, wie eine geringere Anfälligkeit für Antwortverzerrungen, bewirken können, dass Forced-Choice-Messungen im Vergleich zu Rating-Messungen validere Schlussfolgerungen erlauben (Bartram, 2007). Dies gilt insbesondere in Situationen, in denen „viel auf dem Spiel steht“, wie z.B. in Bewerbungssituationen (Christiansen et al., 2005; Martin et al., 2002; Vasilopoulos et al., 2006). Da sich die Testpersonen dieser Arbeit jedoch nicht in solchen Situationen befinden, haben diese Befunde keine Konsequenzen für die Ziele der folgenden Untersuchungen.

Zusammenfassend kann gesagt werden, dass die Skalen des MVSQ von den problematischen Eigenschaften des ipsativen Formats betroffen sein werden, da sie lediglich sieben latente Konstrukte messen. Da es zudem keine parallele normative Messung der Wertesysteme gibt, kann das Ausmaß der vom ipsativen Format bedingten Verzerrungen nicht bestimmt werden. Aufgrund der zitierten Untersuchungen vergleichbarer Instrumente kann davon ausgegangen werden, dass die ipsativen Daten des MVSQ wenig brauchbar für Untersuchungen der Reliabilität und Validität sind. Allerdings gibt es seit kurzem einen Lösungsansatz, der die Probleme ipsativer Daten lösen kann. Dieser wird im folgenden Abschnitt vorgestellt.

3.4 Das Thurstonian IRT-Modell

Vor kurzem entwickelten Brown (2010) und Brown und Maydeu-Olivares (2011, 2013) einen probabilistischen Ansatz, der es ermöglicht, die problematischen Eigenschaften ipsativer Daten

zu umgehen. Dem „Law of Comparative Judgement“ von Louis L. Thurstone (1927) Tribut zollend, wurde das Modell *Thurstonian IRT-Modell*, kurz TIRT-Modell genannt. Der Kern des Ansatzes liegt darin, dass im Gegensatz zum klassischen Scoring die vergleichende Natur von Rangdaten berücksichtigt wird und diese als paarweise Vergleiche modelliert werden. Der Ausgang eines Paarvergleichs geht gemäß dem *Law of Comparative Judgement* auf das Verhältnis der beiden involvierten Merkmalsausprägungen zurück. Ist Merkmal A höher ausgeprägt als Merkmal B, so wird auch Item A über Item B gesetzt.

Im TIRT-Modell werden die Mittelwerte der Paarvergleiche (Utilities), die Faktorladungen auf die je zwei latenten Konstrukte (im MVSQ Wertesysteme) und die Kovarianzen der Paarvergleiche als strukturierte „Fehlerterme“ (Uniquenesses) modelliert. Brown und Maydeu-Olivares (2011, 2013) konnten in einigen Simulationsstudien, wie auch in Untersuchungen mit echten Daten zeigen, dass die Anwendung des TIRT-Modells unter bestimmten Voraussetzungen Schätzungen der Merkmalsausprägungen liefert, die von den Einschränkungen der Ipsativität befreit sind.

Die Herangehensweise der TIRT basiert auf folgenden faktorenanalytischen Gleichungen (Brown & Maydeu-Olivares, 2011, 2013):

$$y^*_{ik} = t_i - t_k \tag{1}$$

wobei t den latenten Nutzwert (engl. *latent utility*) der Items i bzw. k darstellt und y^*_{ik} eine kontinuierliche latente Variable ist, die den Ausgang des Paarvergleichs repräsentiert. Ist $y^*_{ik} \geq 0$, also der Nutzwert von Item i größer als der von Item k, dann muss Item i nach dem „Law of Comparative Judgement“ über Item k gerankt werden. Umgekehrt verhält es sich, wenn $y^*_{ik} < 0$ ist.

Der Nutzwert t von Item i wird nun wie folgt modelliert:

$$t_i = \mu_i + \lambda_i \eta_a + \varepsilon_i \tag{2}$$

mit μ_i als Mittelwert des Nutzwerts des Items (Utilitiy), der Ladung λ_i auf das latente Konstrukt η_a und ε_i als „unique factor“ (Brown & Maydeu-Olivares, 2013, S.41), sozusagen dem nicht zuordenbaren Rest, der als Fehlerterm verstanden werden kann und deshalb als *Uniqueness* bezeichnet wird. Analog gestaltet sich die Gleichung für Item k. Setzt man die Formeln für t_i und t_k nun in die obere Gleichung ein, erhält man die gesamte Gleichung für den latenten Nutzwerts eines Paarvergleich:

$$y^*_{ik} = t_i - t_k = (\mu_i - \mu_k) + (\lambda_i \eta_a - \lambda_k \eta_b) + (\varepsilon_i - \varepsilon_k) \tag{3}$$

wobei Item i dem latenten Konstrukt η_a und Item k η_b zuzuordnen ist. Im Vergleich zur üblichen mathematische Modellformulierung (z.B. τ - kongenerischen Messmodell Eid et al., 2015, S.864) sind dabei zwei zentrale Unterschiede hervorzuheben:

1. Die „manifeste" Variable y^* ist latent, d.h. nicht direkt beobachtbar und nur indirekt über den Ausgang der Paarvergleiche bestimmbar, wobei gilt: Ist $y^*_{ik} \geq 0$ bedeutet dies, dass Item i über k bevorzugt wird und für $y^*_{ik} < 0$ wird Item k über Item i gerankt.

2. y^* hängt von *zwei* latenten Konstrukten (η_a und η_b) ab.

Ferner setzt sich die Uniqueness von y^*_{ik} aus den beiden Residualvarianzen ε_i und ε_k der im Paarvergleich involvierten Items zusammen (Brown & Maydeu-Olivares, 2011). Die Uniquenesses repräsentieren den Fehlerterm pro Paarvergleich.

Zusammenfassend sei gesagt, dass TIRT-Modelle neben den Uniquenesses aus den Parametern Utilities und Faktorladungen bestehen. Für die Utilities (mittlere Nutzenwerte der Items) gilt, je höher dessen Werte, desto höher ist der mittlere Nutzwert und umso leichter ist dieses Item folglich zu bevorzugen. Diese Parameter können demnach als Leichtigkeitsparameter (Eid et al., 2015) oder als inverse Schwierigkeitsparameter verstanden werden. Die Faktorladungen sind die Steigungsparameter in der TIRT-Modellgleichung. Sie beschreiben die Diskriminationsfähigkeit eines Items (Eid et al., 2015) und sind damit die Entsprechung zur Trennschärfe in der klassischen Testtheorie.

Wie eingangs angedeutet, trägt das TIRT-Modell folglich dem kognitiven Prozess des Vergleichens bei der Bearbeitung von FC-Fragebögen Rechnung, indem die Abhängigkeiten zwischen den Items eines Blocks modelliert werden. Dadurch können zufällige Messfehler, die auf das paarweise Vergleichen der Items eines Blocks zurückgehen, mathematisch berechnet und dadurch berücksichtigt werden.

In den folgenden Absätzen werden nun eine Reihe von Eigenschaften des MVSQ im Hinblick auf die Anwendbarkeit des TIRT-Ansatzes untersucht. Zum besseren Verständnis dafür seien hier die relevanten Eigenschaften des MVSQ zusammengefasst: Der MVSQ besteht aus zwei Subskalen, die sich aus jeweils zehn Blöcken mit je sieben Items zusammensetzen. Dadurch ist ebenso sie Zahl der Paarvergleiche mit 21 pro Block und 210 pro Skala festgelegt. Alle Items sind in dieselbe Richtung, d.h. unidirektional kodiert. Die durchschnittliche Korrelation der Merkmale liegt je Skala bei -0.17. Die *wahren* Merkmals-Interkorrelationen sind zwar nicht bekannt, jedoch lassen die teils sehr hohen (> 0.5) ipsativen Merkmals-Interkorrelationen vermuten, dass auch einige der *wahren* Korrelationen positiv sind.

Anzahl der latenten Konstrukte

Laut Brown und Maydeu-Olivares (2011) können die absoluten Merkmalsausprägungen dann gut geschätzt werden, wenn die Zahl der latenten Traits „groß" (S. 495) ist. Bei welcher Zahl

„groß“ beginnt, wird von den Autoren nicht näher spezifiziert. Sie konstatieren jedoch, dass die Schätzung bei zwei gemessen Merkmalen nicht möglich ist, wenn die Items unidirektional kodiert sind, da bei zwei Items pro Block der Rang eines Items automatisch auch den des anderen bedingt. Aus einer Simulationsstudie mit fünf Merkmalen kann jedoch geschlossen werden, dass die Modellparameter bei fünf Merkmalen mit bidirektionaler Kodierung der Items weitestgehend verzerrungsfrei geschätzt werden können. Wenn nur unidirektionale Items verwendet werden, dann hängt die Genauigkeit der Schätzung stark von der Anzahl der gemessenen Merkmale ab, wobei mit zunehmender Anzahl an Merkmalen mit besseren Schätzergebnissen zu rechnen ist (Brown & Maydeu-Olivares, 2011). Bei fünf Merkmalen und unidirektionalen Items (in Vierer-Blöcken) lag die Verzerrung der Schätzungen der Traitkorrelationen bei Brown und Maydeu-Olivares (2011) bei knapp 10%. Faktorladungen, Mittelwerte und Uniquenesses werden jedoch bereits weitestgehend verzerrungsfrei geschätzt. Forschungsbedarf besteht hier bzgl. der Frage, wie groß Verzerrungen bei sieben gemessenen Merkmalen mit sieben Items pro Block sind, wobei aus den eben dargelegten Befunden zumindest geschlussfolgert werden kann, dass die größere Anzahl latenter Traits im MVSQ zu weniger als 10% Verzerrung der Traitkorrelationen führen sollte und Faktorladungen, Utilities und Uniquenesses verzerrungsfrei bestimmt werden sollten.

Kodierung der Items

Brown und Maydeu-Olivares (2011) legen dar, dass Testwerte ungeachtet der Merkmalskorrelationen und der Anzahl der Merkmale dann besser geschätzt werden können, wenn es Items eines Merkmals gibt, die bidirektional, d.h. in positive und negative Richtung kodiert sind. Bzgl. der Kodierung des MVSQ muss deshalb festgestellt werden, dass diese nicht optimal ist, um ein TIRT-Modell zu fitten.

Korrelationen zwischen den Merkmalen

Des Weiteren funktioniert der von Brown und Maydeu-Olivares (2012) angewandte Schätzalgorithmus (DWLS) bei unidirektionaler Kodierung der Items besser, wenn Merkmale nicht miteinander korrelieren, als wenn sie positiv korrelieren (Brown & Maydeu-Olivares, 2011). Noch effektiver wird er, wenn die Merkmale negativ miteinander korrelieren. Da im MVSQ jedoch von einigen stark positiven Merkmalskorrelationen auszugehen ist, gilt auch hier, dass der Aufbau des MVSQ nicht optimal für die effiziente Schätzung eines TIRT-Modells geeignet ist.

Anzahl und Größe der Blöcke

Je mehr Items ein Block enthält, desto mehr Informationen können aus einem Block gewonnen werden (Brown & Maydeu-Olivares, 2011), da mit steigender Itemzahl die Anzahl der

Paarvergleiche – und damit der Informationsträger – exponentiell ansteigt und dadurch eine höhere Messgenauigkeit erreicht werden sollte (Brown & Maydeu-Olivares, 2011). Allerdings ist die Blockgröße aus praktischer Sicht dadurch begrenzt, dass mit steigender Größe auch die kognitive Beanspruchung für den Testbearbeitenden zunimmt. Brown und Maydeu-Olivares (2011) gehen von einer Obergrenze von vier Items pro Block aus. Sie legen dafür allerdings keine Begründung dar. Da es neben dem MVSQ jedoch in der Praxis erprobte Instrumente gibt, wie z.B. den *Rokeach Value Survey* (Feather & Peay, 1975; Rokeach, 1973), in dem mit 18 Werten eine deutlich größere Anzahl Konstrukte vom Bearbeitenden in eine Rangreihenfolge gebracht werden muss, ist fraglich, ob die von Brown und Maydeu-Olivares postulierte Obergrenze wirklich für den praktischen Einsatz gilt. Auch die Praxiserfahrung beim Einsatz des MVSQ zeigt, dass sieben Items pro Block keine größeren Probleme für die Bearbeitenden darstellt.

Ferner konnten Brown und Maydeu-Olivares (2011) zeigen, dass die empirischen Reliabilitäten mit mehr Blöcken verlässlicher geschätzt werden können. Dies verwundert nicht, da mit steigender Anzahl an Blöcken ebenso wie mit steigender Anzahl Items pro Block die Zahl der Paarvergleiche und damit die zur Verfügung stehende Information steigt. Brown und Maydeu-Olivares haben in ihren Simulationsstudien Big-Five-Fragebögen mit unterschiedlichen Blockgrößen und -zahlen simuliert (60, 90, 120 und 180 Paarvergleiche) und konnten bereits bei einer Blockgröße von drei und 60 Paarvergleichen die fünf Merkmale verlässlich messen. Die 210 Paarvergleiche der MVSQ-Subskalen sollten folglich ausreichend Information produzieren, um die Ausprägungen der sieben Merkmale zu schätzen.

Außerdem interagieren Blockgröße und Kodierung der Items. Optimal hinsichtlich Verzerrungen der geschätzten Modellparameter sind nach Brown und Maydeu-Olivares (2011) Simulationsstudien urteilend, Blöcke mit vier Items *und* bidirektional kodierten Items. Hat man ein Modell mit nur einem von beiden (vier Items pro Block *oder* bidirektionale Kodierung), leiden die Genauigkeiten der Schätzungen darunter. Für den MVSQ bedeutet dies, dass die Blockgröße von sieben prinzipiell vorteilhaft ist, da daraus 21 Paarvergleiche pro Block generiert werden können und somit exponentiell mehr Informationen zum Schätzen der Ausprägungen zur Verfügung stehen. Wie oben bereits erläutert, ist die Tatsache, dass die Items im MVSQ unidirektional kodiert sind, nicht optimal für den Schätzvorgang.

Zusammenfassend kann gesagt werden, dass die Anwendung des TIRT-Ansatzes auf den MVSQ möglich sein sollte, wenngleich die Voraussetzungen dafür nicht optimal sind.

3.5 Alternative Ansätze zur Modellierung von FC Daten

Neben dem TIRT-Ansatz gibt es noch einige weitere Ansätze, die der Ipsativität von FC-Daten Rechnung tragen (für einen Überblick s. Brown, 2014). Dazu gehören die Ansätze von Zinnes und Griggs (1974), Stark et al. (2005) und McCloy et al. (2005). Allen drei Ansätze ist gemein, dass sie der Kategorie der *ideal-point*-Modelle zuzuordnen sind. D.h sie gehen davon aus, dass

Items dann von einer Person gewählt werden, wenn sie der tatsächlichen Merkmalsausprägung der Person entsprechen, nicht aber, wenn die vom Item widergespiegelte Merkmalsausprägung stark über oder unter der Merkmalsausprägung liegt. Im Gegenzug dazu handelt es sich beim TIRT-Ansatz um ein *dominance*-Modell, d.h. es wird angenommen, dass ein Item auch dann von einer Person gewählt werden kann, wenn die Merkmalsausprägung höher ist als vom Item indiziert. Für die Praxis bedeutet dies, dass *ideal-point*-Modelle für Items geeignet sind, die nicht nur *dominante* (hoch oder niedrig) Ausprägungen von Merkmalen wiedergeben, sondern auch mittlere Ausprägungen indizieren. Ein solches Item für das Wertesystem ***Verstehen*** könnte z.B. so formuliert sein: „Manchmal denke ich mich gerne in komplexe Fragestellungen hinein". Da es sich bei den Items des MVSQ jedoch um *dominante* Itemformulierungen handelt (z.B. „Ich liebe Aufgaben, bei denen ich eigene theoretische Konzepte entwickle" für ***Verstehen***), sind die Ansätze nicht sinnvoll auf die Daten des MVSQ anwendbar.

3.6 Die Güte psychologischer Fragebögen

Nachdem nun der MVSQ, dessen Eigenschaften und deren testtheoretische Auswirkungen vorgestellt wurden, folgt im kommenden Abschnitt die Einführung des Konzepts der psychometrischen Güte. Die psychometrische Güte eines Fragebogens lässt sich anhand der Hauptgütekriterien Objektivität, Reliabilität und Validität untersuchen (Moosbrugger & Kelava, 2012). Diese Arbeit behandelt eben diese Hauptgütekriterien, die zur einfacheren Lesbarkeit in Abbildung 1 grafisch dargestellt werden.

Darüber hinaus gehören in das Testmanual eines Instruments auch Untersuchung von Nebengütekriterien wie Skalierung[2], Normierung, Testökonomie, Nützlichkeit, Zumutbarkeit, Unverfälschbarkeit und Fairness (Amelang & Schmidt-Atzert, 2006; Lienert & Raatz, 1998; Moosbrugger & Kelava, 2012).

3.6.1 Objektivität

Das Gütekriterium Objektivität ist dann erfüllt, wenn die Messungen eines Tests unabhängig von Testleiter und Testauswerter sind (Bühner, 2011; Moosbrugger & Kelava, 2012). Diese beiden Aspekte werden als Durchführungs- und Auswertungsobjektivität bezeichnet und garantieren, dass die Testergebnisse verschiedener Personen vergleichbar sind, da bei sichergestellter Durchführungs- und Auswertungsobjektivität sowohl der Einfluss von Störfaktoren bei der Bearbeitung als auch Fehler bei der Auswertung ausgeschlossen werden können. Einen dritten wichtigen Aspekt der Objektivität als Gütekriterium stellt die Interpretationsobjektivität dar (Moosbrugger & Kelava, 2012). Sie besagt, dass die Ergebnisse eines Tests unabhängig von der interpretierenden Person auf gleiche Weise interpretiert werden. Abschließend ist zu sagen,

[2]Manche Autoren, z.B. Bühner (2011) zählen Skalierung auch zu den Hauptgütekriterien.

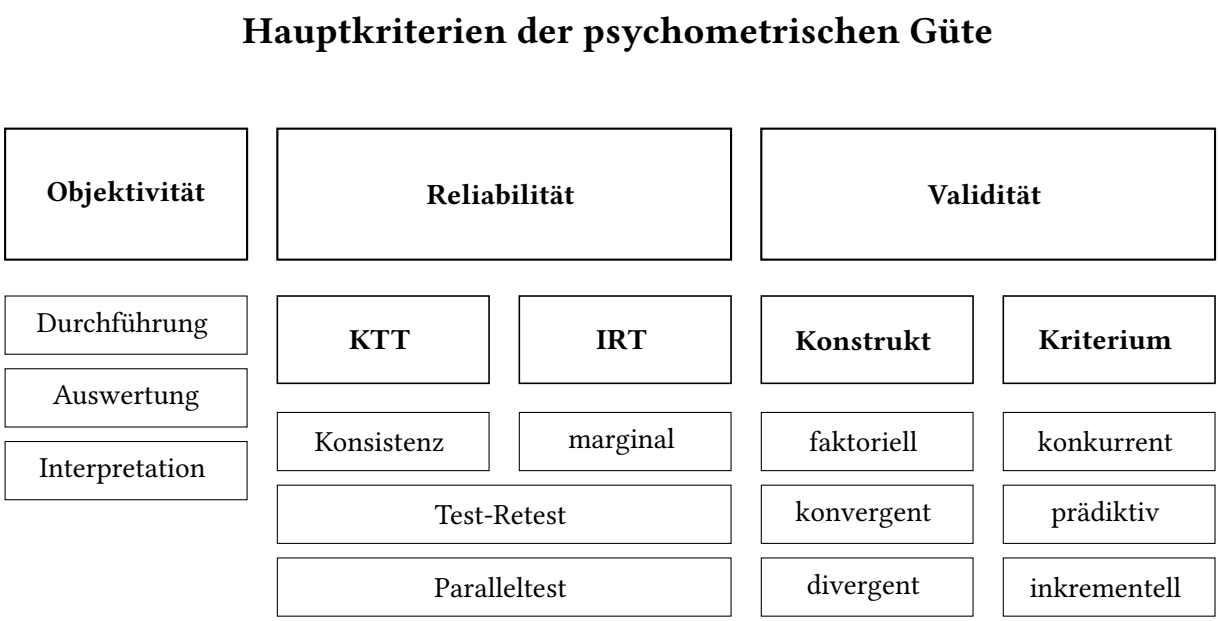

Abbildung 1. Überblick der Hauptgütekritieren und ihrer Elemente.

dass die beiden erstgenannten Aspekte der Objektivität eine Grundvoraussetzung für das nächste Gütekriterium, die Reliabilität, darstellen (Schermelleh-Engel & Werner, 2012). Denn eine Messung kann nur dann hoch reliabel sein, wenn Durchführungs- und Auswertungsobjektivität gegeben sind.

3.6.2 Reliabilität

Die Reliabilität eines Tests bezeichnet dessen Messgenauigkeit bzgl. der zu messenden Konstrukte und zwar ungeachtet der Frage, ob auch das richtige Konstrukt gemessen wird (Amelang & Schmidt-Atzert, 2006; Bühner, 2011). Testtheoretisch gesehen ist die Reliabilität das Verhältnis der True-Score-Varianz zur Gesamtvarianz der Testwerte (Lord et al., 1968), d.h. je ähnlicher sich die *wahre* Varianz und die gesamte Varianz eines Merkmals sind, desto geringer ist folglich der Messfehler und umso höher die Reliabilität. In der Praxis gibt es unterschiedliche Methoden, um die Messgenauigkeit eines Tests zu bestimmen, die je unterschiedliche Annäherungen an die Messgenauigkeit eines Tests darstellen (Amelang & Schmidt-Atzert, 2006; Eid et al., 2015; Schermelleh-Engel & Werner, 2012). In der klassischen Testtheorie (KTT) zählen dazu die Test-Retest-Reliabilität, die Paralleltest-Reliabilität, die Testhalbierungs-Reliabilität und Maße der internen Konsistenz. Die Test-Retest-Reliabilität (auch Test-Wiederholungs-Reliabilität) kann dabei als Maß der (zeitlichen) Stabilität eines Tests verstanden werden (Amelang & Schmidt-Atzert, 2006). Die Paralleltest-Reliabilität „gilt als das beste Verfahren“(Lienert & Raatz, 1998, S.182) der Reliabilitätsbeurteilung, da sich dieser Ansatz auf eine größere Anzahl an Items und damit eine breitere Abbildung der Merkmalsmessung bezieht (Amelang & Schmidt-Atzert,

2006). Sie ist sozusagen die generalisierbarste Methode der Reliabilitätsbestimmung. Des Weiteren gibt es die sogenannte Testhalbierungsreliabilität, bei der Tests in zwei gleichwertige Testhälften zerlegt werden und die Übereinstimmung beider Testhälften die Messgenauigkeit des Tests ausdrückt (Amelang & Schmidt-Atzert, 2006). Maße der Internen Konsistenz (wie z.B. Cronbachs α) stellen eine Verallgemeinerung der Testhalbierungsreliabilität dar, indem darin ein Test nicht nur in zwei, sondern in so viele Teile zerlegt wird, wie es Items gibt (Amelang & Schmidt-Atzert, 2006). Anders formuliert, kann die Testhalbierungs-Reliabilität als Spezialfall der internen Konsistenz gesehen werden. Der wohl größte Vorteil von Konsistenzmaßen im Vergleich zu den anderen Reliabilitätsmaßen liegt darin, dass sie die praktikabelste Berechnung der Reliabilität darstellen (Schermelleh-Engel & Werner, 2012). Das liegt daran, dass eine einzige Testadministration ausreicht, um sie zu berechnen. Es sind keine parallelen Testformen erforderlich und auch die Zuordnung von Testhälften erübrigt sich.

In der IRT können von diesen KTT-Methoden die Paralleltest- sowie die Test-Retest-Reliabilität berechnet werden, Maße der internen Konsistenz hingegen nicht. Der Grund dafür liegt in der Konzeptualisierung des Standardmessfehlers. In der KTT wird angenommen, dass der Standardfehler der Messung für alle Personen einer Population gleich ist, was bedeutet, dass das Merkmal an allen Stellen des Merkmalskontinuums mit dem gleichen Fehler gemessen wird (Embretson & Reise, 2000; Irtel, 1996). Im Gegensatz dazu gilt in der IRT, dass sich der Standardmessfehler an unterschiedlichen Merkmalsausprägungen unterscheiden kann, aber auf die gesamte Bevölkerung verallgemeinerbar ist (Embretson & Reise, 2000). Als Folge kann in der IRT für jede Merkmalsausprägung ein Messgenauigkeitswert bestimmt werden, und zwar sowohl auf Itemebene als auch auf den gesamten Test bezogen (Moosbrugger, 2012). Die entsprechenden Werte können mittels Iteminformationsfunktionen bzw. Testinformationsfunktionen (additive Iteminformationen) berechnet werden und z.B. zum Vergleich zweier Testformen herangezogen werden, indem für unterschiedliche Merkmalsausprägungen die entsprechenden Testinformationen berechnet werden (Moosbrugger, 2012). Die Berechnung einzelner Koeffizienten für den gesamten Test ist damit allerdings nicht möglich und auch nicht unbedingt erforderlich, da die Messgenauigkeit damit sehr detailliert begutachtet werden kann. Allerdings gibt es auch in der IRT die Möglichkeit, einzelne Koeffizienten zu berechnen, die die Messgenauigkeit eines gesamten Tests in einer Zahl widerspiegeln. Sie werden als *marginale* Reliabilitätskoeffizienten bezeichnet und stellen einen Mittelwert der Messgenauigkeit über ein Traitkontinuum dar (Ayala, 2013; Brown & Croudace, 2015; Green et al., 1984). Sie können deshalb als eine Entsprechung zur internen Konsistenz in der KTT verstanden werden. Vorteil solcher Durchschnittsindizes ist, dass sie eine praktische Möglichkeit darstellen, um die Reliabilitäten mehrerer Tests leicht, d.h. anhand einer Zahl, miteinander zu vergleichen, auch wenn dabei möglicherweise Methodeneffekte nicht berücksichtigt werden (Kim, 2012b). Die daraus resultierende einfache Vergleichsmöglichkeit ist vermutlich auch der Grund, warum solche komprimierten Koeffizienten überhaupt berechnet werden. Beispiele für solche Maße in der

IRT sind Andrich's zusammengesetzter Reliabilitätskoeffizient (Andrich, 1988; Embretson & Reise, 2000) und das Konzept der empirischen Reliabilität (Green et al., 1984; Maydeu-Olivares & Brown, 2010).

Für den MVSQ als ipsatives Instrument gilt nun, dass Reliabilitätsmaße, die die Ipsativität nicht berücksichtigen, nicht geeignet sind, um dessen Messgenauigkeit auszudrücken (vgl. Kapitel 3.3.2). Die klassischen Maße Testhalbierungsreliabilität und Cronbachs α berücksichtigen die Ipsativität nicht und sind deshalb nicht sinnvoll interpretierbar. Des Weiteren ist die Bestimmung der Paralleltests-Reliabilität nicht durchführbar, weil parallele Testformen fehlen. In den einschlägigen Datenbanken konnten weder klassische noch ipsative Instrumente gefunden werden, die Wertesysteme gemäß der Gravesschen Theorie messen. Und selbst wenn es ein solches auf der KTT basierendes Instrument gäbe, wäre zu prüfen, ob die unterschiedlichen Erhebungsmethoden der Bedingung der Parallelität gerecht würden (Amelang & Schmidt-Atzert, 2006). Als weiteres Reliabilitätsmaß, dass im vorliegenden Fall nicht bestimmt werden kann, ist Andrich's Reliabilität zu nennen. Diese wurde nicht für IRT-Modelle ipsativer Daten entwickelt und es gibt aktuell keine Umsetzung des Konzepts für TIRT-Modelle. Laut Brown und Maydeu-Olivares (2013) ist die Berechnung eines solchen Maßes bei großen TIRT-Modellen aufgrund der extrem hohen Komplexität und der dafür erforderlichen Rechenleistung derzeit nicht machbar.[3] Die Test-Retest-Reliabilität kann für den MVSQ dann berechnet werden, wenn die Messungen zu den unterschiedlichen Zeitpunkten unter Berücksichtigung der Ipsativität auswertet werden. Dies ist im vorliegenden Fall dann möglich, wenn die Merkmalsausprägungen mit dem TIRT-Ansatz ermittelt werden. Ferner kann auch die empirische Reliabilität des MVSQ über die Schätzung von TIRT-Modellen bestimmt werden (Brown & Maydeu-Olivares, 2013). Dies wird von Entwicklern der des TIRT-Ansatzes (Brown & Maydeu-Olivares, 2013) auch empfohlen.

3.6.2.1 Die empirische Reliabilität

Da es sich bei der empirischen Reliabilität um ein selten verwendetes Maß der Messgenauigkeit handelt, wird es an dieser Stelle vorgestellt. Es handelt sich dabei um einen simulationsbasierten Ansatz, indem, ausgehend von einem an originale Daten angepassten IRT-Modell, *wahre* und *geschätzte* Merkmalsausprägungen ermittelt und miteinander in Bezug gesetzt werden (Brown & Maydeu-Olivares, 2013). Da die geschätzten Scores dabei die Messfehlervarianz enthalten, beschreibt das Verhältnis dieser beiden Scores das Verhältnis der True-Score- zur Gesamt-Varianz und somit die Messgenauigkeit einer Erhebung (Maydeu-Olivares & Brown, 2010). Berechnet werden kann die empirische Reliabilität als quadrierte Korrelation dieser beiden Scores und notiert wird sie üblicherweise als ρ (Maydeu-Olivares & Brown, 2010).

[3] Ein TIRT-Modell, das auf Daten des MVSQ basiert kann mit 210 Paarvergleichen pro Skala als zu komplex eingestuft werden (vgl. Brown & Maydeu-Olivares, 2013).

Die exakte Vorgehensweise zur Berechnung der empirischen Reliabilität setzt sich aus sechs Schritten zusammen, die in Abbildung 2 verbildlicht wurden (Brown & Maydeu-Olivares, 2011, 2013; Maydeu-Olivares & Brown, 2010):

1. Zunächst wird ein IRT-Modell und die entsprechenden Scores auf Basis der originalen Antwortdaten geschätzt. Modell und Scores werden als Original-Modell bzw. Original-Scores bezeichnet, da sie auf den ursprünglichen (originalen) Antwortdaten beruhen.

2. Im zweiten Schritt werden die Korrelationen der Scores berechnet.

3. Auf Basis dieser Korrelationen werden im dritten Schritt normalverteilte Scores simuliert. Das Ergebnis sind folglich Scores, die dieselben Korrelationen wie die Original-Scores aufweisen.

4. Darauf folgend werden unter Verwendung des originalen Modells die den Scores entsprechenden Antwortdaten berechnet. Die simulierten Scores spiegeln somit die vollständig messfehlerfreien Merkmalsausprägungen des berechneten Antwortdatensatzes wider und werden deshalb als *wahr* bezeichnet. Auch die Antwortdaten werden zur einfacheren Zuordnung als *wahr* bezeichnet.

5. Als nächstes werden wie im ersten Schritt wieder IRT-Modell und Scores geschätzt, allerdings auf Basis der *wahren* Antwortdaten.

6. Zum Schluss kann die empirische Reliabilität als quadrierte Korrelation der *geschätzten* und der *wahren* Scores berechnet werden. Je besser das geschätzte IRT-Modell die Beziehung zwischen Antwortdaten und Scores beschreibt, umso ähnlicher sind sich die *wahren* und die geschätzten Scores, umso höher ist deren Korrelation und damit auch die Messgenauigkeit des Fragebogens.

3.6.2.2 Beurteilungsrichtlinien für Reliabilität

Bei der Interpretation von Reliabilitätskoeffizienten ist nach guter wissenschaftlicher Praxis der Kontext zu berücksichtigen (Groth-Marnat, 2003). Zum Beispiel werden in Leistungstests häufig höhere Reliabilitäten erzielt als in Persönlichkeitstests (Amelang & Schmidt-Atzert, 2006; Schermelleh-Engel & Werner, 2012). Laut Amelang und Schmidt-Atzert (2006) liegen interne Konsistenzen von Persönlichkeitstests häufig nur zwischen .60 und .80. Als Faustregel zur Interpretation von Cronbachs α wird in der Regel eine Untergrenze von .70 als akzeptabler Reliabilitätswert angesetzt, wobei zur klinischen Diagnostik Werte von $> .90$ wünschenswert sind (Groth-Marnat, 2003). Des Weiteren spielt die *Breite* des Konstrukts bei der Interpretation von Reliabilitätskoeffizienten eine wesentliche Rolle. Je breiter ein Konstrukt ist, also inhaltlich vielschichtiger, desto heterogener müssen die Items gestaltet sein, die dieses Konstrukt messen

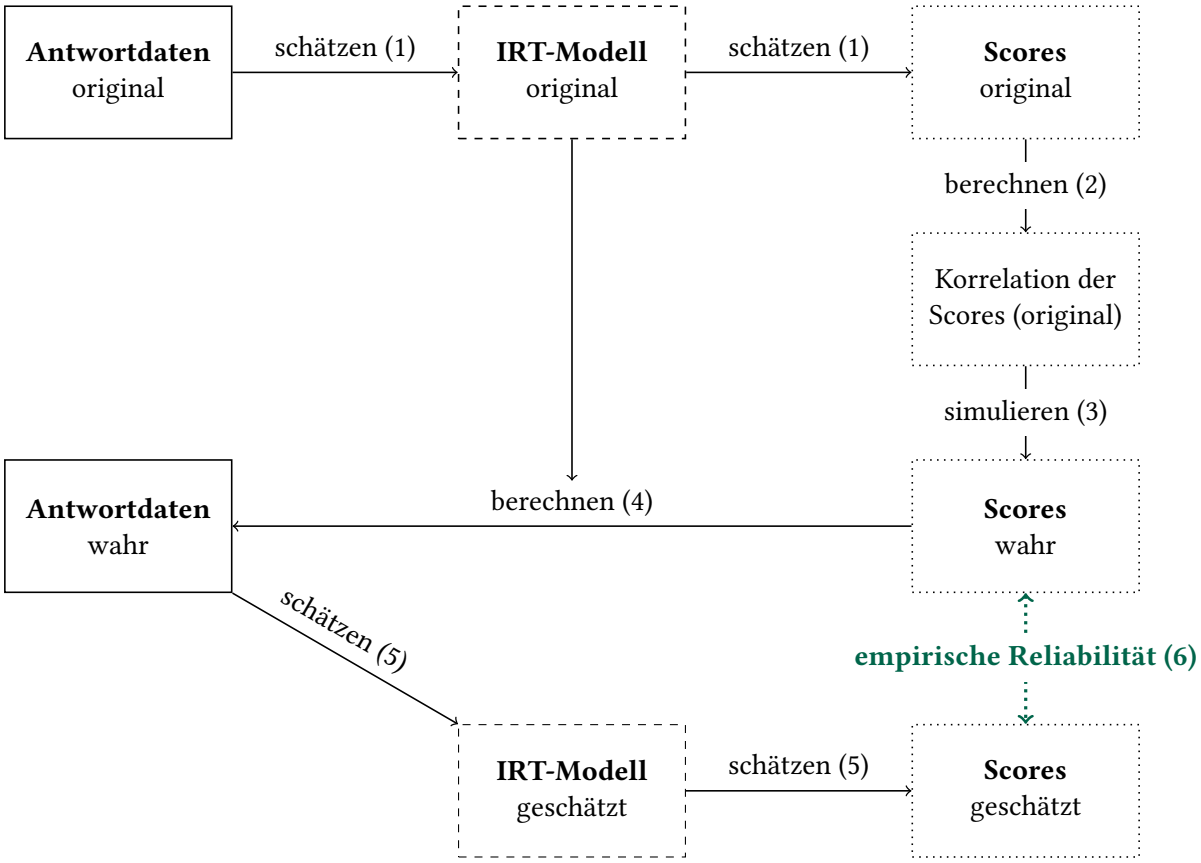

Abbildung 2. Vorgehensweise zur Berechnung der empirischen Reliabilität.

sollen. Das lässt wiederum niedrigere Reliabilitäten erwarten (Bühner, 2011; Lienert & Raatz, 1998; Schermelleh-Engel & Werner, 2012). Im Gegensatz dazu sind sich Items von homogenen (*engen*) Merkmalen inhaltlich ähnlicher und führen deshalb auch eher zu höheren Reliabilitäten. Da Persönlichkeitsmerkmale in der Regel deutlich breiter konzeptualisiert werden als Leistungsmaße, liegt hierin vermutlich auch der Grund, warum Persönlichkeitstests häufig niedrigere Reliabilitäten aufweisen als Leistungstests.

Darauf bezogen sei angemerkt, dass Wertesysteme als konzeptuell breite Konstrukte eingeschätzt werden können (Schwartz, 2003). Vor allem auch die Konzeptualisierung im MVSQ, in dem die Fragen zehn verschiedene Arbeitsaspekte abdecken, lässt diesen Rückschluss zu. In Übereinstimmung damit fallen beobachtete Reliabilitäten von ähnlichen Konstrukten relativ niedrig aus. Für die zehn Werte-Typen nach Schwartz berichten Schmidt et al. (2007) z.B. von einer durchschnittlichen Reliabilität in einer deutschsprachigen Stichprobe von $\alpha = .67$, wobei die Koeffizienten zwischen .48 und .79 schwankten. Auch in anderen Studien (Schwartz, 2005) werden vergleichbare Reliabilitäten der Werte-Typen berichtet, z.B. liegt die mittlere Reliabilität der Wertedimensionen des Schwartz Value Surveys in 13 Stichproben verschiedener Länder bei $\alpha = .70$ und in 14 Stichproben für den Portraits Value Questionnaire bei $\alpha = .68$. Auch bei Instrumenten, die auf anderen Wertetheorien basieren, werden Reliabilitätskoeffizienten in

vergleichbaren Größenordnungen berichtet: z.B. stellte Richins (2004) für die drei Wertedimensionen des *Material Values Scale* über 15 Stichproben durchschnittliche Cronbachs α von .72, .77 und .78 für die normale Version und von .67, .76 und .78 für eine verkürzte Version fest.

Zur Test-Retest-Reliabilität kann gesagt werden, dass diese häufig vom Zeitintervall zwischen den beiden Messungen abhängt und mit zunehmendem Intervall kleiner wird (Amelang & Schmidt-Atzert, 2006). Da sich die berichteten Zeitintervalle häufig unterscheiden, erscheint das Aufführen von Vergleichswerten an dieser Stelle wenig passend.

Zum Vergleich seien an dieser Stelle schließlich noch durchschnittliche Reliabilitätswerte anderer Konstrukte angeführt. In einer Meta-Analyse, die mehr als 1000 α-Koeffizienten der Big Five Persönlichkeitsdimensionen berücksichtigt, wurden für diese durchschnittliche Reliabilitäten von .75, .76, .71, .69 und .72 für Neurotizismus, Extraversion, Offenheit, Verträglichkeit und Gewissenhaftigkeit berichtet (Viswesvaran & Ones, 2000). Auch diese Werte passen ins Bild der zuvor formulierten Größenordnungen von Reliabilitäten in Persönlichkeitstests.

Zum Abschluss des Abschnitts zur Reliabilität bleibt anzumerken, dass die Reliabilität ein Gütekriterium ist, dass die Messgenauigkeit eines Instruments ungeachtet dessen beschreibt, ob der Test valide ist, also ungeachtet dessen, ob gemessen wird, was gemessen werden soll (Amelang & Schmidt-Atzert, 2006; Schermelleh-Engel & Werner, 2012). Obgleich wichtig ist sie kein hinreichendes Kriterium für die „praktische Brauchbarkeit eines Tests“ (Irtel, 1996, S. 33), sondern stellt lediglich eine Voraussetzung zur Untersuchung der Validität eines Tests dar (Amelang & Schmidt-Atzert, 2006).

3.6.3 Validität

Die Validität eines Tests ist „das komplexeste und am schwierigsten zu bestimmende Gütekriterium“ (Hartig et al., 2012, S.144). Dies liegt mitunter daran, dass die Validität eines Tests nur indirekt bestimmt werden kann und es im Prinzip unzählige Außenkriterien gibt, die herangezogen werden können, um auf die Validität eines Tests zu schließen (Bühner, 2011). Dabei ist die Validität eines Tests einfach die Eigenschaft eines Tests, die besagt, „ob der Test das auch wirklich misst, was er zu messen beansprucht“ (Bühner, 2011, S. 61). Bedingung dafür ist lediglich, dass erstens das zu messende Merkmal existiert und zweitens, dass das Merkmal zur Variation innerhalb von Items führt (Borsboom et al., 2004). Während sich das Konzept der Validität also relativ simpel gestaltet, ist es sehr schwierig, die Qualität der Validität allumfassend zu bestimmen. Üblicherweise wird die Validität in die zwei empirisch überprüfbaren Arten Konstruktvalidität und Kriteriumsvalidität zerlegt (Amelang & Schmidt-Atzert, 2006; Bühner, 2011; Cronbach & Meehl, 1955; Hartig et al., 2012). Zudem wird in der Theorie die Inhaltsvalidität als weitere wesentliche Validitätsart angegeben. Manche sehen diese als der Konstruktvalidität untergeordnete Validitätsart an (Haynes et al., 1995), verbreiteter ist hingegen die Ansicht, dass damit die Frage beantwortet wird, wie gut die Items eines Tests,

die zu messenden Merkmale inhaltlich präzise und genau abbilden (Bühner, 2011; Hartig et al., 2012; Lienert & Raatz, 1998). Die Items eines Tests sollten repräsentativ für alle denkbaren Items sein, die das jeweilige Konstrukt indizieren könnten, was nicht empirisch sondern nur anhand logischer Überlegungen beurteilt werden kann und vor allem bei der Testentwicklung zu berücksichtigen ist (Cronbach & Meehl, 1955). In anderen Worten bedeutet dies, dass die Untersuchung der Inhaltsvalidität der Generierung der Items durch Experten vorausgehen muss und deshalb bereits vor der Testentwicklung stattgefunden hat (T. Falter, persönliche Kommunikation, 28.06.2012). Für diese Arbeit erübrigt sich somit die Analyse der Inhaltsvalidität und der Fokus liegt deshalb auf den empirisch überprüfbaren Validitätsarten der Konstrukt- und Kriteriumsvalidität.

Die Konstruktvalidität wird herangezogen, um die Gültigkeit der Interpretation eines gemessenen Testwerts hinsichtlich der theoretischen Definition des Konstrukts zu beurteilen (Hartig et al., 2012). Dies kann vor allem beurteilt werden, wenn die Testwerte in Bezug zu anderen Konstrukten gesetzt werden (Amelang & Schmidt-Atzert, 2006). Dabei kann einerseits untersucht werden, ob ähnliche Konstrukte hohe Zusammenhänge aufweisen und andererseits, ob unähnliche Konstrukte geringe oder keine Zusammenhänge mit dem untersuchten Konstrukt haben. Ersteres wird gemeinhin als konvergente Validität, zweiteres als diskriminante (oder divergente) Validität bezeichnet (Amelang & Schmidt-Atzert, 2006). Beide Annäherungen an die Konstruktvalidität sind sinnvoll und sollten im Rahmen einer Untersuchung der Konstruktvalidität abgedeckt sein. Als dritter Teilaspekt der Konstruktvalidität kann die sogenannte faktorielle Validität untersucht werden (Bühner, 2011). Dabei wird die Dimensionsstruktur inferenzstatistisch, z.B. durch Faktorenanalysen auf ihre Hypothesenkonsistenz überprüft (Hartig et al., 2012). Kann die angenommene Dimensionalität bestätigt werden, ist dies jedoch kein hinreichender Beleg, sondern nur eine Voraussetzung für die Erfüllung der Konstruktvalidität, denn die Bestätigung einer Dimensionsstruktur erlaubt noch keine Aussagen inhaltlicher Natur.

Bei der Kriteriumsvalidität geht es, wie schon der Name andeutet, um den Zusammenhang eines gemessenen Merkmals mit Außenkriterien (Bühner, 2011). Lassen sich hypothesenkonforme Zusammenhänge mit einem Außenkriterium (z.B. Verhalten) feststellen, so erlauben diese einerseits Rückschlüsse auf die Validität der Messung und können andererseits auch als Indiz für die praktische Relevanz des Tests gewertet werden (Hartig et al., 2012). Die Überprüfung der Kriteriumsvalidität erfolgt stets empirisch und kann in folgende Unterkategorien aufgeteilt werden. Je nachdem, ob das Außenkriterium zeitlich nach oder gleichzeitig erhoben wird, spricht man von Vorhersagevalidität (auch prädiktive oder prognostische Validität genannt) oder Übereinstimmungsvalidität (auch konkurrente Validität) (Bühner, 2011). Des Weiteren gibt es auch die Form der retrospektiven Validität, bei der die Erhebungen des Außenkriteriums zeitlich vor der Messung des Merkmals erfolgt. Diese ist jedoch nur dann plausibel, wenn die Annahme sinnvoll ist, dass das Außenkriterium das Merkmal *nicht* beeinflusst. Diese Validitätsarten werden häufig mittels Korrelationen berechnet, wobei im Prinzip jedes inferenzstatistische

Verfahren angebracht sein kann (Groth-Marnat, 2003). Darüber hinaus gibt es die sogenannte inkrementelle Validität (Hartig et al., 2012). Sie bringt zum Ausdruck, in welchem Ausmaß der Test die Varianz des Kriteriums zusätzlich zu einem oder mehreren anderen gemessenen Variablen erklärt (Haynes & Lench, 2003). Zur Bestimmung der inkrementellen Validität bieten sich Mehrebenenanalysen an, in denen die in hierarchisch geschachtelten Regressionsmodellen zusätzlich erklärte Varianz als Maßstab für die inkrementelle Validität verwendet werden kann (Bühner, 2011; Haynes & Lench, 2003).

Ergebnisse aus den Untersuchungen zur Konstrukt- und Kriteriumsvalidität können des Weiteren den Konzepten der internen und externen Validität zugeordnet werden. Interne Validität bezieht sich auf die *interne* Gültigkeit von Ergebnissen und kann durch Experimentalstudien belegt werden. Denn darin können Störvariablen kontrolliert werden und dadurch Verhalten treffgenau und direkt auf das vom Test gemessene Konstrukt zurückgeführt werden (Eid et al., 2015). Die externe Validität ist ein Maß dessen, inwiefern die Testergebnisse generalisiert werden können, d.h. Schlussfolgerungen auf andere Orte, Personen, Situationen und Zeitpunkte übertragen werden können (Eid et al., 2015). Sie ist deutlich aufwändiger zu untersuchen als die interne Validität, da dafür Untersuchungen im Feld durchgeführt werden müssen, wobei idealerweise dieselben (oder zumindest ähnliche) Untersuchungen in Labor und Feld gleichermaßen durchgeführt werden sollten (Eid et al., 2015). Stimmen die Ergebnisse überein, so kann von einer hohen externen Validität ausgegangen werden.

3.7 Verwendete Stichproben

Da zur Schätzung der TIRT-Modelle eine Mindeststichprobengröße von mehreren Hundert Versuchspersonen notwendig ist, damit die Schätzungen konvergieren, können in dieser Arbeit nicht alle Stichproben unabhängig voneinander untersucht werden und hängen in gewissem Maße zusammen. Insgesamt fließen in diese Arbeit drei große Stichproben ein, aus denen teilweise Teilstichproben für einzelne Studien und Analysen verwendet wurden. An die Daten jeder dieser großen Stichproben wurden jeweils TIRT-Modelle angepasst.

Die Stichproben wurden entweder im Hochschulbereich, vor allem der Universität und Ostbayerischen Technischen Hochschule (OTH) Regensburg erhoben oder im Kontext von Beratungs- oder Forschungsprojekten mit größtenteils privatwirtschaftlichen Unternehmen von zertifizierten Nutzern des Fragebogens generiert. Die Ausgabe der Fragebögen erfolgte dabei bei allen Stichproben online und damit unter nicht kontrollierten Bedingungen. Die Verteilung der Zugangsdaten ermöglichte den Teilnehmern stets, den Zeitpunkt der Bearbeitung des Fragebogens selbst zu wählen. Dieses Vorgehen ist damit zu rechtfertigen, dass so einerseits durch die einfachere Durchführung die Anzahl der Probanden leichter erhöht werden konnte. Andererseits kann argumentiert werden, dass die den Teilnehmern dadurch zugewiesene Selbstbestimmung nicht nur negativ (im Sinne von fehlender Kontrolle über

die Durchführungsbedingungen), sondern auch förderlich auf die Motivation hinsichtlich der Testbearbeitung wirken kann. Aufgrund des für alle Testpersonen einheitlichen Web-Interfaces kann außerdem von weitestgehende gleichen und dadurch objektiven Durchführungsbedingungen ausgegangen werden. Darüber hinaus erhielten alle Personen die Möglichkeit, Feedback zu ihren Profilen zu erhalten. Zudem haben alle hier enthaltenen Teilnehmer zugestimmt, dass ihre Daten für wissenschaftliche Zwecke verwendet werden dürfen.

Tabelle 2 gibt einen Überblick der Stichproben inklusive des jeweiligen Umfangs, welche Version des Fragebogens von dieser Stichprobe bearbeitet wurde und welche Teilstichproben zu welchen eigenständigen Stichproben gehören. Im Fließtext folgen dann die Beschreibungen der Stichproben und der Verweis darauf, für welche Untersuchung die jeweilige Stichprobe verwendet wurde.

Tabelle 2. Überblick der Stichproben und Teilstichproben.

Stichprobe	I	II	III		
N (gesamt)	729	618	698		
Alter M (SD)	28.3 (10)	34.4 (10.3)	30.4 (10.4)		
n Weiblich (%)	433 (59.4)	255 (41.3)	357 (51.1)		
n studierend (%)	412 (56.5)	130 (21)	319 (45.7)		
Teilstichprobe	Ia	IIa	IIIa	IIIb	IIIc
n	39	31	104	166	62
Alter M (SD)	28.3 (10)	34.4 (10.3)	23.2 (2.6)	23.1 (2.7)	23.3 (2.7)
n Weiblich (%)	16 (41)	9 (29)	75 (72.1)	105 (63.3)	41 (66.1)
Personengruppe	B	B	S	S	S
Version	1	2	2	2	2

Anmerkung. Alter in Jahren; Personengruppe: B = Berufstätige, S = Studierende; Version bezieht sich auf die Version des Fragebogens.

Stichprobe I (N = 729) setzt sich aus $n = 316$ Berufstätigen und $n = 412$ Studierenden zusammen. Die Daten der Berufstätigen wurden im Rahmen von Beratungs- und Coaching-Projekten verschiedener Berater im Zeitraum zwischen Dezember 2011 und März 2013 erhoben. Der Studierendendatensatz ($n = 412$) wurde im selben Zeitraum entweder im Rahmen von Vorlesungen der Betriebswirtschaft (OTH Regensburg) oder als freiwilliger Test mit Feedback als Gegenleistung an OTH oder Universität Regensburg durch den Verfasser der Arbeit erhoben.

Von der gesamten Stichprobe waren $n = 433$ Personen weiblich (59.4%) und $n = 296$ Personen männlich (40.6%). Das Durchschnittsalter betrug 28.3 Jahre (SD = 10) bei einer Alterspanne von 17 bis 67. Der größte Teil gab als Nationalität (96%) und Muttersprache (95.1%) Deutsch an.

Die Studiengänge der Studierenden verteilten sich auf mehr als 26 verschiedene Studiengänge, wobei die einzigen Studiengänge mit nennenswerten Größen Betriebswirtschaft ($n = 299$) und Psychologie ($n = 42$) waren.

Zu den Zeitpunkten der Erhebung betrug die Berufserfahrung in der gesamten Stichprobe im Durchschnitt 5.7 Jahre (SD = 8.3), wobei die Berufstätigen durchschnittlich deutlich mehr Erfahrung aufwiesen (M = 10.8, SD = 10.1) als die Studierenden (M = 1.8, SD = 3). In Stufen verteilte sich die Berufserfahrung von beiden Gruppen zusammen wie folgt: 23.2% haben zwischen drei Monaten[4] und einem Jahr in einem Vollzeitjob gearbeitet, 34.3% zwei bis fünf Jahre, 12.9% sechs bis 15 Jahre und 12.9% mehr als 15 Jahre. 16.7% gaben an, über keine Berufserfahrung zu verfügen.

Die Aufgabenbereiche der Berufstätigen verteilen sich auf verschiedene Abteilungen und Funktionen. Nennenswert sind hier die drei am häufigsten gemachten Angaben: Unternehmensführung ($n = 27$), Vertrieb ($n = 21$) sowie Forschung & Entwicklung ($n = 16$). Ferner ist bekannt, welchen Managementebenen die Berufstätigen angehörten: 17 Angestellte (5.4%) gaben an, zur Geschäftsführung zu gehören, 45 (14.2%) befanden sich in der Organisationshierarchie eine Ebene unter der Geschäftsführung, 53 (16.8%) zwei Ebenen und 117 (37%) drei und mehr Ebenen unter der Geschäftsführung. 67 Personen (21.2%) gaben an, Selbstständig zu sein und 16 (5.1%) machten keine Angabe.

Daten dieser Stichprobe wurden zunächst in Kapitel 5 verwendet, um eine Itemanalyse durchzuführen und darauf basierend Empfehlungen zur Überarbeitung des Fragebogens abzugeben. Ferner wurden Antwortdaten dieser Stichprobe in Kapitel 8 im Rahmen des Versionsvergleichs zwei TIRT-Modelle angepasst sowie die entsprechenden empirische Reliabilitäten geschätzt.

Stichprobe Ia ist eine Stichprobe von $n = 39$ Auszubildenden und dual Studierenden, deren Daten zur Analyse der divergenten Validität herangezogen wurden. 16 Personen waren davon weiblich (41%). Das Alter lag im Durchschnitt bei 21.8 Jahren ($SD = 2.1$), die Altersspanne betrug 18 bis 29 Jahre und die durchschnittliche Berufserfahrung belief sich auf 2.2 Jahre ($SD = 0.9$). Alle Teilnehmer erhielten die Möglichkeit, Feedback zu den Ergebnissen der Bearbeiteten Fragebögen zu erhalten.

[4]Drei Monate ist der kleinste auswählbare Wert, der größer als Null ist.

Stichprobe II (N = 618) besteht aus $n = 488$ Berufstätigen und $n = 130$ Studierenden, die auch im Rahmen von Kundenprojekten verschiedener Anwender des Fragebogens sowie an Universität und OTH Regensburg erhoben wurden.

Insgesamt waren bei dieser Stichprobe $n = 255$ Personen weiblich (41.3%) und $n = 363$ Personen männlich (58.7%). Im Vergleich zu Stichprobe I hat diese Stichprobe insgesamt ein etwas ausgeglicheneres Geschlechterverhältnis, wobei nun der männliche Anteil proportional höher ist. Im Durchschnitt waren die Probanden 34.4 Jahre alt (SD = 10.3) und die Alterspanne ging von 16 bis 63. Erneut gab eine sehr große Mehrheit Deutsch sowohl als Nationalität (93.7%) als auch als Muttersprache (92.2%) an.

Die Verteilung der Studiengänge war mit mehr als 20 verschiedenen Studiengängen etwas weniger breit gefächert als in Stichprobe I, wobei auch diesmal Betriebswirtschaft mit $n = 95$ der zahlenmäßig größte Studiengang war. Kein weiterer Studiengang wie mehr als 20 Studierende auf.

Die Berufserfahrung der gesamten Stichprobe belief sich im Durchschnitt auf 10.6 Jahre (SD = 9.6), wobei die Gruppe der Berufstätigen erneut deutlich mehr Erfahrung aufwies (M = 13, SD = 9.4) als die Gruppe der Studierenden (M = 1.8, SD = 2). Auch hier wieder die Abstufungen der Berufserfahrung von beiden Gruppen zusammen: 14.3% zwischen drei Monate und einem Jahr, 19.9% zwei bis fünf Jahre, 32.9% sechs bis 15 Jahre und 27.8% mehr als 15 Jahre. 5.1% verfügten über keine Berufserfahrung.

Da die Anzahl der Berufstätigen in dieser Stichprobe deutlich höher ist als zuvor, gibt es auch mehrere nennenswerte Aufgabenbereiche. In abnehmender Reihenfolge waren die Aufgabenbereiche der Berufstätigen mit $n > 20$: Vertrieb ($n = 74$), Forschung und Entwicklung ($n = 61$), IT ($n = 49$), Produktion ($n = 34$), Personal ($n = 33$) und Marketing ($n = 22$).

Die Zugehörigkeiten zu den Hierarchieebenen verteilten sich wie folgt: 15 Angestellte (3.1%) gehörten zur Geschäftsführung, 57 (11.7%) eine Ebene unter der Geschäftsführung, 110 (22.5%) zwei Ebenen und 247 (50.6%) drei und mehr Ebenen unter der Geschäftsführung. 40 Personen (8.2%) waren Selbstständig und 19 (3.9%) machten keine Angabe.

Diese Stichprobe wurde für mehrere Analysen herangezogen. Zunächst wurden damit die Tuning-Parameter bestimmt und die ersten TIRT-Modelle geschätzt (Kapitel 6), die dann mit den TIRT-Modellen der ersten Fragebogenversion verglichen wurden (Kapitel 8). Als wichtigste Untersuchung mit dieser Stichprobe kann die Berechnung der empirischen Reliabilitäten und die Bestimmung dessen Schätzgenauigkeit gesehen werden (Kapitel 7). Außerdem wurde auch die faktorielle Validität anhand der Daten dieser Stichprobe untersucht (Kapitel 9.1). Eine Teilstichprobe von $n = 31$ Vertriebsmitarbeitern wurde außerdem dazu verwendet, um die inkrementelle Validität der MVSQ-Scores zu untersuchen (Kapitel 10.3).

Teilstichprobe IIa beinhaltet die Daten einer Stichprobe von N = 31 Vertriebsmitarbeitern, wovon 9 (29%) weiblich waren. Männer waren mit 71% in der Geschlechterverteilung demnach

deutlich überrepräsentiert. Das Durchschnittsalter betrug 33.1 Jahre ($SD = 6.9$) und die mittlere Berufserfahrung lag bei 7.7 Jahre ($SD = 6$). Diese Stichprobe wurde zur Untersuchung der prädiktiven und inkrementellen Validität verwendet.

Stichprobe III (N = 698) setzt sich aus $n = 379$ Berufstätigen und $n = 319$ Studierenden zusammen. Die Berufstätigen wurden wieder im Rahmen mehrerer Kundenprojekte verschiedener Anwender des Fragebogens und die Studierenden an Universität und OTH Regensburg erhoben. Diese Stichprobe wurde nun hauptsächlich für Untersuchungen der Validität herangezogen.

Insgesamt verteilten sich die Geschlechter wie folgt: $n = 357$ Personen waren weiblich (51.1%) und $n = 341$ Personen waren männlich (48.9%). Das Verhältnis der Geschlechter war somit ausgeglichen. Das Durchschnittsalter betrug 30.4 Jahre (SD = 10.4) bei einer Spanne von 18 bis 67. Erneut gab eine sehr große Mehrheit Deutsch sowohl als Nationalität (94.7%) als auch als Muttersprache (93.3%) an.

Die Verteilung der Studiengänge war mit mehr als 31 verschiedenen Studiengängen breiter als in den vorherigen Stichproben. Wieder war Betriebswirtschaft mit $n = 162$ der am häufigsten vertretene Studiengang. Daneben war Informatik ($n = 23$) der einzige Studiengang mit mehr als 20 Studierenden.

Die Berufserfahrung lag im Mittel bei 7.5 Jahren (SD = 9.3) und wenig überraschend hatten die Berufstätigen wieder deutlich mehr Erfahrung (M = 12.3, SD = 10.3) als die Studierenden (M = 1.8, SD = 2). Die Stufen der Berufserfahrung von beiden Gruppen zusammen verteilten sich wie folgt: 24.1% hatten zwischen drei Monate und einem Jahr, 25.1% zwei bis fünf Jahre, 18.9% sechs bis 15 Jahre und 19.1% mehr als 15 Jahre Erfahrung. 12.7% verfügten über keine Berufserfahrung.

Die Aufgabenbereiche der Berufstätigen in dieser Stichprobe mit $n > 20$ verteilten sich wie folgt: Vertrieb ($n = 66$), Forschung und Entwicklung ($n = 61$), Personal ($n = 37$) und Logistik ($n = 23$). Die Zugehörigkeiten zu den Hierarchieebenen gestaltete sich so, dass 18 Angestellte (4.7%) gehörten zur Geschäftsführung, 44 (11.6%) eine Ebene unter der Geschäftsführung, 58 (15.3%) zwei Ebenen und 237 (62.5%) drei und mehr Ebenen unter der Geschäftsführung arbeiteten. 18 Personen (4.7%) waren Selbstständig und bei 4 Personen (1.1%) lag keine Angabe vor.

Diese Stichprobe wurde gemeinsam mit Stichprobe II dazu verwendet, in Kapitel 10.1 die konkurrente Validität zu untersuchen. Mehrere Teilstichproben daraus wurden zur Untersuchung der konvergenten (Kapitel 9.2), divergenten (Kapitel 9.3) und prädiktiven Validität (Kapitel 10.2) herangezogen. Die entsprechenden Teilstichproben werden im folgenden präsentiert:

Teilstichprobe IIIa wurde im Rahmen einer Studie erhoben, an der ursprünglich 117 Studierende von Universität und OTH Regensburg teilnahmen. Aufgrund unvollständiger Bearbeitung wurden 13 Personen von der Analyse ausgeschlossen. Die verbleibende Stichprobe setzte sich

somit aus 104 Studierenden (75 Frauen, 72.1%) zusammen, deren Durchschnittsalter 23.2 Jahre (SD = 2.6, Spanne von 18 bis 33) betrug. Der einzige Studiengang mit mehr als 20 Studierenden war Betriebswirtschaft ($n = 45, 43.3\%$). Die durchschnittliche Berufserfahrung belief sich auf 1.6 Jahre (SD = 1.9). Als Gegenleistung für die Teilnahme an der Untersuchung erhielten die Teilnehmer Feedback den Ergebnissen der involvierten Fragebögen.

Teilstichprobe IIIb ist ebenso eine rein studentische Stichprobe ($n = 166$), die an Universität und OTH Regensburg akquiriert wurde. Sie wurde verwendet, um die divergente und kriteriumsbezogene Validität zu analysieren. Es waren 105 Personen weiblich (60%) und das Durchschnittsalter betrug 23.1 Jahre (SD = 2.7 bei einer Altersspanne von 18 bis 33 Jahren. 98.8% gaben Deutsch als ihre Nationalität und 97.6% als Muttersprache an. Auch in dieser Teilstichprobe war der am häufigsten vertretene Studiengang Betriebswirtschaft mit $n = 74$ Studierenden (44.6%). Die mittlere Berufserfahrung betrug 1.6 Jahre (SD = 2). Auch hier erhielten die Teilnehmer als Gegenleistung für die Teilnahme an der Untersuchung Feedback zu ihren Ergebnissen des MVSQ.

Teilstichprobe IIIc wurde aus der Teilstichprobe IIIb rekrutiert und dazu herangezogen wurde, die Test-Retest-Reliabilität zu bestimmen. Im Nachgang an die experimentelle Untersuchung, der Teilstichprobe IIIb zugrunde lag, wurden die Teilnehmer dieser Untersuchung nach zehn Wochen gebeten, den MVSQ erneut zu bearbeiten. Von den 166 eingeladenen haben $n = 62$ Personen den MVSQ noch ein mal bearbeitet. Zwei Drittel dieser Personen waren weiblich ($n = 41$, 66.1%) und entsprechend ein Drittel männlich ($n = 21$, 33.9). Das Durchschnittsalter betrug 23.3 ($SD = 2.7$) bei einer Spanne von 19 bis 33 Jahre. Alle gaben Deutsch als Nationalität und Muttersprache an und die Berufserfahrung lag zwischen 0 und 9 Jahren, ($M = 1.5$, $SD = 2.1$). Häufigster Studiengang war Betriebswirtschaft mit 25 Studierenden (40.3%).

3.8 Zusammenfassung

In diesem Kapitel wurde der MVSQ vorgestellt und insbesondere auf sein Format hin analysiert. Dabei wurden die Vorteile und Nachteile des ipsativen Formats dargestellt und die daraus resultierende Notwendigkeit, ein TIRT-Modell zu schätzen, wenn Reliabilität und Validität auf wissenschaftlich korrekte Art untersucht werden wollen. Darüber hinaus wurden die Hauptgütekriterien der psychometrischen Güte vorgestellt, die in dieser Arbeit untersucht werden. Zum Schluss erfolgt der Überblick sowie die gesammelte detaillierte Beschreibung aller in dieser Arbeit verwendeten Stichproben.

Kapitel 4

Objektivität

Als erste Analyse des MVSQ wird die Objektivität des Instruments anhand der drei Formen der Durchführungs-, Auswertungs- und Interpretationsobjektivität untersucht. Damit dies auf sinnvolle Art und Weise vollzogen werden kann, werden zuerst die Anforderungen an die drei Objektivitätsarten dargelegt, um danach zu prüfen, inwiefern diese vom MVSQ erfüllt werden.

4.1 Methode

Die Durchführungsobjektivität ist dann hoch, wenn die Durchführungsbedingungen für alle Testpersonen gleich und frei von subjektiven Einflüssen einer Testleitung sind. Konkret können in der einschlägigen Literatur (Amelang & Schmidt-Atzert, 2006; Bortz & Döring, 2006; Bühner, 2011; Lienert & Raatz, 1998; Schermelleh-Engel & Werner, 2012) zwei wesentliche Kriterien ausgemacht werden:

- Erstens sollte die Durchführung möglichst unabhängig von der Testleitung sein, d.h. die Testleitung sollte keinen Einfluss auf die Bearbeitung der Testperson haben. Dies kann dadurch erreicht werden, indem die sozialen Interaktionen zwischen der Testleitung und Testperson so gering wie möglich gehalten werden und indem ein hohes Maß an Standardisierung eingerichtet wird (z.B. durch schriftliche Anleitungen und Regeln für die Bearbeitung des Tests).
- Ein weiterer wesentlicher Aspekt betrifft die Testsituation. Maximale Durchführungsobjektivität ist dann gewährleistet, wenn die Testsituation für alle Testpersonen gleich ist. Dazu muss diese unter kontrollierten Bedingungen, d.h. mit denselben Materialien und ohne erhebliche Einflüsse durch Außenkriterien wie z.B. Lärm durchgeführt werden.

Die quantitative Bestimmung der Objektivität ist schwierig. Man könnte z.B. versuchen, den selben Test von derselben Person bei unterschiedlichen Testleitern durchführen zu lassen und die Ergebnisse zu interkorrelieren. Allerdings wären hier Erinnerungseffekte nur schwer

zu kontrollieren und auch würde die Reliabilität die Ergebnisse beeinflussen, da diese nicht davon getrennt werden kann.

Die Auswertungsobjektivität ist gegeben, wenn die Auswertung unabhängig vom Testauswerter erfolgt. An dieser Stelle sind vor allem zwei Fälle zu erwähnen, die negative Auswirkungen auf die Auswertungsobjektivität haben können. Auch diese beruhen auf den oben genannten Literaturquellen.

- Wenn die Auswertung per Hand erfolgt, d.h. der Testleitung Fehler unterlaufen können.
- Wenn die Auswertung eine subjektive Einschätzung des Testleiters erfordert. Dies ist bei projektiven Tests der Fall und kann dann wiederum durch Auswertungsregeln kontrolliert werden.

Quantitativ kann die Auswertungsobjektivität dadurch bestimmt werden, wenn die Antworten derselben Testperson mehreren Auswertern vorgelegt werden. Der Grad der Übereinstimmung der Auswertungsergebnisse stellt dann ein Maß der Auswertungsobjektivität dar. Dies könnte allerdings auch unter den Aspekt der Messgenauigkeit fallen (Amelang & Schmidt-Atzert, 2006).

Die Interpretationsobjektivität wird erreicht, wenn dieselben Auswertungsergebnisse unabhängig vom Testinterpreten zu den gleichen Schlussfolgerungen führen. Die Testinterpretation muss also unabhängig von der interpretierenden Person gleich lauten. Dazu können nach den oben genannten Quellen folgende Maßnahmen ergriffen werden:

- Standardisierung der Interpretation durch Interpretationsleitfäden und Normierung.
- Ausbildung der Testinterpreten, sodass diese zu denselben Ergebnisinterpretation gelangen.

Anhand dieser Kriterien kann die Interpretationsobjektivität argumentativ abgeschätzt werden. Eine quantitative Bemessung könnte so überprüft werden, dass man unterschiedlichen Testinterpreten dieselben Ergebnisse inklusive Interpretationsrichtlinien bzw. nach Ausbildung vorlegt und dann prüft, ob diese zu denselben Ergebnissen kommen. Diese Methode weist jedoch zwei Probleme auf. Erstens kann es schwer sein, eine ausreichend große Zahl an ausgebildeten Testinterpreten zu akquirieren und zweitens würde das Ergebnis wesentlich vom Komplexitätsgrad der Interpretationsergebnisse abhängen. Einfache Richtlinien würden zu höheren Übereinstimmungen führen als komplexe, was wiederum bedeutet, dass es angebracht wäre die Qualität der Interpretationsrichtlinien zu überprüfen. Eine anfängliche logisch-argumentative Überprüfung der Interpretationsobjektivität wird demnach in der Regel bevorzugt. So auch hier.

4.2 Ergebnisse

Zur Durchführungsobjektivität

Die Testdurchführung kann beim MVSQ als weitestgehend unabhängig von der Testleitung angesehen werden, denn in der Regel kommunizieren Testleiter und Testperson vor der Testdurchführung nur via E-Mail. Für die Durchführung gibt es zudem eine standardisierte und relativ ausführliche Anleitung, die für alle Personen gleich ist. Neben einer schriftlichen Einweisung steht den Teilnehmern ein professionelles Screen-Cast-Video zur Verfügung, das zeigt, wie die Antwortprozedur des Fragebogens funktioniert. Falls dieses Video einen Einfluss haben sollte, wäre dies ein systematischer Einfluss, der für alle Personen gleich ist. Der Einfluss einer Testleitung kann insgesamt als äußerst gering angesehen werden.

Die Testsituation ist insofern standardisiert, dass die Bedienungsoberfläche für alle Teilnehmer exakt gleich ist. Benutzt man die traditionelle Ausdrucksweise, dann kann gesagt werden, dass die „Materialien“ standardisiert sind. Nicht kontrolliert sind jedoch die räumlichen und zeitlichen Bedingungen, in denen die Testpersonen den Fragebogen bearbeiten. Da der Fokus bei der Bearbeitung jedoch auf den Bildschirm gerichtet ist und in der Anleitung der Hinweis erfolgt, den Fragebogen in einer ruhigen Umgebung und möglichst ohne Störeinflüsse zu absolvieren, kann auch diese Bedingung als weitestgehend erfüllt angesehen werden. Die Durchführungsobjektivität kann beim MVSQ somit als sehr hoch eingestuft werden.

Zur Auswertungsobjektivität

Die Auswertung erfolgt beim MVSQ voll automatisch und damit weder per Hand noch ist die subjektive Einschätzung eines Testleiters erforderlich. Damit ist die Auswertungsobjektivität vollständig gegeben und eine quantitative Bestimmung damit nicht erforderlich.

Zur Interpretationsobjektivität

Im Vergleich zur Durchführung und Auswertung ist beim MVSQ (wie bei vielen Persönlichkeitsfragebögen) der Einfluss der Testleitung bei der Interpretation am größten. Zur Standardisierung der Interpretation gibt es einen ca. 20 Seiten umfassenden teilindividualisierten Ergebnisbericht, der für jedes Profil erstellt wird und zahlreiche Domänen der Interpretation abdeckt. Die Erstellung der Ergebnisberichte ist für alle Testpersonen gleich, birgt somit ein hohes Maß an Standardisierung und kann als Argument pro Interpretationsobjektivität aufgeführt. Als weiteres Argument muss genannt werden, dass die Anwender des MVSQ eine mehrtägige Ausbildung inklusive Supervision absolvieren müssen (G. Singer, persönliche Kommunikation, 19.05.2012). Auch dies ist im Sinne einer hohen Interpretationsobjektivität zu werten. Für beide Fälle gilt jedoch, dass nicht überprüft wird, inwiefern sich die Testleiter an die vorgegebenen

Richtlinien halten. Kritisch ist ferner anzumerken, dass es kein Testmanual gibt und keine veröffentlichten Informationen zur Entwicklung des Instruments verfügbar waren.

4.3 Diskussion

In diesem kurzen Kapitel wurden die Teilaspekte der Objektivität untersucht. Insgesamt können diese als weitestgehend gegeben angesehen werden, was bedeutet, dass der MVSQ inklusive Testunterlagen, Testdarbietung, Auswertung und Interpretation so exakt festgelegt ist, dass er von verschiedenen Testleitern an unterschiedlichen Orten und Zeitpunkten administriert werden kann und dabei zum selben Ergebnis kommen würde (Moosbrugger & Kelava, 2012). Damit ist des Weiteren eine wesentliche Voraussetzung zur Untersuchung der Messgenauigkeit des MVSQ erfüllt (Schermelleh-Engel & Werner, 2012).

Als einschränkend kann aufgeführt werden, dass der Test ausschließlich online administriert werden kann, was dazu verleitet, dass er ohne Aufsicht einer Testleitung bearbeitet wird. Dies wiederum hat zur Folge, dass in den meisten Fällen keine Kontrolle möglicher Störfaktoren wie z.B. Lärm oder Ablenkung erfolgt, andererseits aber Versuchsleitereffekte bei der Testdurchführung ausschließt. Im Gegenteil, die Durchführungsbedingungen sind durch die Unveränderlichkeit des Online-Interfaces vermutlich sogar objektiver, als wenn es eine Testleitung gäbe.

Eine interessante Fragestellung könnte sein, ob die Drag & Drop-Umsetzung im MVSQ einen Einfluss auf die Ergebnisse im Vergleich zu einer Papier-und-Bleistift-Lösung hätte. Insbesondere bei Ranking-Verfahren könnten sich Unterschiede zeigen, da die Umsetzung von Ranking-Verfahren mit Papier und Bleistift aufwändiger durchzuführen und auszuwerten wäre. Ungeachtet des Antwortformats kann allerdings gesagt werden, dass in einigen Studien geringe bis keine Unterschiede zwischen Online- und Offline-Tests gefunden wurden (Ihme et al., 2009; Mead & Drasgow, 1993; Meade et al., 2007) und die Fragestellung deshalb als den Untersuchungsgebieten von Reliabilität und Validität nachrangig behandelt werden kann.

Kapitel 5

Deskriptivstatistische Evaluation der Items

Standardmäßig wird bei der Fragebogenentwicklung als erster Schritt nach der Entwicklungsphase eines Instruments eine deskriptivstatistische Evaluation – auch Itemanalyse genannt – durchgeführt (Bühner, 2011; Kelava & Moosbrugger, 2012). Die Itemanalyse dient der Beurteilung der Qualität einzelner Items, die vor allem an den Kennwerten der Itemschwierigkeit und Trennschärfe festgemacht wird (Bortz & Döring, 2006; Kelava & Moosbrugger, 2012). Darüber hinaus können Itemvarianzen und Testwertverteilungen herangezogen werden, um Rückschlüsse auf die Qualität des Tests zu ziehen und Ansatzpunkte zur Revision einzelner Items abzuleiten (Bühner, 2011; Kelava & Moosbrugger, 2012). Ziel dieser Analyse war die Durchführung einer Itemanalyse und dadurch Identifizierung ungeeigneter Items, die dann von den Fragebogenentwicklern überarbeitet werden können.

5.1 Methode

Diese Itemanalyse wird auf Basis der Daten aus Stichprobe I (siehe Kapitel 3.7) durchgeführt und beinhaltet die Berechnung, den Bericht und die Beurteilung der Itemschwierigkeiten, Trennschärfen, Itemvarianzen und Testwertverteilungen. Folgende Auflistung stellt kurz die verwendeten Kennwerten vor:[1]

- Die Itemschwierigkeit P_i ist ein Maß dafür, wie schwierig oder leicht ein Item „gelöst“ werden kann (Kelava & Moosbrugger, 2012). Bei Rating-Skalen bedeutet dies, je schwieriger ein Item ist, desto niedriger ist die durchschnittlich erreichte Punktzahl. In Ranking-Skalen erreichen schwierigere Items folglich durchschnittlich niedrigere Ränge.
- Die Trennschärfe r_{it} eines Items gibt an, „wie groß der korrelative Zusammenhang zwischen den Itemwerten [...] und den Testwerten der Probanden ist“ (Kelava & Moosbrugger,

[1] Die Notation ist Kelava und Moosbrugger (2012) entnommen, wobei i das Item indiziert.

2012, S.84). Sie ist also ein Zusammenhangsmaß, das zum Ausdruck bringt, inwiefern ein Item mit dem Summen- (oder Mittel-) Wert der übrigen Items desselben latenten Konstrukts übereinstimmt.

- Die Itemvarianz V_i kann als Index für die „Differenzierungsfähigkeit eines Items“ (Kelava & Moosbrugger, 2012, S. 83) gesehen werden. Je größer die Varianz der Bewertungen eines Items ist, umso größere Unterschiede können anhand dieses Items zwischen Probanden gemessen werden.
- Die Testwertverteilungen können anhand deskriptiver Maße wie Mittelwert, Median, Modalwert, Varianz, Spannweite, Schiefe und Kurtosis beurteilt werden, wobei die Normalverteilungsannahme gilt (Kelava & Moosbrugger, 2012; Lienert & Raatz, 1998).
- In FC-Fragebögen können des Weiteren für jeden Paarvergleich die relativen Häufigkeiten der Bevorzugung berechnet werden. Eine solche relative Häufigkeit drückt aus, wie häufig ein Item höher gerankt wird als ein zweites Item. Es ist also ein Maß der Schwierigkeit, das sich auf zwei Items bezieht. Dieser Kennwert kann deshalb als *paarweise Itemschwierigkeit* (P_{pi}) bezeichnet werden.

Da es sich beim MVSQ um einen FC-Fragebogen handelt, sind die daraus generierten Daten ipsativ. Auf die Durchführung einer deskriptivstatistische Evaluation hat die Ipsativität eine Reihe von Auswirkungen. Im Folgenden werden Auswirkungen von Ipsativität bei der Berechnung der einzelnen Kennwerte erläutert. Wenn im Folgenden von *klassischen* Itemkennwerten die Rede ist, dann sind damit die ohne Einfluss von Ipsativität berechneten Koeffizienten gemeint, die z.B. in Rating-Fragebögen berechnet werden können.

Auswirkungen von Ipsativität auf Itemschwierigkeit

Der Schwierigkeitsindex P_i eines Items berechnet sich klassischerweise als „Quotient aus der bei diesem Item tatsächlich erreichten Punktsumme aller Probanden und der maximal erreichbaren Punktsumme“ (Kelava & Moosbrugger, 2012, S.76). Demzufolge ist der Schwierigkeitsindex bei leichten Items hoch und bei schwierigen Items niedrig. Von diesem Quotienten ist der Zähler von der Ipsativität des FC-Formats betroffen. Genauer gesagt hängt der Zähler – also die tatsächlich erreichte Punktsumme – von allen Items desselben Blocks ab und enthält somit auch Informationen aller Items desselben Blocks. Der „klassisch“ berechnete Schwierigkeitsindex eines Items beinhaltet folglich Informationen relativ zu den anderen Items desselben Blocks und erlaubt ergo vor allem darüber Aussagen, welches Item im Vergleich mit *allen* anderen Items des Blocks schwieriger zu beantworten ist. In absoluten Werten ist eine Tendenz zur Mitte hin zu erwarten, da der durchschnittliche Itemwert bei sieben Items pro Block bei 3 liegt und die durchschnittliche Itemschwierigkeit pro Block dadurch auf 0.5 (bei maximal 6 erreichbaren Punkten für Rang 1) festgesetzt ist. Als Folge daraus dürfte die von Bühner (2011, S. 81)

vorgeschlagene Interpretationsrichtlinie, dass Items zwischen $.20$ und $.80$ als „mittel“ eingestuft werden, im vorliegenden Fall zu liberal sein. Um der Tendenz zur Mitte hin Rechnung zu tragen, empfiehlt es sich in der folgenden Analyse die Beurteilungsrichtlinien zu verschärfen. Eben der Umstand, dass *alle* Items eines Blocks zum Schwierigkeitsindex eines Items beitragen, könnte zu fehlerhaften Rückschlüssen führen. Dies wird deutlich, wenn man sich vor Augen führt, dass sich der Rang eines Items auch als Summe der „gewonnenen“ Paarvergleiche des involvierten Items mit den übrigen Items berechnen lässt. Jeder Paarvergleich wiederum kann als die relative Häufigkeit ausgedrückt werden, mit der ein Item über ein zweites Item gerankt wird. Im MVSQ, der sieben Items pro Block und sechs Paarvergleiche pro Item hat, könnte sich so die mittlere Itemschwierigkeit eines Items aus drei sehr niedrigen und drei sehr hohen paarweisen Vergleichen, d.h. relativen Häufigkeiten zusammensetzen. Diese mittlere Itemschwierigkeit würde nur widerspiegeln, wie sich die Schwierigkeit des involvierten Items zu *allen* Items des Blocks verhält. Betrachtet man hingegen die Paarvergleiche, lässt sich ein differenzierteres Urteil über die Schwierigkeit eines Items ableiten und dadurch einzelne Items punktgenauer angepasst werden. Konsequenterweise werden in der folgenden Untersuchung die relativen Häufigkeiten als *paarweise Itemschwierigkeiten* (P_{pi}) als zusätzliches Maß der Schwierigkeit eines Items aufgenommen. Bei der Umformulierung eines einzelnen Items gilt gleichwohl, dass das Anpassen dieses Items in einer überarbeiteten Version immer auch Auswirkungen auf die paarweisen Itemschwierigkeiten im Vergleich mit dem betreffenden Item im selben Block haben kann.

Ipsativität und Itemvarianzen

„Die Itemvarianz informiert über die Differenzierungsfähigkeit eines Items“ (Kelava & Moosbrugger, 2012, S. 83). Auch hierauf haben die Abhängigkeiten der Items einen erheblichen Einfluss, denn das FC-Format bedingt, dass niedrige Bewertungen eines Items mit höheren Bewertungen eines anderen Items einhergehen. Items differenzieren deshalb erzwungenermaßen zwischen Merkmalen. Als Folge sind extreme Werte der Varianzen nur eingeschränkt möglich und die Varianzen können vor allem relativ zueinander sinnvoll beurteilt werden. Es gilt dennoch, dass höhere Itemvarianzen wünschenswert sind, da größere Itemvarianzen bedeuten, dass diese Items besser zwischen unterschiedlichen Merkmalsausprägungen differenzieren. Obgleich Itemvarianzen seltener zur Itemselektion oder -beurteilung herangezogen werden, da Teile ihrer Ausprägungen bereits durch die Itemschwierigkeit abgedeckt sind (Kelava & Moosbrugger, 2012), werden sie in dieser Analyse einbezogen, da aufgrund der Ipsativität der Daten jede weitere Informationsquelle genutzt werden sollte, um ein möglichst differenziertes Bild der Kennwerte zu erhalten.

Auswirkungen von Ipsativität auf Trennschärfe

Die Trennschärfe als Maß beziffert den korrelativen Zusammenhang des Itemwerts mit dem Testwert, d.h. allen übrigen Items desselben Konstrukts. Für den MVSQ bedeutet dies, dass Trennschärfen Informationen des *gesamten* Tests enthalten. Denn der Rang eines Items kommt im Zusammenspiel mit allen Items desselben Blocks zustande, hängt also von allen Items desselben Blocks ab. Wird nun das betreffende Item in Bezug mit den übrigen Items desselben Konstrukts gesetzt, die ebenfalls wieder von allen Items in den jeweiligen Blöcken abhängen, dann hängt die Trennschärfe *eines* Items direkt und indirekt von *allen* Items des Tests ab. Die Interpretation der Trennschärfen auf Itemebene wird dadurch unmöglich. Denkbar ist lediglich, dass Tendenzen über Blöcke oder Konstrukte eine gewisse Aussagekraft haben können. Sind z.B. Trennschärfen nur eines Blocks im Verhältnis niedrig, könnte dies ein Indiz dafür sein, dass ebendieser Block von besonders minderer Qualität ist. Obgleich dieser Rückschluss nur dann gelten würde, wenn die Items in den übrigen Blöcken ähnlich homogen (Kelava & Moosbrugger, 2012) und von höherer Qualität sind, was wiederum im forced-choice-Format nicht prüfbar ist. Aufgrund dieser Unwägbarkeiten empfiehlt es sich, die Bedeutung der Trennschärfen, die üblicherweise als zentrales Kriterium in Itemanalysen gelten (Lienert & Raatz, 1998) herabzustufen und hinter die weniger Items betreffenden Itemschwierigkeiten und paarweisen Itemschwierigkeiten zu stellen.

Ipsative Testwertverteilungen

Testwerte werden im vorliegenden Fall als Rangsummen der Items über alle Blöcke berechnet, wobei das Item mit dem höchsten Rang 6 Punkte, das niedrigste Item 0 Punkte erhält. Insgesamt ist die maximale Ausprägung der Testwerte auf eine Spanne von 60 Punkte begrenzt. Die einfache Aufsummierung der Itemränge bedeutet, dass im Grund angenommen wird, dass es sich bei den Items um essentiell τ-äquivalente Variablen im Sinne der klassischen Testtheorie handelt (Eid et al., 2015) und alle Ränge so behandelt werden, als hätten Sie die Ladung 1. Diese Annahme kann berechtigterweise in Frage gestellt werden, stellt im Moment jedoch die einzig sinnvolle Berechnung dar, weil ipsative Daten offenkundig nicht für die Spezifizierung eines klassischen Testmodells geeignet sind (Brown & Maydeu-Olivares, 2011). Ipsative Testwertverteilungen können folglich aufgrund der Itemabhängigkeiten innerhalb der Blöcke auch durch das FC-Format verzerrt sein. Geht man davon aus, dass die Reihenfolgen der Wertesystemhierarchien in der Gesamtbevölkerung normalverteilt sind, die zugrunde liegende Stichprobe repräsentativ für die Gesamtbevölkerung ist und die Qualitäten der Items eines ipsativen Tests homogen sind, dann wäre der Vergleich der Testwertverteilungen auf ihre Übereinstimmungen mit der Normalverteilung angebracht. Bei der vorliegenden Stichprobe muss die Repräsentativität allerdings schon deshalb in Frage gestellt werden, da ein großer Teil der Stichprobe aus Studierenden besteht.

Fazit der Auswirkungen der Ipsativität auf die Itemkennwerte

Unter Berücksichtigung der dargestellten Auswirkungen der Ipsativität auf die Itemkennwerte, kann gesagt werden, dass die beiden Schwierigkeitsindizes und die Itemvarianzen diejenigen Kennwerte sind, die am direktesten mit der Qualität eines Items zusammenhängen und deshalb am nachvollziehbarsten sind. Bei den paarweisen Itemschwierigkeiten sind die Abhängigkeiten auf zwei Items und bei den normalen Itemschwierigkeiten sowie den Itemvarianzen auf die Items desselben Blocks begrenzt. Trennschärfen und Testwertverteilungen hingegen beinhalten Informationen des gesamten Tests. Zur Evaluation der Items sind deshalb die Schwierigkeitskennwerte und Varianzen geeigneter als Trennschärfen und Testwertverteilungen. Die nun folgende Analyse und Ableitung potenziell verbesserungswürdiger Items fokussiert sich deshalb auf die erstgenannten Maße, wobei dennoch alle Kennwerte berichtet werden.

Richtlinien zur Beurteilung der Itemkennwerte

Bei der Beurteilung von Itemkennwerten haben sich bestimmte Faustregeln etabliert. Diese werden hier aufgegriffen und, um der Ipsativität der Daten Rechnung zu tragen, leicht modifiziert. Tabelle 3 zeigt die gesammelten Beurteilungsrichtlinien, die sich an der Darstellung von Bühner (2011, S.81) orientieren.

Tabelle 3. Beurteilungsrichtlinien für Kennwerte der Itemanalyse.

Kennwert	Kürzel	Niedrig	Mittel	Hoch
Schwierigkeit	P_i	$> .75$	$.75 - .25$	$< .25$
Paarweise Schwierigkeit	P_{pi}	$> .80$	$.80 - .20$	$< .20$
Trennschärfe	r_{it}	$< .30$	$.30 - .50$	$> .50$
		Min	M (SD)	Max
Itemvarianz	V_i	1.76	3.2 (0.6)	5.13
Trennschärfe	r_{it}	.00	.34 (.13)	.59

Anmerkung. Beurteilungsrichtlinien in Anlehnung an Bühner (2011) und auf Basis der im Fließtext dargelegten Zusammenhänge.

Bei Itemschwierigkeiten gilt allgemein, dass extreme Werte, d.h. sehr niedrige oder hohe Werte andeuten, dass ein Item wenig differenziert und folglich als Item mit geringerer Qualität eingestuft werden kann (Kelava & Moosbrugger, 2012). Wie oben beschrieben, sind ipsative Itemschwierigkeiten zur Mitte hin verzerrt. Um dieser Verzerrung Rechnung zu tragen, wurden die Grenzwerte um $.05$ zur Mitte hin verschoben. Für die paarweisen Itemschwierigkeiten gilt die Verzerrung zur Mitte hin nicht, weswegen die gebräuchlichen Richtwerte von $.80$ und

.20 beibehalten wurde. Des Weiteren sei angemerkt, dass der optimale Wert einer paarweisen Itemschwierigkeit bei $P_{pi} = .50$ liegt, da in diesem Fall die aus diesem Paarvergleich gewonnene Information am größten ist.

Für die Trennschärfen wurden ebenfalls die in Bühner (2011) berichteten Richtwerte verwendet, wobei Folgendes zu bedenken ist: Trennschärfen kommen allgemein eher bei höheren Itemvarianzen zu Stande (Kelava & Moosbrugger, 2012). Im vorliegenden Fall sind diese allerdings durch die Ipsativität künstlich begrenzt, weswegen auch die Höhen der Trennschärfen niedriger als bei klassischen Instrumenten ausfallen dürften. Dies bedeutet zum einen, dass die verwendeten Richtwerte als konservativ gesehen werden können und auch vor allem der relative Vergleich untereinander aufschlussreich sein kann. Für diesen Vergleich bietet es sich an, Mittelwert, Standardabweichung und die Extremwerte aller Trennschärfen als Referenzwerte heranzuziehen. Diese sind auch in Tabelle 3 aufgeführt.

Da Itemvarianzen keine standardisierten Kennwerte sind, ist es naheliegend, darauf das selbe Verfahren anzuwenden, zumal für die Itemvarianzen auch keine *textbook*-Richtlinien gefunden werden konnten. Demnach wurden in Tabelle 3 auch Mittelwert, Standardabweichung und Extremwerte aller Itemvarianzen als Referenzpunkte angegeben. Grundsätzlich gilt jedoch, je höher die Varianz, umso mehr differenziert das Item zwischen unterschiedlichen Merkmalsausprägungen und umso größer ist die Information, die aus dem Item bezogen auf das gemessene latente Konstrukt gezogen werden kann (Kelava & Moosbrugger, 2012).

Um Testwertverteilungen zu beurteilen, werden diese gemäß Kelava und Moosbrugger (2012) anhand von Mittelwert, Median, Modal, Testwertvarianz, Spanne, Schiefe und Exzess untersucht.

5.2 Ergebnisse

Im Ergebnisteil werden nun die Itemkennwerte der beiden Subskalen zuerst berichtet und dann beurteilt. Tabelle 4 zeigt zunächst die Itemschwierigkeiten und Tabelle 5 exemplarisch die paarweisen Itemschwierigkeiten des ersten Blocks der Annäherungsskala. Die komplette tabellarische Ansicht der paarweisen Itemschwierigkeiten ist aus Gründen der Übersichtlichkeit in Anhang A dargestellt. Danach folgen die Itemvarianzen (Tabelle 6), die graphische Darstellung des Zusammenhangs zwischen Itemschwierigkeiten und Itemvarianzen (Abbildung 3), Trennschärfen (Tabelle 7) und Testwertverteilungen (Tabelle 8), jeweils der Annäherungsskala. Anschließend werden in der gleichen Reihenfolge die entsprechenden Kennwerte der MVSQ^V-Skala berichtet (Tabellen 9 bis 12). Zum Abschluss des Ergebnisteils werden die überarbeitungswürdigen Items in tabellarischer Form zusammengefasst.

5.2.1 Itemkennwerte

Tabelle 4 enthält die Itemschwierigkeiten der Annäherungswertesysteme. Dabei sei wiederholt, dass der Mittelwert pro Block dem FC-Format geschuldet auf den Wert von $.50$ festgelegt ist. Würde man die oben besprochenen, nicht verschärften Beurteilungsrichtlinien für Itemschwierigkeiten anwenden, so würden lediglich drei Itemschwierigkeiten diese Grenzwerte (Tabelle 3) über- bzw. unterschreiten. Nach den verschärften Richtlinien müssen insgesamt acht und davon jeweils vier Itemschwierigkeiten als zu niedrig bzw. zu hoch eingestuft werden. Konkret sind die Items GB_1^A, GB_7^A, NA_6^A sowie NA_8^A besonders schwer und die Items GL_5^A, GL_{10}^A, VE_6^A sowie VE_{10}^A besonders leicht zu bevorzugen. Dabei fällt auf, dass nur die Blöcke 2, 3 und 4 nicht von extremen Schwierigkeiten betroffen sind und sich die möglicherweise problematischen Items relativ gleich auf die übrigen Blöcke verteilen.

Tabelle 4. Itemschwierigkeiten der $MVSQ^A$-Skala.

	Block									
Wertesystem	1	2	3	4	5	6	7	8	9	10
GB^A	.21	.31	.46	.27	.37	.52	.18	.40	.33	.36
MA^A	.46	.37	.52	.26	.52	.37	.32	.41	.43	.40
GW^A	.50	.46	.58	.39	.34	.43	.59	.68	.35	.40
ER^A	.70	.52	.30	.70	.59	.61	.68	.52	.65	.30
GL^A	.72	.71	.68	.59	.77	.67	.55	.74	.69	.78
VE^A	.63	.68	.68	.62	.57	.78	.74	.54	.62	.81
NA^A	.29	.46	.28	.67	.34	.12	.45	.22	.43	.44

Anmerkung. Wertesysteme: GB = Geborgenheit; MA = Macht; GW = Gewissheit; ER = Erfolg; GL = Gleichheit; VE = Verstehen; NA = Nachhaltigkeit; A = Annäherung.

Die Tabelle 5 zeigt nun die paarweisen Itemschwierigkeiten des ersten Blocks. Dabei ist leicht zu erkennen, dass sich die normalen Itemschwierigkeiten aus den Mittelwerten der sechs paarweisen Itemschwierigkeiten berechnen lassen. Zum Beispiel berechnet sich die Itemschwierigkeit des Items „Geborgenheit“ (P_i von $GB_1^A = .21$) aus dem Mittelwert der sechs Werte in der ersten Zeile in Tabelle 4 oder als Eins minus dem Mittelwert der ersten Spalte. Des Weiteren ist zu beachten, dass die paarweisen Schwierigkeiten von Zeile zu Spalte zu lesen sind. Zum Beispiel wurde das Item MA_1^A in 75% der Fälle *über* das Item GB_1^A angeordnet und das GB_1^A-Item dementsprechend nur zu 25% über MA_1^A gerankt. Man könnte auch sagen, dass das MA_1^A-Item in der Stichprobe häufiger als GB_1^A und folglich „leichter“ zu bevorzugen war. Zwei weitere Vorteile der paarweisen Itemschwierigkeiten sollen am Beispiel von GB_1^A

Tabelle 5. Paarweise Itemschwierigkeiten des ersten Blocks der MVSQ^A-Skala.

	GB^A	MA^A	GW^A	ER^A	GL^A	VE^A	NA^A
GB^A		.25	.22	.12	.09	.17	.38
MA^A	.75		.50	.28	.26	.33	.67
GW^A	.78	.50		.33	.30	.40	.69
ER^A	.88	.72	.67		.50	.55	.87
GL^A	.91	.74	.70	.50		.58	.86
VE^A	.83	.67	.60	.45	.42		.80
NA^A	.62	.33	.31	.13	.14	.20	

Anmerkung. Wertesysteme: GB = Geborgenheit; MA = Macht; GW = Gewissheit; ER = Erfolg; GL = Gleichheit; VE = Verstehen; NA = Nachhaltigkeit; A = Annäherung.

verdeutlicht werden. Erstens kann man sehen, dass die paarweisen Schwierigkeiten tiefer gehende Differenzierungen ermöglichen als die klassischen Schwierigkeiten. Die paarweise Itemschwierigkeit des GB_1^A-Items mit GL_1^A ($P_{pi} = .09$) und ER_1^A ($P_{pi} = .12$) sind hoch problematisch, da diese Paarvergleiche aufgrund der Einseitigkeit nur sehr wenig Information über die latenten Konstrukte liefern. Im Vergleich mit Item NA_1^A ($P_{pi} = .38$) ist der Wert jedoch unbedenklich. Diese Unterschiede werden aus der normalen Itemschwierigkeit nicht ersichtlich. Zweitens kann anhand desselben Beispiels gesehen werden, dass die Verschärfung der Beurteilungsrichtlinie für die normalen Itemschwierigkeiten insofern sinnvoll ist, da die Itemschwierigkeit von GB_1^A mit $P_i = .21$ unter den ursprünglichen Grenzwerten als akzeptabel eingestuft würde, die paarweisen Itemschwierigkeiten jedoch klar zeigen, dass es sich um ein problembehaftetes Item handelt.

Von allen paarweisen Itemschwierigkeiten des ersten Blocks sind nach der Beurteilungsrichtlinie zwölf Schwierigkeiten problematisch, die aufgrund der bivariaten Natur der Kennwerte nur sechs Paarvergleiche betreffen. Bei diesen Paarvergleichen handelt es sich um $GB_1^A - GW_1^A$ ($P_{pi} = .22$), $GB_1^A - ER_1^A$ ($P_{pi} = .12$), $GB_1^A - GL_1^A$ ($P_{pi} = .09$), $GB_1^A - VE_1^A$ ($P_{pi} = .17$), $GL_1^A - NA_1^A$ ($P_{pi} = .86$) und $VE_1^A - NA_1^A$ ($P_{pi} = .80$). Da vier dieser Werte das Item GB_1^A betreffen, kann mit hoher Wahrscheinlichkeit davon ausgegangen werden, dass es sich dabei um ein schlechtes Item handelt. Zwei der Paarschwierigkeiten involvieren das Item NA_1^A. Dies könnte ebenfalls darauf hindeuten, dass dieses Item tendenziell schwächer ist.

Zu Beurteilung der Itemvarianzen der MVSQ^A-Skala (Tabelle 6) können zunächst die Referenzwerte aus Tabelle 3 herangezogen werden. Demnach liegt insgesamt eine Itemvarianz mehr als zwei Standardabweichungen unter dem Durchschnitt aller Itemvarianzen und zwar die Varianz von Item NA_6^A mit $V_i = 1.87$. Des Weiteren liegen 18 Itemvarianzen zwischen

Tabelle 6. Itemvarianzen der MVSQ^A-Skala.

	Block										
Wertesystem	1	2	3	4	5	6	7	8	9	10	M
GB^A	2.05	2.96	4.18	2.54	3.65	2.47	2.08	2.96	3.02	3.05	2.90
MA^A	3.02	3.66	4.05	2.53	3.72	3.36	2.59	3.48	3.93	3.23	3.36
GW^A	3.77	3.35	3.05	3.20	3.46	2.78	3.45	3.15	3.23	2.82	3.23
ER^A	2.54	3.23	2.95	3.14	2.80	2.50	2.59	3.43	3.05	3.03	2.93
GL^A	2.48	3.41	2.40	2.88	2.65	2.46	2.65	2.45	2.46	2.15	2.60
VE^A	3.45	3.15	2.77	2.85	3.17	2.39	2.46	3.55	3.47	1.76	2.90
NA^A	2.26	3.29	2.69	3.28	2.91	1.87	3.58	2.16	4.07	2.83	2.89
M	2.79	3.29	3.16	2.92	3.19	2.55	2.77	3.03	3.32	2.70	2.97

Anmerkung. Wertesysteme: GB = Geborgenheit; MA = Macht; GW = Gewissheit; ER = Erfolg; GL = Gleichheit; VE = Verstehen; NA = Nachhaltigkeit; A = Annäherung; M = Mittelwert.

einer und zwei Standardabweichungen unter dem Mittelwert. An dieser Stelle wird darauf verzichtet, diese einzeln zu berichten, da die Werte Tabelle 6 entnommen werden können. Erwähnt sei allerdings, dass insbesondere der gesamte Block 6 problematisch erscheint, da die mittlere Itemvarianz dieses Blocks ($V_{i6} = 2.55$) um mehr als eine Standardabweichung vom Gesamtmittel abweicht. Als ebenso schwächer – bezogen auf die Differenzierungsfähigkeit – können die Blöcke 10 ($V_{i10} = 2.70$), 7 ($V_{i7} = 2.77$) und 1 ($V_{i1} = 2.79$) eingestuft werden. Auf Merkmalsebene kann gesagt werden, dass insbesonders Items des Wertesystems ***Gleichheit*** auffallen, da sie eine relativ niedrige durchschnittlichen Itemvarianz von 2.60 aufweisen. Das bedeutet, dass Items dieses Wertesystems verhältnismäßig wenig zwischen hohen und niedrigen Ausprägungen unterscheiden können. Auf Skalenebene sei zudem hinzugefügt, dass die durchschnittliche Varianz aller Items dieser Skala mit 2.97 unter dem Durchschnittswert des gesamten Fragebogens (3.20, vgl. Tabelle 3) liegt.

In Abbildung 3 ist der Zusammenhang der Itemschwierigkeiten und Itemvarianzen der MVSQ^A-Subskala zu sehen. Dieser gestaltet sich tendenziell so wie von Kelava und Moosbrugger (2012) beschrieben, nämlich dass die Itemvarianz bei mittleren Itemschwierigkeiten am höchsten ist. Die Abbildung wurde hier aus zweierlei Gründen aufgenommen. Zum einen sollte graphisch überprüft werden, ob dieser Zusammenhang auch für ipsative Maße gilt und zum anderen kann daran grob abgeschätzt werden, welcher Anteil der Items als potenziell unbrauchbar beurteilt werden kann. Im vorliegenden Fall sind vor allem die Items an den Rändern mit Schwierigkeiten >.25 und <.75 gemeint (gestrichelte Linien), wobei zu sehen ist,

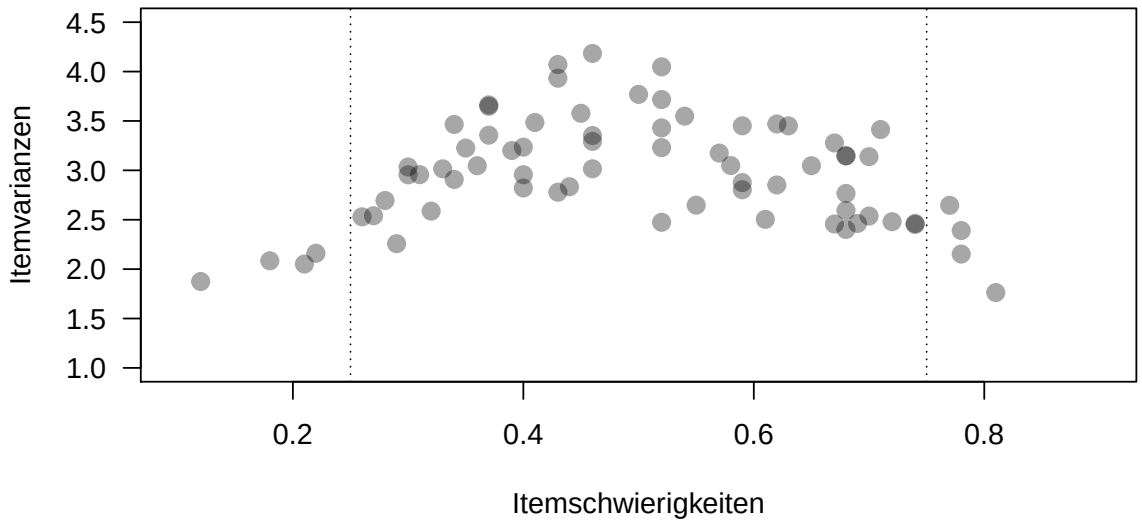

Abbildung 3. Zusammenhang der Itemschwierigkeiten und Itemvarianzen der MVSQ^A-Skala. Gestrichelte Linien sind die Beurteilungsrichtlinien der Itemschwierigkeiten.

dass hohe Itemschwierigkeiten nicht zwangsweise mit niedrigen Itemvarianzen einhergehen, da es einige Items mit niedrigeren Varianzen, d.h. niedrigeren Differenzierungsfähigkeiten gibt als das „problematische" Item, das am nächsten an der 0.75-Linie liegt. Insgesamt ist der typisch kurvilineare Zusammenhang zwischen Itemvarianzen und Schwierigkeiten (Kelava & Moosbrugger, 2012) jedoch zu erkennen.

Weiter folgen nun die Trennschärfen und Testwertverteilungen der MVSQ^A-Skala. Diese Kennwerte sind wie oben erläutert weniger geeignet, um einzelne Items zu beurteilen, stattdessen sollte eher auf die Qualität einzelner Blöcke, Merkmale und der Skala als Ganzes rückgeschlossen werden. Bei den Trennschärfen (Tabelle 7) fällt zunächst auf, dass keiner der mittleren Trennschärfen pro Block oder Merkmal im als „hoch" eingestuften Bereich, d.h. über .50 liegt. Andererseits liegen auch nur zwei Blöcke und ein Merkmal im niedrigen Bereich unter .30. Block 6 weist die mit Abstand niedrigste Trennschärfe von $r_{it} = .16$ auf. Auch Block 10 liegt mit $r_{it} = .29$ unterhalb der wünschenswerten Grenze von .30. Auf Merkmalsebene sticht besonders die durchschnittliche Trennschärfe der ***Erfolg***^{A}-Items mit $r_{it} = .26$ hervor, was mit Abstand den niedrigsten Wert darstellt und bedeutet, dass die Items dieses Merkmals als sehr heterogen angesehen werden können. Alle anderen durchschnittlichen Trennschärfen pro Merkmal liegen im mittleren Bereich.

Aus Tabelle 8 kann abgelesen werden, dass mit Ausnahme der Testwerte von ***Erfolg***^{A} alle Testwertverteilungen leicht schief sind. Die Verteilungen von ***Geborgenheit***^{A}, ***Macht***^{A}, ***Gewissheit***^{A} und ***Nachhaltigkeit***^{A} sind dabei rechtsschief, die von ***Gleichheit***^{A} und ***Verstehen***^{A}

Tabelle 7. Trennschärfen der MVSQ^{A}-Skala.

	Block										
Wertesystem	1	2	3	4	5	6	7	8	9	10	M
GB^{A}	.42	.39	.42	.28	.32	.09	.44	.52	.42	.35	.37
MA^{A}	.41	.32	.48	.41	.52	.21	.42	.36	.52	.38	.40
GW^{A}	.57	.51	.26	.49	.46	.16	.35	.41	.57	.23	.40
ER^{A}	.15	.36	.27	.22	.24	.19	.35	.39	.37	.11	.26
GL^{A}	.43	.38	.40	.35	.53	.25	.49	.46	.53	.44	.43
VE^{A}	.41	.52	.39	.49	.39	.22	.42	.47	.41	.17	.39
NA^{A}	.30	.27	.47	.22	.41	.00	.57	.50	.59	.33	.37
M	.38	.39	.38	.35	.41	.16	.43	.45	.49	.29	.37

Anmerkung. Wertesysteme: GB = Geborgenheit; MA = Macht; GW = Gewissheit; ER = Erfolg; GL = Gleichheit; VE = Verstehen; NA = Nachhaltigkeit; A = Annäherung; M = Mittelwert.

Tabelle 8. Testwertverteilungen der MVSQ^{A}-Skala.

Wertesystem	M	Md	Mod	SD	Min	Max	Schiefe	Exzess
GB^{A}	20.37	20	17	8.88	2	50	0.42	-0.22
MA^{A}	24.35	24	22	10.04	3	56	0.22	-0.37
GW^{A}	28.39	28	25	9.86	3	55	0.14	-0.46
ER^{A}	33.38	33	32	7.88	10	57	-0.03	-0.36
GL^{A}	41.39	42	49	9.08	12	59	-0.44	-0.35
VE^{A}	40.04	41	45	9.23	10	60	-0.20	-0.61
NA^{A}	22.09	21	19	8.95	2	50	0.41	-0.26

Anmerkung. M = Mittelwert; Md = Median; Mod = Modalwert; SD = Standardabweichung; Wertesysteme: GB = Geborgenheit; MA = Macht; GW = Gewissheit; ER = Erfolg; GL = Gleichheit; VE = Verstehen; NA = Nachhaltigkeit; A = Annäherung.

linksschief. Dementsprechend gestalten sich auch Mittelwert, Median, Modalwert und die Spannweite, d.h. je rechtsschiefer eine Verteilung, umso niedriger, je linksschiefer, umso höher sind die entsprechenden Werte. ***Macht***^{A} weist die höchste und ***Erfolg***^{A} die niedrigste Streuung auf, wobei alle Verteilungen negative Kurtosi aufweisen, d.h. im Zentrum weniger „spitz“ und an den Rändern weniger breit als die Normalverteilung sind (DeCarlo, 1997). Insgesamt

weichen die Schiefe und Exzess-Werte jedoch nur gering von denen einer Normalverteilung ab, weswegen diese als zumindest approximativ normalverteilt bezeichnet werden können.

Nachdem nun die Kennwerte der MVSQ^A-Skala berichtet wurden, folgt nun die Darstellung der Kennwerte der MVSQ^V-Skala. Von allen Itemschwierigkeiten (Tabelle 9) liegen lediglich die Koeffizienten der Items GW_9^V ($P_i = .76$), GL_8^V ($P_i = .19$) und VE_2^V ($P_i = .22$) außerhalb der festgelegten Grenzwerte. Im Vergleich mit der MVSQ^A-Skala sind die Schwierigkeiten demnach gleichmäßiger verteilt. Die paarweisen Itemschwierigkeiten wurden aufgrund des Umfangs wieder in den Anhang A verschoben. Auf sie wird in komprimierter Form bei den Überarbeitungsempfehlungen eingegangen.

Tabelle 9. Itemschwierigkeiten der MVSQ^V-Skala.

	Block									
Wertesystem	1	2	3	4	5	6	7	8	9	10
GB^V	.68	.64	.50	.44	.49	.65	.48	.32	.33	.60
MA^V	.41	.68	.67	.56	.61	.35	.56	.67	.62	.46
GW^V	.59	.55	.72	.42	.57	.51	.44	.65	.76	.53
ER^V	.60	.65	.48	.73	.29	.33	.57	.50	.44	.57
GL^V	.44	.31	.38	.43	.64	.58	.48	.19	.49	.59
VE^V	.33	.22	.41	.53	.48	.54	.40	.73	.37	.34
NA^V	.46	.45	.35	.39	.44	.55	.57	.43	.49	.40

Anmerkung. Wertesysteme: GB = Geborgenheit; MA = Macht; GW = Gewissheit; ER = Erfolg; GL = Gleichheit; VE = Verstehen; NA = Nachhaltigkeit; V = Vermeidung.

Tabelle 10 zeigt die Itemvarianzen der MVSQ^V-Skala. Auch hier ergibt sich der Eindruck, dass die Kennwerte besser ausfallen als in der MVSQ^A-Skala, da sie sowohl insgesamt höher liegen, als auch einzeln gleichmäßiger verteilt sind. Kein Wert ist weiter als zwei Standardabweichungen vom Referenz-Mittelwert entfernt und nur sechs Werte liegen mehr als eine Standardabweichung unter dem Referenz-Mittelwert. Zum Vergleich, in der MVSQ^A-Skala waren von diesen Maßstäben insgesamt 19 Items betroffen. Auffallend sind einzig die Itemvarianzen innerhalb von Block 8, die mit Ausnahme des Items ER_8^V verhältnismäßig niedrig ausfallen (P_{pi} zwischen 2.04 und 2.85). Dieser Block besitzt demnach eine deutlich geringere Differenzierungsfähigkeit. Die Auftragung der Itemvarianzen gegen die Itemschwierigkeiten der MVSQ^V-Skala (Abbildung 4) verbildlichen die beschriebenen Ergebnisse. Insgesamt liegen nur drei Items außerhalb der Grenzen der Itemschwierigkeiten. Der kurvilineare Zusammenhang ist auch in dieser Abbildung zu erkennen.

Tabelle 10. Itemvarianzen der MVSQ^V-Skala.

	Block										
Wertesystem	1	2	3	4	5	6	7	8	9	10	M
GB^V	3.32	2.56	3.46	4.37	4.11	3.65	5.13	2.04	2.76	3.77	3.52
MA^V	3.95	3.24	3.92	3.80	3.56	4.02	4.03	2.54	3.57	4.29	3.69
GW^V	3.35	2.95	2.76	3.64	4.09	3.91	3.37	2.73	3.19	4.10	3.41
ER^V	3.93	3.71	3.81	2.95	2.80	2.50	3.58	4.26	2.65	3.54	3.37
GL^V	3.66	3.05	3.36	3.83	3.47	3.92	3.73	2.34	3.51	3.58	3.44
VE^V	3.40	2.51	3.42	2.95	3.15	3.18	3.94	2.62	3.57	2.89	3.16
NA^V	3.09	3.04	2.90	3.40	3.81	3.76	3.23	2.85	4.06	3.70	3.38
M	3.53	3.01	3.38	3.56	3.57	3.56	3.86	2.77	3.33	3.69	3.43

Anmerkung. Wertesysteme: GB = Geborgenheit; MA = Macht; GW = Gewissheit; ER = Erfolg; GL = Gleichheit; VE = Verstehen; NA = Nachhaltigkeit; V = Vermeidung; M = Mittelwert.

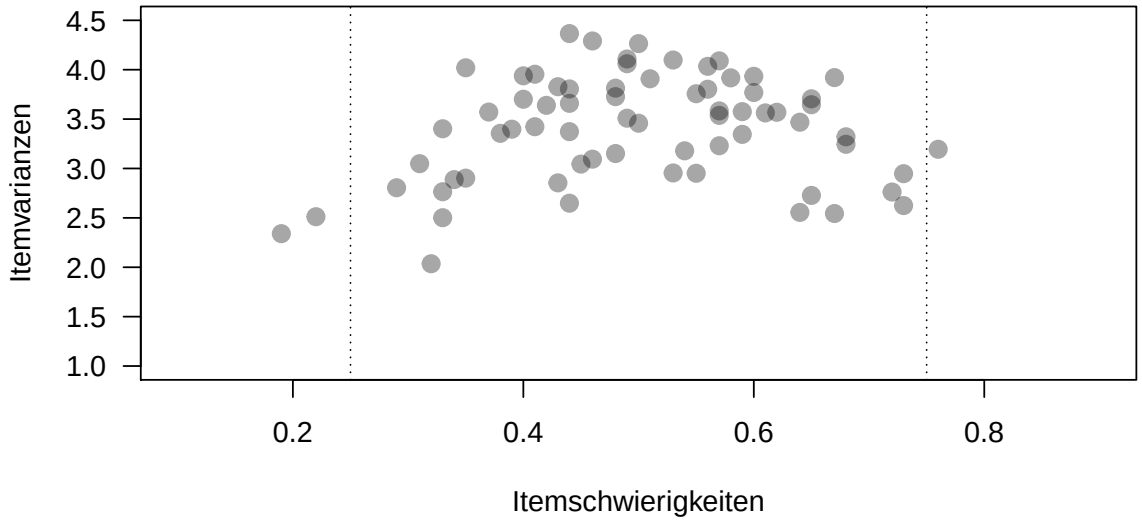

Abbildung 4. Zusammenhang der Itemschwierigkeiten und Itemvarianzen der MVSQ^V-Skala. Gestrichelte Linien sind die Beurteilungsrichtlinien der Itemschwierigkeiten.

Die Trennschärfen (Tabelle 11) passen überraschenderweise nicht in das Bild, das die bisherigen Itemkennwerte vermitteln, nämlich dass die Items der MVSQ^V-Skala besser sind als die der MVSQ^A-Skala. Denn diese sind im Durchschnitt deutlich niedriger ($r_{it} = .30$)

Tabelle 11. Trennschärfen der MVSQ^{V}-Skala.

	Block										
Wertesystem	1	2	3	4	5	6	7	8	9	10	M
GB^{V}	.45	.31	.30	.21	.35	.46	.29	.04	.15	.36	.29
MA^{V}	.46	.42	.34	.34	.50	.38	.44	.16	.40	.27	.37
GW^{V}	.45	.38	.45	.34	.58	.45	.37	.34	.32	.57	.42
ER^{V}	.30	.06	.32	.24	.19	.07	.07	.24	.17	.04	.17
GL^{V}	.55	.38	.43	.28	.40	.44	.26	.23	.24	.27	.35
VE^{V}	.28	.17	.26	.30	.09	.26	.21	.19	.14	.23	.21
NA^{V}	.37	.32	.26	.34	.37	.26	.32	.17	.04	.20	.26
M	.41	.29	.34	.29	.35	.33	.28	.20	.21	.28	.30

Anmerkung. Wertesysteme: GB = Geborgenheit; MA = Macht; GW = Gewissheit; ER = Erfolg; GL = Gleichheit; VE = Verstehen; NA = Nachhaltigkeit; V = Vermeidung; M = Mittelwert.

Tabelle 12. Testwertverteilungen der MVSQ^{V}-Skala.

Wertesystem	M	Md	Mod	SD	Min	Max	Schiefe	Exzess
GB^{V}	30.72	31	24	8.96	6	51	-0.07	-0.53
MA^{V}	33.53	34	35	10.17	4	57	-0.20	-0.34
GW^{V}	34.48	34	32	10.42	7	60	-0.06	-0.36
ER^{V}	30.93	31	32	7.42	10	50	-0.17	-0.28
GL^{V}	27.08	27	26	9.52	1	57	-0.15	-0.28
VE^{V}	26.14	26	26	7.62	2	47	-0.17	0.03
NA^{V}	27.11	27	23	8.42	3	52	0.05	-0.43

Anmerkung. M = Mittelwert; Md = Median; Mod = Modalwert; SD = Standardabweichung; Wertesysteme: GB = Geborgenheit; MA = Macht; GW = Gewissheit; ER = Erfolg; GL = Gleichheit; VE = Verstehen; NA = Nachhaltigkeit; V = Vermeidung.

und es gibt auch mehr Blöcke und Merkmale, die geringere Trennschärfen als .30 aufweisen. Konkret sind dies die Blöcke 2 und 4 (jeweils $r_{it} = .29$), Block 7 ($r_{it} = .28$), Block 8 ($r_{it} = .20$), Block 9 ($r_{it} = .21$) und Block 10 ($r_{it} = .28$), d.h. sechs der zehn Blöcke haben sehr niedrige Trennschärfen. Auch auf Merkmalsebene sind die Trennschärfen niedriger als in der MVSQ^{A}-Skala: ***Geborgenheit***^{V} ($r_{it} = .29$), ***Erfolg***^{V} ($r_{it} = .17$), ***Verstehen***^{V} ($r_{it} = .21$) und

Nachhaltigkeit^V ($r_{it} = .26$). Dies kann so interpretiert werden, dass die Items je desselben Wertesystems weniger homogen als bei der MVSQ^A-Skala sind.

Im Gegensatz dazu ergeben die Testwertverteilungen der MVSQ^V-Skala (Tabelle 12) wieder ein ausgewogeneres Bild. Im Verglichen mit den Kennwerten der MVSQ^A-Skala sind die Mittelwerte, Mediane und Modale merklich ausgeglichener. Auch die Spannweiten sind mit Ausnahme der von ***Erfolg***^{V} untereinander ähnlicher und die Standardabweichungen sind, was die Ausprägungen betrifft, in etwa vergleichbar. Die Schiefe-Kennwerte sind deutlich geringer als in der MVSQ^A-Skala und auch die Exzesse sind kleiner. Demnach kann gesagt werden, dass die Verteilungen der Vermeidungswertesysteme der Normalverteilung ähnlicher sind als die der Annäherungswertesysteme.

5.2.2 Itembeurteilung

Zur besseren Übersichtlichkeit werden die Empfehlungen zur Überarbeitung der Items zuerst vollständig für die MVSQ^A-, danach für die MVSQ^V-Skala dargelegt.

Auf Blockebene sticht bei der MVSQ^A-Skala besonders Block 6 hervor, denn er weist sowohl deutlich unterdurchschnittliche Itemvarianzen (vgl. Tabelle 6) als auch sehr niedrige Trennschärfen auf (vgl. Tabelle 7). Von den Itemschwierigkeiten dieses Blocks (vgl. Tabellen 4 und 78) liegen allerdings lediglich zwei Werte (der Items VE_6^A und NA_6^A) außerhalb des gewünschten Bereichs und zwar sowohl bzgl. der klassischen als auch der paarweisen Schwierigkeiten. Auf Blockebene hat zudem noch Block 10 mit der durchschnittlichen Trennschärfe von .29 einen Wert unter der Grenze von .30.

Vergleicht man die Items derselben Wertesysteme miteinander, kann einerseits gesagt werden, dass die ***Erfolg***^{A}-Items im Durchschnitt eine niedrige Trennschärfe ($r_{it} = .26$) aufweisen. Bei den Itemschwierigkeiten liegen zwar einige Itemschwierigkeiten knapp an den Grenzen, jedoch keine einzige Itemschwierigkeit im extremen Bereich.

Die Empfehlung auf Block- und Itemebene lautet demzufolge, dass Block 6 und die Items des ***Erfolg***^{A}-Wertesystems einer Überprüfung unterzogen werden sollen. Um nun spezifischere Empfehlungen dahingehend abgeben zu können, welche Einzelitems überarbeitet werden sollen, werden im Folgenden die anhand der paarweisen Itemschwierigkeiten als besonders problematisch identifizierten Items dargelegt. Diejenigen Items, dessen paarweise Itemschwierigkeiten die die Grenzwerte aus Tabelle 3 mehr als drei mal über- bzw. unterschreiten, werden dabei in die Kategorie *problematische* Items eingeordnet. Alle Items, die in nur zwei solchen paarweisen Itemschwierigkeiten involviert sind, werden in der Kategorie *fragwürdig* zusammengefasst. Diese Einteilung erscheint nach logischen Überlegungen in diesem Stadium der Testentwicklung angemessen, da es nicht sinnvoll ist, zu viele Items auf einmal zu überarbeiten. Auch die Veränderung weniger FC-Items birgt bereits die Gefahr, dass die Auswirkungen der Veränderungen eines Items auf die übrigen Items des Blocks nicht mehr nachvollziehbar sind

(Brown & Maydeu-Olivares, 2012). Bei FC-Fragebögen gilt aufgrund der zahlreichen Interdependenzen zwischen den Items das Prinzip, pro Schritt besser weniger Items zu verändern und stattdessen mehrere Revisionsschritte durchzuführen. Items mit nur einer Itemschwierigkeit können dann in späteren Revisionen berücksichtigt werden.

Tabelle 13 fasst folglich diejenigen Items der MVSQ^A-Skala zusammengefasst, die als „problematisch" eingestuft und deshalb als dringend revisionsbedürftig angesehen werden können. In Block 1 sind also die Items ***Geborgenheit***^{A} und ***Nachhaltigkeit***^{A} zu nennen, die beide verglichen mit ***Erfolg***^{A}, ***Gleichheit***^{A} und ***Verstehen***^{A} zu schwierig sind. Eventuell sind auch diese drei letztgenannten Items zu leicht, da sie jeweils zwei sehr niedrige Itemschwierigkeiten aufweisen (siehe Tabelle 14). Des Weiteren sollte das Item ***Macht***^{A} in Block 4 so angepasst werden, dass es leichter zu bevorzugen wird, ***Gleichheit***^{A} in Block 5 schwieriger, ***Geborgenheit***^{A} in Block 7 und ***Nachhaltigkeit*** in Block 8 beide leichter, sowie ***Gleichheit***^{A} und ***Verstehen***^{A} in Block 10 schwieriger.

Tabelle 13. Problematische Items der MVSQ^A-Skala.

		Paarweise Itemschwierigkeiten P_{pi}						
Item	P_i	GB^A	MA^A	GW^A	ER^A	GL^A	VE^A	NA^A
GB$^A_1$	.21				.12	.09	.17	
NA$^A_1$	.29				.13	.14	.20	
MA$^A_4$	.26				.13		.16	.17
GL$^A_5$	.77	.82		.85				.87
VE$^A_6$	.78		.82	.82				.94
NA$^A_6$	.12	.11		.15	.09	.09	.06	
GB$^A_7$	.18			.15	.11	.14	.10	
NA$^A_8$	.22			.14		.10	.20	
GL$^A_{10}$	.78	.84	.82	.84	.85			.83
VE$^A_{10}$	.81	.85	.85	.88	.89			.86

Anmerkung. P_i = Itemschwierigkeit; P_{pi} = Paarweise Itemschwierigkeit; Wertesysteme: GB = Geborgenheit; MA = Macht; GW = Gewissheit; ER = Erfolg; GL = Gleichheit; VE = Verstehen; NA = Nachhaltigkeit; A = Annäherung; Zahlen indizieren die Blocknummer.

Von den als mit *fragwürdiger* Güte eingestuften Items ist vor allem Block 10 auffällig, denn hier gibt es eine deutliche Zweiteilung des Blocks in leichte Items (***Gleichheit***^{A} und ***Verstehen***^{A}) sowie schwere Items (die verbleibenden Items). Tabelle 14 zeigt hier, dass die Items der beiden Wertesysteme im Vergleich zu allen anderen Items zu einfach zu bevorzugen

Tabelle 14. Fragwürdige Items der MVSQ^A-Skala.

		Paarweise Itemschwierigkeiten P_{pi}						
Item	P_i	GB^A	MA^A	GW^A	ER^A	GL^A	VE^A	NA^A
ER$_1^A$	.70	.88						.87
GL$_1^A$	.72	.91						.86
VE$_1^A$	.63	.83						.80
ER$_3^A$	.30					.19	.19	
GL$_3^A$	.68				.81			.85
VE$_3^A$	.68				.81			.85
NA$_3^A$	.28					.15	.15	
GB$_4^A$	.27				.17	.20		.16
ER$_4^A$	.70	.83	.87					
NA$_4^A$	.67	.84	.83					
GW$_6^A$	.43						.18	.85
MA$_7^A$	.32				.16		.14	
ER$_7^A$	.68	.89	.84					
VE$_7^A$	.74	.90	.86					
GL$_8^A$	.74	.87						.90
GB$_{10}^A$	.36					.16	.15	
MA$_{10}^A$	.40					.18	.15	
GW$_{10}^A$	.40					.16	.12	
ER$_{10}^A$	.30					.15	.11	
NA$_{10}^A$	.44					.17	.14	

Anmerkung. P_i = Itemschwierigkeit; P_{pi} = Paarweise Itemschwierigkeit; Wertesysteme: GB = Geborgenheit; MA = Macht; GW = Gewissheit; ER = Erfolg; GL = Gleichheit; VE = Verstehen; NA = Nachhaltigkeit; A = Annäherung; Zahlen indizieren die Blocknummer.

sind. Ganz allgemein kann den Itemschwierigkeiten tendenziell entnommen werden, dass die Items der Wertesysteme ***Gleichheit***^{A} und ***Verstehen***^{A} zu leicht und Items der Wertesysteme ***Geborgenheit*** und ***Nachhaltigkeit*** zu schwer zu bevorzugen sind. Dieselbe Tendenz spiegelt sich auch in den Testwertverteilungen (Tabelle 8) wieder.

Bei der MVSQ^V-Skala ist gemessen an Itemvarianzen (Tabelle 10) und Trennschärfen (Tabelle 11) Block 8 derjenige mit den schwächsten Werten. Den Itemvarianzen nach zu urteilen, sind die übrigen Blöcke in Ordnung. Bei den Trennschärfen fällt noch Block 9 mit einem sehr niedrigen Wert auf. Auf Konstruktebene ist möglicherweise ***Erfolg***^{V} und ***Verstehen***^{V} nicht ausreichend trennscharf, allerdings sind die Werte der Testwertverteilungen und Itemvarianzen dieser beiden Wertesysteme nicht auffällig.

Die Tabellen 15 und 16 zeigen, dass es deutlich weniger problematische und fragwürdige Items in der MVSQ^V-Skala verglichen mit der MVSQ^V-Skala gibt. Als einzige unbedingt zu überarbeitende Items sind hier ***Verstehen***^{V} in Block 2 sowie ***Geborgenheit***^{V} und ***Gleichheit***^{V} in Block 8 einer Revision zu unterziehen. Da Block 8 außerdem drei weitere Items von fragwürdiger Güte enthält, ist eventuell die Überarbeitung des gesamten Blocks sinnvoll. In Block 2 könnten ferner das Item ***Geborgenheit***^{V} und in Block 3 das Item ***Gewissheit***^{V} leichter formuliert werden.

Tabelle 15. Problematische Items der MVSQ^V-Skala.

		Paarweise Itemschwierigkeiten P_{pi}						
Item	P_i	GB^V	MA^V	GW^V	ER^V	GL^V	VE^V	NA^V
VE$_2^V$	.22	.16	.14	.18	.18			
GB$_8^V$	.32		.17	.18			.13	
GL$_8^V$	.19		.14	.12			.10	

Anmerkung. P_i = Itemschwierigkeit; P_{pi} = Paarweise Itemschwierigkeit; Wertesysteme: GB = Geborgenheit; MA = Macht; GW = Gewissheit; ER = Erfolg; GL = Gleichheit; VE = Verstehen; NA = Nachhaltigkeit; V = Vermeidung; Zahlen indizieren die Blocknummer.

Außerdem sollte in der MVSQ^V-Skala die komplette Skala auf die Konsistenz der Items mit den Konstrukten überprüft werden, schließlich deuten die verhältnismäßig niedrigen Trennschärfen darauf hin, dass zumindest in einigen Blöcken, allen voran Block 8 und Block 9, die Items wenig mit den Testrohwerten korrelieren.

Konkret bietet sich die Vorgehensweise an, zuerst die einzelnen als problematisch identifizierten Items (Tabellen 13 und 15) anzupassen. Als zweites sollten die genannten problematischen Blöcke (Block 6 in MVSQ^A und Block 8 in MVSQ^V) verbessert werden und als drittes auf Konstruktebene die Itemformulierungen auf ihre inhaltliche Repräsentativität bzgl. der entsprechenden Konstrukte hin zu überprüfen (Items von ***Erfolg***^{A}). Zum Schluss können gegebenenfalls die fragwürdigen Items (Tabellen 14 und 16) unter Berücksichtigung der bereits veränderten Itemformulierungen angepasst werden.

Tabelle 16. Fragwürdige Items der MVSQ^V-Skala.

		Paarweise Itemschwierigkeiten P_{pi}						
Item	P_i	GB^V	MA^V	GW^V	ER^V	GL^V	VE^V	NA^V
GB$_2^V$	.64					.80	.84	
GW$_3^V$	.72					.81		.81
MA$_8^V$	.67	.83				.86		
GW$_8^V$	.65	.82				.88		
VE$_8^V$	.73	.87				.90		

Anmerkung. P_i = Itemschwierigkeit; P_{pi} = Paarweise Itemschwierigkeit; Wertesysteme: GB = Geborgenheit; MA = Macht; GW = Gewissheit; ER = Erfolg; GL = Gleichheit; VE = Verstehen; NA = Nachhaltigkeit; V = Vermeidung; Zahlen indizieren die Blocknummer.

5.3 Diskussion

In diesem Kapitel wurde eine deskriptivstatistische Evaluation der Items durchgeführt. Dazu wurden die klassischen Itemkennwerte Itemschwierigkeit, Itemvarianzen, Trennschärfen und Testwertverteilungen berechnet und um den Kennwert der paarweisen Itemschwierigkeit erweitert, der aus der ipsativen Natur der Daten entsprang. Alle Kennwerte wurden unter Berücksichtigung der Restriktionen, die aus der Ipsativität der Fragebogendaten resultieren, beurteilt. Daraus ergaben sich Empfehlungen darüber, welche Items überarbeitungswürdig sind. Diese Informationen wurden den Testentwicklern zur Verfügung gestellt, die daraufhin den Fragebogen einer Revision unterzogen.

Beim Vergleich der Subskalen wurde festgestellt, dass sich merklich weniger überarbeitungsbedürftige Items in der MVSQ^V-Skala als der MVSQ^A-Skala ergaben. Andererseits fielen insbesondere die Trennschärfekoeffizienten der MVSQ^V- deutlich niedriger als die der MVSQ^A-Skala aus. Allerdings konnten die Trennschärfekoeffizienten aufgrund der Ipsativität der Daten nur eingeschränkt anhand ihrer absoluten Werte beurteilt werden, da nicht klar ist, inwiefern diese Format-bedingt verzerrt sind. Gerade bezogen auf die Trennschärfen hat sich gezeigt, wo die Grenzen der klassischen Itemanalyse auf ipsative Instrumente liegen. Denn im vorliegenden Fall ist nicht erklärbar, warum die Trennschärfen der MVSQ^V-Skala niedriger, die Itemschwierigkeiten und Itemvarianzen jedoch deutlich besser als bei der MVSQ^A-Skala ausfielen. Die Trennschärfen wurden deshalb nicht auf Itemebene herangezogen, um die Qualität einzelner Items zu beurteilen, sondern lediglich auf Block- und Konstruktebene als Indikator verwendet.

Einschränkend muss bezogen auf die Stichprobe gesagt werden, dass es sich aufgrund des hohen Anteils Studierender (59.7%) nicht um eine bevölkerungsrepräsentative Stichprobe handelt. Auch der Vergleich der Testwertverteilungen, vor allem die der MVSQ^A-Wertesysteme,

zeigt, dass die Wertesystemausprägungen in der Stichprobe nicht gleichverteilt sind. Die in dieser Stichprobe überproportional ausgeprägten Wertesysteme ***Gleichheit***^{A} und ***Verstehen***^{A} passen inhaltlich besser zur Gruppe der Studierenden als z.B. zur Gruppe der im produzierenden Gewerbe arbeitenden Bevölkerung. Dies kann als weiterer Indikator für fehlenden Repräsentativität der Stichprobe gewertet werden. Eine größere Stichprobe und insbesondere mit einem höheren Anteil der arbeitenden Bevölkerung aus allen Altersschichten wäre an dieser Stelle wünschenswert, denn auch die Verteilung von Testwerten kann Itemkennwerte, wie z.B. die Trennschärfekoeffizienten beeinflussen (Kelava & Moosbrugger, 2012).

Der nächste Schritt in der Testentwicklung ist konsequenterweise die Untersuchung der überarbeiteten Fragebogenversion (Version 2) auf ihre Güte und insbesondere die Begutachtung der Veränderungen der Itemkennwerte.

Kapitel 6

Thurstonian IRT-Modelle

In Kapitel 3 wurde gezeigt, dass der MVSQ ipsative Daten produziert, die, wenn klassisch ausgewertet, weder für die Berechnung von verzerrungsfreien klassischen Reliabilitätskoeffizienten, noch für sinnvolle Untersuchungen zur Validität verwendet werden können. Ein Lösungsansatz, durch den die Daten von den problematischen Eigenschaften der Ipsativität befreit werden können, stellt die Thurstonian Item-Response-Theorie (TIRT) dar (siehe Kapitel 3.4). In diesem Kapitel werden nun zwei TIRT-Modelle an die Daten des MVSQ angepasst (pro Subskala ein Modell). Damit lassen sich einerseits die psychometrischen Eigenschaften des Instruments auf Itemebene beurteilen und andererseits Merkmalsausprägungen berechnen, die dann für die Validitätsuntersuchungen herangezogen werden können. Hierfür wurde die von den Testentwicklern überarbeitete Version des Fragebogens verwendet.

6.1 Methode

Zunächst ist anzumerken, dass die Daten, denen die TIRT-Modelle angepasst werden, aus Stichprobe II stammen. Wie in Kapitel 3.7 beschrieben, handelt es sich dabei um eine Stichprobe, die größtenteils aus Berufstätigen besteht. Für die Kennwerte der Stichprobe sei auf diesen Abschnitt verwiesen. Zudem sei angemerkt, dass die Daten mit der überarbeiteten Version des MVSQ (Version 2) erhoben wurden.

Das für dieses Kapitel zentrale Konzepte des TIRT-Ansatzes wurde bereits in Kapitel 3.4 eingeführt. In diesem Kapitel werden die TIRT-Modelle nun geschätzt. Dazu wird zunächst im Methodenteil auf die Vorgehensweise bei der Spezifizierung und Schätzung von TIRT-Modellen eingegangen. Im Ergebnisteil werden die Modellparameter berichtet und zum Abschluss erfolgt wie bei Brown und Maydeu-Olivares (2013) die Überprüfung der Plausibilität der so ermittelten Scores.

6.1.1 Modellschätzung mittels DWLS und MAP

Im Artikel „Fitting a Thurstonian IRT model to forced-choice data using M*plus* “ beschreiben Brown und Maydeu-Olivares (2012) detailliert die Vorgehensweise zur Spezifizierung und Schätzung eines TIRT-Modells unter Verwendung des Statistik-Programms M*plus* (Muthén & Muthén, 1998-2012), in dem die Modellparameter mit einem Diagonally Weighted Least Squares (DWLS) Schätzer und die Scores mit einem Maximum a Posteriori (MAP) Schätzer geschätzt werden. Die von Brown und Maydeu-Olivares (2011, 2013) skizzierten Voraussetzungen für die Anpassung eines TIRT-Modells an die MVSQ-Daten mit M*plus* waren im vorliegenden Fall jedoch nicht hinreichend erfüllt, da damit nur TIRT-Modelle geschätzt werden können, die Blöcke mit maximal vier Items haben. Die Blockgröße im MVSQ von sieben stellt damit ein Ausschlusskriterium für die Schätzung von TIRT-Modellen mit M*plus* dar. Der Grund dafür liegt darin, dass bei mehr als zwei Items pro Block lokale Abhängigkeiten entstehen, die für Blockgrößen mit mehr als vier Items nicht mehr ignoriert werden können. Laut A. Brown (persönliche Kommunikation, 21.10.2013) würden die lokalen Abhängigkeiten in einem MVSQ-TIRT-Modell auf Grund der additiven Formulierung der Testinformationsfunktion (Brown & Maydeu-Olivares, 2011) so viel Varianz erklären, dass die Testinformation massiv überschätzt würde und die Ergebnisse unbrauchbar wären. Um diese Problematik zu umgehen, kann eine Weiterentwicklung der Schätzung von TIRT-Modellen verwendet werden, die im folgenden Absatz vorgestellt wird.

6.1.2 Modellschätzung mittels Metaheuristic Stochastic Search

Ende 2014 veröffentlichten Zes et al. (2014) das R-Paket `kcirt`, dessen wesentliche Erweiterung zu DWLS und MAP in der Schätzmethode bestand. Zwar handelt es sich bei der Umsetzung von Zes et al. (2014) um eine leichte Generalisierung des TIRT-Modells nach Brown und Maydeu-Olivares (2011), aber abgesehen von einer anderen Terminologie unterscheidet sich das k-cube Thurstonian IRT-Modell (kcTIRT-Modell) von Brown und Maydeu-Olivares' TIRT-Modell nur dahingehend, dass die Faktorladungen nicht nur als eindimensionaler Vektor, sondern als Matrix konzipiert werden können. Dadurch könnten die Faktorladungen eines Items auf mehrere latente Traits modelliert werden. Ignoriert man diese Einstellungsmöglichkeit, dann sind kcTIRT- und TIRT-Modelle identisch.

Genau genommen werden zur Schätzung der TIRT-Modellparameter im `kcirt`-Paket zwei Algorithmen angewendet. Ein Expectation-Maximization Algorithmus, der Startwerte für den zweiten Algorithmus, die metaheuristische stochastische Suche (MSS) bereit stellt. Die MSS ist den Ridge Regression-Verfahren zuzuordnen und macht Gebrauch von der Shrinkage-Technik (Zes et al., 2014). Shrinkage ist eine Methode, die beim Schätzvorgang die Varianzen der Parameterschätzungen künstlich durch sogenannte *Tuning-Parameter* gegen Null schrumpft, um dadurch die Modellanpassung zu verbessern (James et al., 2014). Gerade in Regressionsmodellen

mit vielen korrelierten Variablen – wie dem TIRT-Modell, das im Vergleich zu einfachen Regressionsmodellen z.B. zusätzlich die Interkorrelationen der Uniquenesses beinhaltet (Brown & Maydeu-Olivares, 2011) – können Parameterschätzwerte schlecht bestimmt werden und leicht zu hohe Varianzen aufweisen (Hastie et al., 2013). Durch die Regularisations-Parameter (*Tuning-Parameter*), die die Varianzen der Schätzungen künstlich schrumpfen, kann das Ausmaß dieses Problems verringert werden und Parameter verlässlicher geschätzt werden (Hastie et al., 2013). Die Höhe der Tuning-Parameter bestimmt dabei proportional, wie stark die Modellparameter gegen Null geschrumpft werden, wodurch wiederum die Anpassung des Modells an die Daten verändert wird (Hastie et al., 2013). Die Qualität der Anpassung eines Modells hängt also von der Auswahl adäquater, d.h. zu den Daten und zum Modelldesign passender Tuning-Parameter ab.

Der Vorteil dieser Umsetzungen gegenüber Brown und Maydeu-Olivares' Vorgehen unter Verwendung des MAP-Schätzers besteht darin, dass Verzerrungen der Schätzergebnisse, die auf die strukturierten lokalen Abhängigkeiten der Items zurückgehen, durch die Anpassung des MSS-Schätzers (via der Tuning-Parameter) vermieden werden können (D. Zes, persönliche Kommunikation, 14.04.2014). Dadurch kann das oben geschilderte Problem bei der Schätzung eines TIRT-Modells mittels M*plus* umgangen werden. Andererseits müssen dazu zuerst passende Tuning-Parameter ermittelt werden. Dies stellt den ersten Teil dieses Kapitels dar.

Anstatt der in `kcirt` verwendeten Terminologie – Faktorladungen werden dort als „Hyperparameter“ und Uniquenesses als „System Shocks“(Zes et al., 2014, S. 1) bezeichnet – wird hier die von Brown und Maydeu-Olivares (2011) eingeführte Terminologie festgehalten, die auch schon in Kapitel 3.4 verwendet wurde und psychologischen Konventionen folgt.

6.1.3 Modellspezifizierung

Ähnlich der Umsetzung bei Brown und Maydeu-Olivares (2012) werden im `kcirt`-Rahmen die Anzahl und kodierte Richtung der Faktorladungen sowie die Anzahl der Utilities spezifiziert.[1] Pro Modell werden hier je 70 Faktorladungen und 70 Utilities eingestellt. Als Startwerte der Faktorladungen werden dem EM-Algorithmus dabei in engen Grenzen um 1 normalverteilte Zufallswerte zugegeben. Die Uniquenesses und deren Startwerte werden automatisch im `kcirt`-Paket erzeugt und müssen nicht extra angelegt werden. Außerdem kann spezifiziert werden, ob die Faktorladungsmatrix als diagonale oder als volle Matrix spezifiziert wird. Hier wird sie wie bei Brown und Maydeu-Olivares (2012) als diagonale Matrix spezifiziert. Des Weiteren muss festgelegt werden, welche Items in welchem Block auf welches Konstrukt laden und um welche Fragetypen es sich handelt (reines Ranking oder „Most/Least like me“). Beim MVSQ wurde ein reines Ranking gewählt und die Reihenfolge der Faktorladungen der Items auf die Wertesysteme wurde gemäß der Reihenfolge der Rohdaten eingestellt.

[1] Für Beispiele siehe die Paketdokumentation in R.

Darüber hinaus muss festgelegt werden, wie viele Iterationen der `kcirt`-Schätzalgorithmus MSS durchlaufen soll. Hierfür wurden 50 Iterationen gewählt. In Anhang B ist zu sehen, dass dies ausreichend viele Iterationen waren, damit die Schätzung konvergieren konnte.

Abschließend sei angemerkt, dass es sich bei den hier spezifizierten Modellen um „große" Modelle handelt, da pro Modell 210 Paarvergleiche modelliert werden. Brown und Maydeu-Olivares (2013) sprechen bereits bei 196 Paarvergleichen von einem „großen" Modell und begründen damit aufgrund zu hoher Rechenintensität das Fehlen von Standardfehler und Anpassungsgüte. Das Modell ist also zu komplex, um damit eine deterministische Berechnung der Anpassungsgüte durchzuführen. In solchen Fällen stellt häufig die stochastische Simulation eine plausible Alternative dar, um die Anpassungsgüte abzuschätzen (Dekking et al., 2005). Eine solche Herangehensweise wird im folgenden Abschnitt skizziert.

6.1.4 Anpassungsgüte

Da die MVSQ-TIRT-Modelle zu umfangreich sind, um ein Anpassungsmaß zu berechnen, dieses aber benötigt wird, um die Modellgüte zu beurteilen, wird es mittels Simulation approximativ geschätzt. Dazu kann laut D. Zes (persönliche Kommunikation, 06.11.2014) die bereits in Kapitel 3.6.2.1 vorgestellte, simulationsbasierte Vorgehensweise zur Reliabilitätsbestimmung verwendet werden, um damit die mittlere Abweichung der Schätzung (Root Mean Square Error) zu berechnen. Der Root Mean Square Error (RMSE) ist eine konventionelle Methode, um durchschnittliche Abweichungen und damit den Modellfit zu berechnen, wobei sich die Abweichung auf den Unterschied einer beobachteten und einer vorhergesagten Zufallsvariable bezieht (Barreto & Howland, 2006). Im vorliegenden Fall handelt es sich bei dem Modell, das an die Originaldaten angepasst wurde, um das beobachtete Modell und das Modell auf Basis der simulierten Daten beschreibt das vorhergesagte Modell. Letzteres wurde in Abbildung 2 als *geschätztes* Modell bezeichnet. Für die TIRT-Modelle gilt ferner, dass je ein RMSE für die relevanten Parameterschätzungen, d.h. für Utilities (RMSE_μ), Faktorladungen (RMSE_λ) und Varianz der Merkmalsscores (RMSE_η), und ebenso ein kombinierter Wert für die Anpassungsgüte des gesamten Modells berechnet werden kann (D. Zes, persönliche Kommunikation, 06.11.2014). Die entsprechenden Formeln zur Berechnung der RMSEs lauten:

$$\text{RMSE} = \sqrt{\text{RMSE}_\mu^2 + \text{RMSE}_\lambda^2 + \text{RMSE}_\eta^2} \tag{4}$$

$$\text{RMSE}_\mu = \frac{1}{u} \sum_{i=1}^{u} \frac{(\mu_i - \hat{\mu}_i)^2}{\mu_i^2 + \hat{\mu}_i^2} \tag{5}$$

$$\text{RMSE}_\lambda = \frac{1}{f} \sum_{i=1}^{f} \frac{(\lambda_i - \hat{\lambda}_i)^2}{\lambda_i^2 + \hat{\lambda}_i^2} \tag{6}$$

$$\text{RMSE}_\eta = \frac{1}{m} \sum_{i=1}^{m} \frac{|V(\eta_i) - V(\hat{\eta}_i)|}{|V(\eta_i) + V(\hat{\eta}_i)|} \tag{7}$$

wobei u, f und m die Anzahl der Utilities, Faktorladungen bzw. Merkmale beschreiben, die im vorliegenden Fall in beiden Modellen der Subskalen jeweils 70 für Utilities und Faktorladungen sowie 7 für die Merkmale betragen. Damit die Berechnung des kombinierten RMSE erfolgen kann, wurden die RMSEs der Modellparameter jeweils durch die Summe der Quadrate standardisiert (D. Zes, persönliche Kommunikation, 06.11.2014).

Bei der Interpretation des RMSE gilt, je kleiner der Wert ist, umso kleiner ist auch die Abweichung des geschätzten vom beobachteten Modell und umso besser ist die Anpassungsgüte des Modells bezogen auf die Daten. Es handelt sich dabei um ein relatives Maß der Anpassungsgüte, d.h. die absoluten Werte müssen im Kontext der Größenordnungen der entsprechenden Variablen interpretiert werden.

Zudem muss an dieser Stelle angefügt werden, dass die Höhe des RMSE in der hier dargestellten Berechnungsweise aufgrund der Zufallsgenerierung der Scores (Schritt 3 in Abbildung 2) einer gewissen Variabilität ausgesetzt ist. Um Effekte der Zufallsgenerierung auszugleichen, kann das der Monte-Carlo-Methode entstammende Prinzip der Replikation angewendet werden (Harwell et al., 1996). Das bedeutet, dass die Berechnung mehrfach durchgeführt wird und am Ende ein Mittelwert aus den Ergebnissen der Replikationen berechnet wird. Um darüber hinaus die Reproduzierbarkeit der Ergebnisse zu gewährleisten, empfiehlt es sich, bei der Erzeugung der zufallsgenerierten Daten die deterministische Zufallsgeneration anzuwenden (Schumacker, 2014).

6.1.5 Eignung der Tuning-Parameter

In Abschnitt 6.1.2 wurde beschrieben, dass der im `kcirt`-Paket eingesetzte Schätzer (metaheuristische stochastische Suche) Gebrauch von der Shrinkage-Technik macht, die durch sogenannte Tuning-Parameter gesteuert wird. Die Tuning-Parameter (TP) bestimmen dabei den Grad der Schrumpfung der geschätzten Variablen und beeinflussen dadurch maßgeblich die Anpassungsgüte einer Modellschätzung. Da bisher keine Musterbeispiele zur Anwendung der MSS auf ein Modell der hier erforderlichen Größenordnung veröffentlicht wurden[2], muss zunächst eine geeignete Kombination von TP gefunden werden. Die MSS weist im Gegensatz zu anderen Shrinkage-Verfahren wie normale Ridge-Regression oder LASSO-Regression (vgl. Hastie et al., 2013; James et al., 2014) die Besonderheit auf, dass darin drei Tuning-Parameter zum Einsatz kommen, die sich gegenseitig beeinflussen[3]. Diese drei Tuning-Parameter schrumpfen je entweder Utilities (μ), Faktorladungen (λ) oder Merkmalsausprägungen (η). Entsprechend der Modellnotation werden auch die Tuning-Parameter in der folgenden Analyse indiziert.

[2] Die einzig verfügbaren Beispiele können der Dokumentation des `kcirt`-Pakets entnommen werden. Sie behandeln Modelle mit maximal vier Merkmalen.

[3] Siehe hierzu den Quellcode der Funktion `kcirt.fitMSS`.

Zur Bestimmung einer geeigneten TP-Kombination ist laut D. Zes (persönliche Kommunikation, 06.11.2014) die Methode der Wahl, eine Vielzahl von TIRT-Modellen mit unterschiedlichen TP-Kombinationen zu schätzen und diese anhand der dazugehörigen RMSEs zu vergleichen. Dabei gilt, dass niedrigere RMSEs für einen besseren Modellfit und folglich für besser geeignete TP sprechen. Aufgrund der hohen Rechenintensität wird diese Vorgehensweise hier iterativ durchgeführt. D.h. es werden abwechselnd (je nach verfügbarer Rechenleistung) eine bestimmte Anzahl TIRT-Modelle geschätzt und deren RMSEs miteinander verglichen. Darauf basierend kann eine Entscheidung für die Auswahl der zu testenden TP in der folgenden Iteration getroffen werden. Sollten mehrere TP-Kombinationen vergleichbare RMSEs liefern, können wieder mehrere Replikationen desselben Schätzvorgangs durchgeführt werden. Dadurch werden Verzerrungen, die auf die Zufallsgenerierung der Daten zurückgehen, ausgeglichen und somit robustere Einschätzungen der Eignung der TP erreicht. Für den vorliegenden Fall können erneut zehn Replikationen als ausreichend angesehen werden, da in dieser Größenordnung einerseits die Effekte der Zufallsgenerierung deutlich sichtbar werden und sich zum anderen der Rechenaufwand in einem annehmbaren Rahmen hält.

Um Vergleichbarkeit und Reproduzierbarkeit der Modellschätzungen mit unterschiedlichen Tuning-Parametern zu gewährleisten, wurde bei allen Schätzvorgängen die deterministische Zufallsgenerierung (Schumacker, 2014) angewandt. Das heißt, es wurde für alle Modellschätzungen derselben *seed*-Wert verwendet, wodurch die zufallsgenerierten Zahlen stets sehr ähnliche Eigenschaften aufwiesen. Als *seed*-Wert wurde dafür 1 gewählt. Darüber hinaus wurde aus denselben Gründen auch im Falle des Replikationsvorgangs die deterministische Zufallsgenerierung eingesetzt, wobei für jede Replikation ein anderer *seed*-Wert verwendet wurde. Bei zehn Replikationen wurden dementsprechend *seed*-Werte von 1 bis 10 gewählt.

Da der MVSQ aus zwei unabhängigen Teilfragebögen besteht, d.h. zwei TIRT-Modelle geschätzt werden sollen, werden gegebenenfalls zwei TP-Kombinationen ermittelt. Idealerweise würde dieselbe TP-Kombination zur Schätzung beider Modelle verwendet werden, weil das den Vergleich der Modellparameter bzgl. ihrer absoluten Höhen erlauben würde.

6.1.6 Beurteilung der TIRT-Modellparameter

Bei der Beurteilung der Utilities und Faktorladungen können folgende Kriterien formuliert werden. Für die Utilities gilt einerseits, dass sie zwischen den Wertesystemen in vergleichbarer Höhe sein sollten, da dies die Vergleichbarkeit der Wertesysteme bzgl. der Beantwortungsschwierigkeit gewährleistet. Für die zehn Items eines Wertesystems gilt jedoch, dass es wünschenswert ist, dass sie in den Höhen variieren, denn dadurch würden unterschiedliche Stellen des Merkmalskontinuums abgebildet und die Wahrscheinlichkeit erhöht, das Wertesystem ungeachtet der Ausprägung der Testperson genauerer zu messen (Embretson & Reise, 2000).

Auch für die Faktorladungen können testtheoretische Kriterien abgeleitet werden. Je kleiner die Faktorladung, umso weniger unterscheidet das Item zwischen den latenten Traits. Prinzipiell sind also höhere Faktorladungen erstrebenswert. Ladungen kleiner Null sind komplett unbrauchbar, da sie keine Information mehr über das Wertesystem liefern (Embretson & Reise, 2000). Abgesehen davon, muss bei der Beurteilung der Utilities und Faktorladungen berücksichtigt werden, dass deren Ausprägungen maßgeblich durch die Höhen der Tuning-Parameter beeinflusst werden, dementsprechend keine absoluten Richtlinienwerte zur Verfügung stehen und sie in ihren absoluten Ausprägungen nur relativ zueinander interpretiert werden können.

Die Uniquenesses werden nicht berichtet und folglich nicht untersucht, da es erstens pro Modell in `kcirt` $210 \times 210 = 44100$ Werte gibt und diese Anzahl schlicht zu umfangreich ist, sowohl um sie zu berichten, als auch um sie sinnvoll zu sichten und zu interpretieren. Außerdem werden die wesentlichen Informationen des Modells bereits in den Faktorladungen und Utilities wiedergegeben (D. Zes, persönliche Kommunikation, 06.11.2014). Die Uniquenesses enthalten nur den nicht zuordenbaren Rest an Information und können deshalb vernachlässigt werden.

6.1.7 Analyse der TIRT-Scores

Eine ausführliche Suche in den einschlägigen Datenbanken psychologischer Literatur *PsycINFO*, *PSYNDEX* und *PsycBOOKS* sowie den alle Fächergruppen umfassenden Datenbanken *Web of Science / Social Sciences Citation Index* und *Google Scholar* nach Verwendung des `kcirt`-Ansatzes ergab keine Treffer. Da zudem der TIRT-Ansatz allgemein relativ neu ist und wenig Vergleichsstudien existieren, empfiehlt sich eine Überprüfung der Plausibilität der geschätzten TIRT-Scores, wie auch von den Entwicklern der Ansatzes durchgeführt (Brown & Maydeu-Olivares, 2013).

Dazu kann der direkte Vergleich der mittels KTT ausgewerteten (ipsativen) und mittels TIRT ausgewerteten (normativen) Scores herangezogen werden. Hierbei sollten sich die beiden Score-Arten sowohl auf der globalen Ebene (Skaleninterkorrelationen), als auch auf der individuellen Ebene (Profilkorrelationen und mittlere Ausprägungen) unterscheiden. Das Ausmaß der Unterschiede kann mit den von Brown und Maydeu-Olivares (2013) festgestellten Unterschieden verglichen werden. Zur sinnvolleren Vergleichbarkeit werden TIRT-und KTT-Scores dabei z-standardisiert.

Ipsative Scores sind davon gekennzeichnet, dass die Summe aller Merkmalsausprägungen (Profil) bei allen Personen gleich ist, da die Anzahl der Ränge bei allen Personen gleich ist. Dies gilt ebenso für den Mittelwert. Bezogen auf den MVSQ bedeutet das, dass die Summe der ipsativ ausgewerteten Wertesystemausprägungen bei allen Personen bei 210[4] und der Mittelwert bei 21 liegt. Außerdem gehen im FC-Format hohe Ränge eines Items zwangsweise mit niedrigeren Rängen auf einem anderen Item einher, weil nicht alle Items auf Rang 1 gerankt werden können.

[4] 10 Blöcke $\times$ 6! (Summe der Ränge).

Dies hat zum einen zur Folge, dass die klassisch ausgewerteten Items und damit auch die Scores negativ miteinander korrelieren. Im MVSQ beträgt diese Korrelation $r = -.17$ (siehe Kapitel 3.3.2). Zum anderen bedeutet dies, dass eine Person nicht auf allen Merkmalen gleichzeitig hohe Ausprägungen haben kann. Die normativen Scores sollten von diesen Restriktionen befreit sein. Es werden deshalb folgende Eigenschaften erwartet:

1. Die mittlere Korrelation aller Scores ist bei beiden Skalen nicht mehr auf $r = -.17$ festgesetzt.
2. Die mittlere Ausprägung aller Wertesystemausprägungen einer Person (Profil) kann sich zwischen Personen unterscheiden.
3. Es sind Profile möglich, in denen Person ausschließlich hohe oder ausschließlich niedrige Ausprägungen auf allen Wertesysteme aufweisen. Diese Profile werden als „all-high"- bzw. „all-low"-Profile bezeichnet (Brown & Maydeu-Olivares, 2013).

Diese Restriktionen können überprüft werden. Sind sie in den normativen Scores aufgehoben, spricht dies für die erfolgreiche Anwendung des TIRT-Ansatzes und die Qualität der normativen Scores.

6.2 Ergebnisse

Gemäß der drei Ziele dieses Kapitels gliedert sich auch der Ergebnisteil in drei Teile. Zuerst wird darstellt, wie eine geeignete Kombinationen von Tuning-Parametern (TP) zur Schätzung der TIRT-Modelle ermittelt wurde. Danach erfolgt der Bericht der geschätzten TIRT-Modelle und zum Abschluss die Analyse der TIRT-Scores hinsichtlich ihrer Plausibilität.

6.2.1 Geeignete Tuning-Parameter

Es erfolgt nun die Darstellung der iterativen Vorgehensweise bis zur finalen Auswahl der Tuning-Parameter. An dieser Stelle sei daran erinnert, dass für die Schätzung in `kcirt` drei Tuning-Parameter ermittelt werden müssen. Die Kürzel für die Tuning-Parameter der Utilities sind dabei TP_μ, für die Faktorladungen TP_λ und für die Scores TP_η.

Iteration 1

In der ersten Iteration wurden für die Tuning-Parameter TP_μ, TP_λ und TP_η Werte aus folgender Menge eingesetzt: $\text{TP} \in \{0.1; 0.3; 0.5; 0.7; 0.9\}$. Diese Auswahl der Werte orientiert sich an der Standardeinstellung des `kcirt`-Pakets, wo die Werte für die TP jeweils Zahlen mit einer Nachkommastelle sind. Aus den fünf möglichen Werten ergeben sich in Summe 125 TP-Kombinationen und dementsprechend wurden 125 TIRT-Modelle geschätzt. Die Modelle

wurden jeweils an die Daten der MVSQ^A-Skala angepasst. Tabelle 17 enthält die zehn TP-Kombinationen mit den niedrigsten RMSEs, also die zehn am besten passenden Modelle. Es kann ihr entnommen werden, dass die Modellgüte tendenziell mit abnehmenden TP zunimmt, wobei die Größen der λ- und η-TP einen größeren Einfluss auf die Modellgüte zu haben scheinen als die TP$_\mu$, da unter den fünf besten Modellen alle fünf möglichen Werte für TP$_\mu$ scheinbar mit nur geringen Änderungen der Anpassungsgüte einhergehen. TP$_\lambda$ und TP$_\eta$ führten bei dieser Auswahl an Werten zu besseren Schätzungen, je kleiner sie gewählt wurden.

Tabelle 17. Güte der TIRT-Modell Schätzung bei unterschiedlichen Tuning-Parametern (Iteration 1 und 2).

Iteration 1				Iteration 2			
TP$_\mu$	TP$_\lambda$	TP$_\eta$	RMSE	TP$_\mu$	TP$_\lambda$	TP$_\eta$	RMSE
0.1	0.1	0.1	0.41	0.01	0.01	0.09	0.07
0.3	0.1	0.1	0.43	0.01	0.09	0.01	0.08
0.5	0.1	0.1	0.50	0.01	0.03	0.05	0.09
0.7	0.1	0.1	0.57	0.01	0.05	0.03	0.09
0.9	0.1	0.1	0.62	0.01	0.07	0.03	0.09
0.1	0.5	0.7	0.87	0.01	0.05	0.01	0.09
0.1	0.7	0.7	0.91	0.01	0.03	0.03	0.09
0.5	0.9	0.9	0.92	0.01	0.01	0.05	0.09
0.5	0.5	0.3	0.92	0.01	0.03	0.07	0.09
0.3	0.3	0.7	0.93	0.01	0.05	0.05	0.10

Anmerkung. Die Tabelle enthält jeweils die zehn besten Modellschätzungen aus Iteration 1 und 2. TP = Tuning-Parameter; RMSE = Root Mean Square Error.

Die absoluten Größen der zehn besten RMSE-Werte liegen zwischen 0.41 und 0.93. Geht man die Berechnungsformel des RMSE zurück, dürften Unterschiede der Modellparameter zwischen den *wahren* und den *geschätzten* Werten auch im besten Modell noch bei über 0.5 liegen. Dies ist unter Berücksichtigung der zu erwartenden Ausprägungen der Utilities, Faktorladungen und Merkmalsvarianzen als relativ groß einzuschätzen. Der Tendenz entsprechend, dass mit kleineren Tuning-Parametern die RMSE besser wurden, wurden in der folgenden Iteration kleinere Tuning-Parameter untersucht.

Iteration 2

Für die zweite Iteration wurden Werte für TP $\in \{0.01; 0.03; 0.05; 0.07; 0.09\}$ gewählt. Auch hier ergeben sich somit 125 TP-Kombinationen und 125 TIRT-Modelle, die geschätzt werden müssen. Die komprimierten Ergebnisse dieser Iteration wurden auch in Tabelle 17 aufgelistet. Als erstes kann festgestellt werden, dass die zehn besten Schätzungen deutlich bessere RMSEs aufweisen. Das bedeutet, dass die Tuning-Parameter zwischen 0.01 und 0.09 bessere Modellschätzungen liefern können als Tuning-Parameter zwischen 0.1 und 0.9. Allerdings gilt die Tendenz, dass kleinere TP_λ und TP_η zu besseren Schätzungen führen, nicht mehr. Vielmehr scheint es eine Tendenz zu geben, dass wenn einer der beiden klein ist, der andere größer sein muss, um einen guten Modellfit zu erreichen. Über die TP_μ kann gesagt werden, dass sie bei den zehn besten Schätzungen alle den kleinstmöglichen Wert dieser Iteration haben. Insgesamt liegt in diesem Schritt die geeignetste TP-Kombination mit dem niedrigsten RMSE von 0.07 für $\text{TP}_\mu = 0.01$, $\text{TP}_\lambda = 0.01$ und $\text{TP}_\eta = 0.09$ vor, wobei sich die RMSEs deutlich geringfügiger unterscheiden als in Iteration 1. Bezogen auf die absoluten Werte des RMSE kann gesagt werden, dass die Abweichungen der wahren und geschätzten Variablen in einem Bereich um 0.2 bewegen, also deutlich weniger variieren als in Iteration 1.

Um den Zusammenhang von TP_λ und TP_η zu verdeutlichen, wurden diese in Abbildung 5 bei konstantem $\text{TP}_\mu = 0.01$ verbildlicht. Darin ist zu erkennen, dass wenn beide Werte für TP_λ und TP_η zu klein (z.B. 0.01) oder zu groß (z.B. 0.09) gewählt werden, die dazugehörigen RMSEs im Vergleich schlechter ausfallen. Bessere Anpassungsgüten werden anscheinend dann erreicht, wenn TP_λ und TP_η aufeinander abgestimmt sind. Um zu untersuchen, ob die Abstimmungsverhältnisse einem Muster folgen, wurde für die folgende Iteration TP_μ festgehalten, um die Durchführung der Tests für größere Anzahlen von Werten für TP_λ und TP_η zu vereinfachen.

Iteration 3

In Iteration 3 wurde nun TP_λ bei 0.01 fixiert und die Werte für TP_λ und TP_η aus allen Kombinationen der Wertemenge $\{0.001; 0.005; 0.0075; 0.01; 0.02; 0.03; 0.04; 0.05; 0.06; 0.07; 0.08; 0.09; 0.1; 0.11; 0.12\}$ gewählt. Insgesamt ergab das 225 Modelle. Die dazugehörigen RMSEs wurden in Abbildung 6 wieder farblich und nach TP-Kombination geordnet dargestellt. Dabei tritt das Muster der Abstimmung zwischen TP_λ und TP_η, das sich in Abbildung 5 angedeutet hat, wesentlich deutlicher hervor und zeigt einen leicht kurvilinearen Zusammenhang zwischen TP_λ und TP_η. Zwar liegen sehr viele RMSEs in einem sehr niedrigen Bereich um 0.1, doch zwei RMSEs fallen in der farblichen Darstellung auf. Die hellsten Felder treten bei der Kombination der Parameter $\text{TP}_\lambda = 0.01$ und $\text{TP}_\eta = 0.09$ sowie bei $\text{TP}_\lambda = 0.0075$ und $\text{TP}_\eta = 0.1$ auf. Die erste Kombination war schon in Iteration 2 die Beste. Tabelle 18 zeigt die entsprechenden Zahlenwerte der RMSEs und es kann festgestellt werden, dass keine der untersuchten TP-Kombinationen zu einem besseren RMSE im Vergleich zu Iteration 2 geführt hat. Allerdings

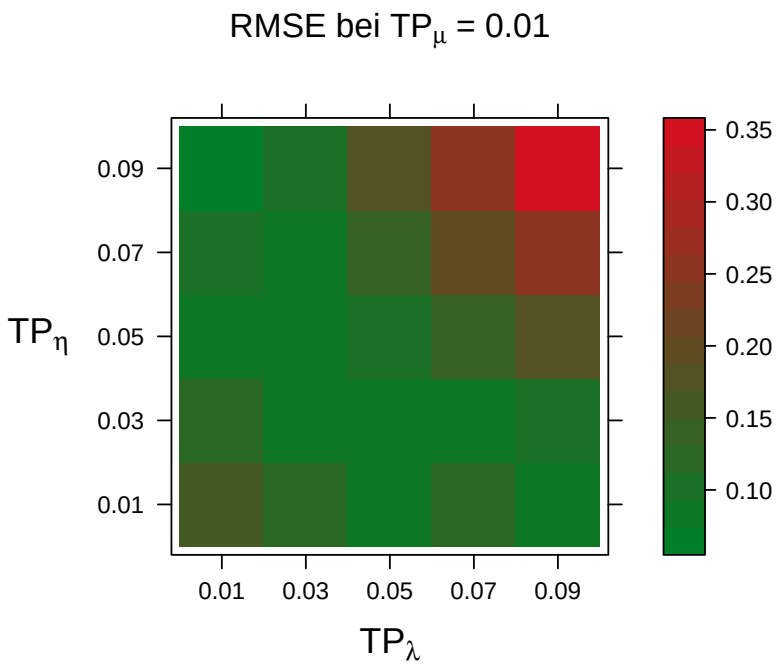

Abbildung 5. Zusammenhang von TP_λ und TP_η bei konstantem $TP_\mu = 0.01$. Die farbliche Schattierung repräsentiert dabei die Ausprägung des zur TP-Kombination gehörigen RMSE. Grün indiziert niedrige Werte, rot hohe Werte, gemäß der am rechten Rand angeführten Skala. TP = Tuning-Parameter; RMSE = Root Mean Square Error.

liegen die besten zehn Parameter-Kombinationen in einer sehr ähnlichen Größendimension (RMSE zwischen 0.0734 und 0.0776) wie der beste RMSE aus Iteration 2. Zudem fällt auf, dass bei acht der zehn besten Kombinationen TP_λ niedriger als TP_η ist.

Iteration 4

Die zehn besten Modellschätzungen aus Iteration 3 traten bei einem TP_μ von 0.01 auf, dem kleinsten untersuchten Wert. Deshalb kann vermutet werden, dass kleinere TP_μ zu noch adäquateren Modellschätzungen führen. In der vierten Iteration werden als Folge noch kleinere Werte für TP_μ untersucht, wobei die Werte für TP_λ und TP_η in relativ engen Grenzen um die beste Kombination dieser beiden Werte aus Iteration 3 gewählt werden: $TP_\mu \in \{0.001; 0.002; 0.003; 0.004; 0.005; 0.006; 0.007; 0.008; 0.009; 0.010\}$, $TP_\lambda \in \{0.0075; 0.0100; 0.0125\}$ und $TP_\eta \in \{0.09; 0.11; 0.10\}$. Mit jeweils drei Werten für TP_λ und TP_η und zehn Werten für TP_μ ergeben sich so 90 TP-Kombinationen, mit denen je ein TIRT-Modell angepasst wird.

Tabelle 18 zeigt die Ergebnisse dieser Iteration, die zu dem Rückschluss führen, dass kleinere Tuning-Parameter der Utilities (TP_μ) bessere Schätzergebnisse liefern. Auch die RMSE haben sich bei den 10 besten Modellen verbessert und liegen mit 0.04 bei der besten TP-Kombination ($TP_\mu = 0.002$, $TP_\lambda = 0.0075$ und $TP_\eta = 0.09$) noch mal deutlich besser als die zuvor erzielten Anpassungsgüten.

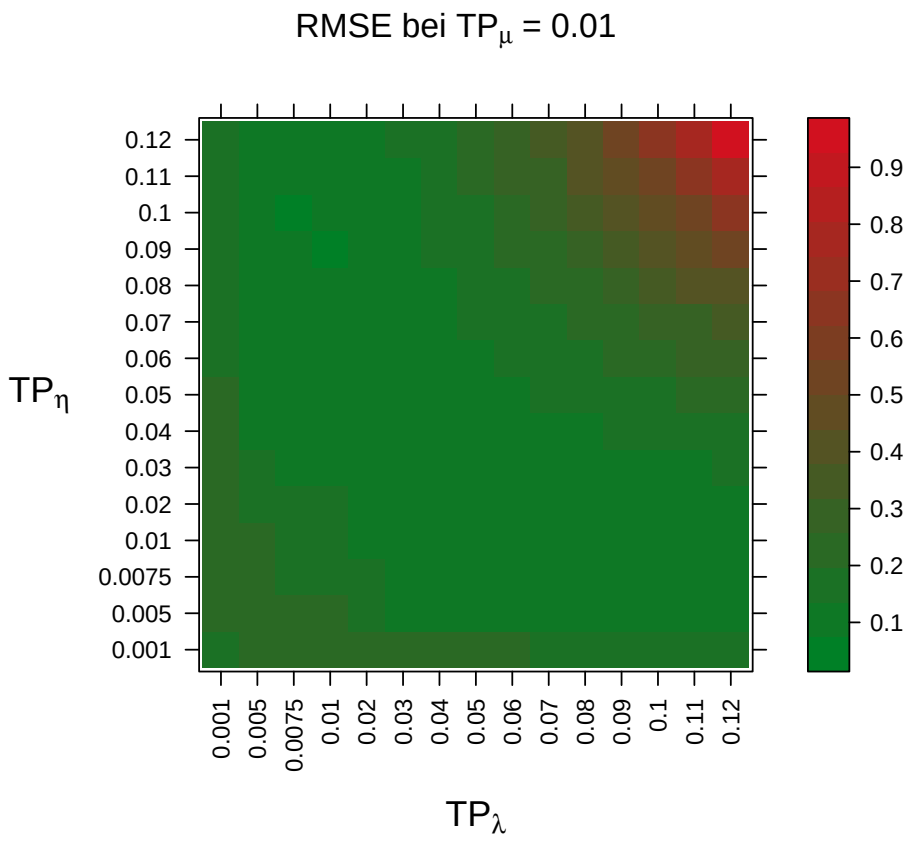

Abbildung 6. Detaillierter Zusammenhang von TP_λ und TP_η bei konstantem $TP_\mu = 0.01$. Die farbliche Schattierung repräsentiert dabei die Ausprägung des zur TP-Kombination gehörigen RMSE. Grün indiziert niedrige Werte und rot hohe Werte, gemäß der am rechten Rand angeführten Skala. TP = Tuning-Parameter; RMSE = Root Mean Square Error.

Da die Berechnungsweise des RMSE simulationsbasiert ist, für alle Simulationsschritte der *seed*-Wert von 1 gesetzt wurde und mehrere TP-Kombinationen zu ähnlichen RMSEs führten, wurde zur Kontrolle möglicher Effekte der Zufallsgenerierung ein weitere Überprüfung der RMSE durchgeführt. Es wurden zehn Replikationen mit zehn verschiedenen *seeds* für die besten drei TP-Kombinationen geschätzt und daraus der mittlere RMSE sowie die Spanne der RMSEs berechnet. Die Ergebnisse dieses Analyse sind in Tabelle 19 dargestellt. Darin können zwei Beobachtungen gemacht werden: Erstens sind die Mittelwerte der RMSEs etwas höher als die in Iteration 4 ermittelten RMSEs und zweitens sind die Unterschiede sehr gering, so dass im Endeffekt jede der drei besten TP-Kombinationen verwendet werden könnte und alle drei zu vergleichbaren Ergebnissen führen würden. Die finale Wahl der Tuning-Parameter viel auf die Kombination $TP_\mu = 0.001$, $TP_\lambda = 0.0075$ und $TP_\eta = 0.09$, da bei dieser Kombination die Spanne der RMSEs, bei gleichem Mittelwert mit der ersten Kombination, am engsten war.

Tabelle 18. Güte der TIRT-Modell Schätzung bei unterschiedlichen Tuning-Parametern (Iteration 3 und 4).

Iteration 3				Iteration 4			
TP_μ	TP_λ	TP_η	RMSE	TP_μ	TP_λ	TP_η	RMSE
0.01	0.0100	0.0900	0.073	0.002	0.0075	0.09	0.036
0.01	0.0075	0.1000	0.074	0.001	0.0075	0.09	0.037
0.01	0.0100	0.1000	0.075	0.001	0.0075	0.10	0.038
0.01	0.0200	0.0500	0.075	0.002	0.0075	0.11	0.042
0.01	0.0100	0.1200	0.076	0.002	0.0075	0.10	0.043
0.01	0.0300	0.0300	0.077	0.003	0.0075	0.10	0.043
0.01	0.0400	0.0200	0.077	0.003	0.0100	0.09	0.045
0.01	0.0800	0.0100	0.077	0.002	0.0100	0.10	0.046
0.01	0.1000	0.0075	0.077	0.004	0.0075	0.11	0.047
0.01	0.0500	0.0200	0.078	0.002	0.0100	0.09	0.048

Anmerkung. Die Tabelle enthält jeweils die zehn besten Modellschätzungen aus Iteration 3 und 4. TP = Tuning-Parameter; RMSE = Root Mean Square Error.

Tabelle 19. Vergleich der RMSEs über 10 Replikationen bei den drei besten TP-Kombinationen.

Tuning-Parameter			RMSE		
μ	λ	η	M	min	max
0.002	0.0075	0.09	0.049	0.034	0.068
0.001	0.0075	0.09	0.049	0.036	0.063
0.001	0.0075	0.10	0.055	0.039	0.075

Anmerkung. RMSE = Root Mean Square Error.

Iteration 5

Im folgenden Abschnitt werden geeignete Tuning-Parameter für das $MVSQ^V$-TIRT-Modell gesucht. Es ist naheliegend, zunächst die Tuning-Parameter des Annäherungsmodells zu verwenden, da die Subskalen dasselbe Format aufweisen. Wie zuvor wird der RMSE als Maß der Eignung der Tuning-Parameter berechnet.

Wenn für beide Modelle dieselben Tuning-Parameter verwendet würden, hätte dies den Vorteil, dass die Modellparameter jeweils im selben Ausmaß und im gleichen Verhältnis zueinander

geschrumpft würden. Dies würde den Vergleich der absoluten Höhen der Modellparameter zwischen den Modellen plausibel machen. Würden unterschiedliche Tuning-Parameter zur Schätzung der Modelle verwendet, wären lediglich Aussagen innerhalb der Modelle sinnvoll, da der Einfluss unterschiedlicher Tuning-Parameter auf das Ausmaß der Unterschiedlichkeit der Modellparameter nicht abzuschätzen ist.

Der RMSE des MVSQ^V-Modells, das mit denselben Tuning-Parametern wie das MVSQ^A-Modell geschätzt wurde (TP$_\mu = 0.001$, TP$_\lambda = 0.0075$ und TP$_\eta = 0.09$), betrug im Mittel über zehn Replikationen 0.074 (von 0.031 bis 0.108). Zwar ist der Mittelwert der RMSE etwas höher als der Vergleichswert der MVSQ^A-Skala bei denselben TPs, befindet sich jedoch in derselben Größenordnung.

Für beide TIRT-Modelle kann somit dieselbe Parameterkombination gewählt werden, obgleich die Anpassungsgüte beim MVSQ^V mit diesen Tuning-Parametern etwas schlechter ist. Die gewählten Tuning-Parameter passen demnach besser zu den MVSQ^A-Daten, was nicht verwunderlich ist, da die iterative Suche nach einer geeigneten Kombination von den MVSQ^A-Daten ausging. Zum Zweck der besseren Vergleichbarkeit wurde diese Parametereinstellung jedoch auch für das MVSQ^V-Modell beibehalten.

6.2.2 MVSQ Thurstonian IRT-Modelle

Mit den zuvor ermittelten Tuning-Parametern wurden nun die TIRT-Modelle an die Daten beider MVSQ-Skalen angepasst. Die entsprechenden Modellparameter werden im Folgenden präsentiert und analysiert. Die Tabellen 20 und 21 zeigen die Utilities der beiden Subskalen und Tabellen 22 und 23 enthalten die Faktorladungen.

Bei den Annäherungs-Utilities fallen besonders die Wertesysteme ***Geborgenheit***^{A} und ***Verstehen***^{A} auf, da sie deutlich extremere Werte aufweisen als die übrigen Wertesysteme. Die ***Geborgenheit***^{A}-Utilities sind im Schnitt besonders niedrig (M = -0.71), d.h. schwer zu beantworten, die ***Verstehen***^{A}-Utilities besonders hoch (M = 0.93), d.h. leicht zu bevorzugen. Auch für ***Gleichheit***^{A} (M = 0.5) und ***Nachhaltigkeit***^{A} (M = -0.43) sind die Utilities im Schnitt extremer ausgeprägt, was auch am Vergleich mit den Vermeidungs-Utilities festgemacht werden kann, denn hier liegt die höchste bzw. niedrigste durchschnittliche Utility pro Wertesystem deutlich näher an Null (vgl. Tabellen 20 und 21). Insgesamt kann daraus der Schluss gezogen werden, dass die Utilities der MVSQ^A-Skala im Vergleich der Wertesysteme sehr ungleich sind, was bei der nächsten Revision der Items beachtet werden sollte. Die Items von ***Geborgenheit***^{A} und ***Nachhaltigkeit***^{A} sollten dabei leichter und die von ***Verstehen***^{A} und ***Gleichheit***^{A} schwerer gemacht werden.

Andererseits ist zu sagen, dass die Varianzen der Utilities von den vier Annäherungswertesystemen von ***Geborgenheit***^{A} mit 0.35, ***Gewissheit***^{A} mit 0.38, ***Erfolg***^{A} mit 0.48 und ***Gleichheit***^{A} mit 0.25 deutlich höher als bei den Vermeidungs-Utilities sind. Im Vergleich

Tabelle 20. Utilities der MVSQ^{A}-Skala.

	Block										
	1	2	3	4	5	6	7	8	9	10	M
GB^{A}	-1.21	-0.66	-0.22	-1.63	-0.57	-1.54	0.25	-0.62	-0.57	-0.34	-0.71
MA^{A}	0.04	-0.43	0.63	0.36	0.56	-1.09	-0.37	0.18	-0.06	-0.50	-0.07
GW^{A}	-0.19	-0.40	0.46	-0.74	-0.80	0.26	0.39	0.37	-1.39	-0.13	-0.22
ER^{A}	0.06	-0.51	-1.59	0.70	0.58	0.40	0.45	0.41	0.51	-0.17	0.08
GL^{A}	1.00	0.77	0.75	0.60	0.78	0.70	-0.31	-0.44	0.86	0.29	0.50
VE^{A}	0.80	1.04	0.86	1.04	0.48	1.26	0.73	0.57	0.92	1.55	0.93
NA^{A}	-0.68	0.20	-0.63	-0.36	-0.84	0.31	-0.99	-0.69	-0.34	-0.33	-0.43

Anmerkung. GB = Geborgenheit; MA = Macht; GW = Gewissheit; ER = Erfolg; GL = Gleichheit; VE = Verstehen; NA = Nachhaltigkeit; A = Annäherung; M = Mittelwert.

Tabelle 21. Utilities der MVSQ^{V}-Skala.

	Block										
	1	2	3	4	5	6	7	8	9	10	M
GB^{V}	0.83	0.23	0.15	0.30	-0.15	0.77	0.02	0.14	0.76	0.28	0.33
MA^{V}	-0.57	0.22	0.54	0.28	0.29	-1.19	0.03	0.46	-0.27	-0.21	-0.04
GW^{V}	0.61	-0.11	0.49	-0.02	0.58	0.25	-0.41	0.18	0.60	0.12	0.23
ER^{V}	0.06	0.54	-0.25	-0.70	0.13	-0.69	0.15	-0.10	0.03	0.29	-0.06
GL^{V}	0.22	-0.47	-0.13	0.02	0.49	0.34	0.01	-0.20	-0.44	0.43	0.03
VE^{V}	-0.99	-0.26	-0.22	0.29	-0.88	0.32	-0.00	0.20	-0.85	-0.35	-0.27
NA^{V}	-0.14	-0.35	-0.58	-0.16	-0.36	0.35	0.23	-0.77	0.41	-0.38	-0.18

Anmerkung. GB = Geborgenheit; MA = Macht; GW = Gewissheit; ER = Erfolg; GL = Gleichheit; VE = Verstehen; NA = Nachhaltigkeit; V = Vermeidung; M = Mittelwert.

schwanken die mittleren Varianzen der Utilities der Vermeidungssysteme zwischen 0.12 und 0.29. Dies zeigt, dass die Vermeidungs-Items die Vermeidungssysteme weniger breit bzgl. des Merkmalskontinuums messen.

Bei den Faktorladungen der Annäherungswertesysteme fallen zwei Blöcke auf. Block 6 hat deutlich niedrigere Faktorladungen und Block 10 weist drei Ladungen auf, die deutlich unter dem Durchschnitt liegen. Im Vergleich zwischen den Wertesystemen liegen alle Faktorladungen

ungefähr auf einem Niveau, wobei die Faktorladungen von ***Gewissheit***^{A} deutlich besser und die von ***Nachhaltigkeit***^{A} ein wenig besser als die der anderen Wertesysteme sind.

Tabelle 22. Faktorladungen der MVSQ^{A}-Skala.

	Block										
	1	2	3	4	5	6	7	8	9	10	M
GB^{A}	1.37	0.99	2.47	0.58	1.45	2.06	0.50	2.32	1.44	0.63	1.38
MA^{A}	2.02	0.62	2.38	0.84	1.92	0.80	0.99	1.55	1.75	1.29	1.42
GW^{A}	1.98	2.06	1.43	1.55	1.70	1.15	2.12	1.78	1.62	0.86	1.62
ER^{A}	1.79	1.28	1.51	2.26	0.64	0.53	2.08	1.31	1.73	1.51	1.46
GL^{A}	0.70	2.86	1.06	0.94	0.71	0.99	1.80	0.95	1.41	1.54	1.30
VE^{A}	2.62	1.72	0.67	1.61	1.12	0.67	1.34	1.19	1.89	0.80	1.36
NA^{A}	0.71	0.88	1.41	1.82	1.34	0.29	2.01	0.94	2.83	2.22	1.44
M	1.60	1.49	1.56	1.37	1.27	0.93	1.55	1.44	1.81	1.26	1.43

Anmerkung. GB = Geborgenheit; MA = Macht; GW = Gewissheit; ER = Erfolg; GL = Gleichheit; VE = Verstehen; NA = Nachhaltigkeit; A = Annäherung; M = Mittelwert.

Tabelle 23. Faktorladungen der MVSQ^{V}-Skala.

	Block										
	1	2	3	4	5	6	7	8	9	10	M
GB^{V}	1.77	1.40	1.25	1.01	1.30	2.22	1.24	1.11	1.48	1.75	1.45
MA^{V}	1.96	1.04	1.18	1.01	1.66	1.96	1.65	0.34	2.65	1.06	1.45
GW^{V}	1.38	1.43	1.19	0.96	2.44	2.10	1.37	1.08	1.90	2.14	1.60
ER^{V}	2.45	-0.05	1.51	1.53	2.32	0.48	-0.14	0.57	2.52	-0.50	1.07
GL^{V}	2.25	1.65	1.38	0.77	1.71	1.81	1.17	0.68	0.93	1.61	1.40
VE^{V}	2.67	1.06	1.20	1.11	0.75	0.96	1.39	0.41	2.46	1.54	1.36
NA^{V}	2.10	2.05	1.27	1.45	1.16	0.74	1.28	1.14	1.04	1.21	1.34
M	2.08	1.23	1.28	1.12	1.62	1.47	1.14	0.76	1.85	1.26	1.38

Anmerkung. GB = Geborgenheit; MA = Macht; GW = Gewissheit; ER = Erfolg; GL = Gleichheit; VE = Verstehen; NA = Nachhaltigkeit; V = Vermeidung; M = Mittelwert.

Bei den MVSQ^V-Faktorladungen fallen im Vergleich zwischen den Merkmalen ebenso zwei Merkmale auf. Die Ladungen des ***Gewissheit***^{V}-Wertesystems sind auch hier deutlich höher. Zudem liegen die Ladungen des ***Erfolg***^{V}-Wertesystems deutlich unter dem Durchschnitt. Im Vergleich zwischen den Fragen fallen Block 1 und Block 9 mit deutlich überdurchschnittlichen Ladungen sowie Block 8 mit deutlich unterdurchschnittlichen Ladungen auf.

Utilities und Faktorladungen sind größtenteils in einer akzeptablen Größenordnung, jedoch kann empfohlen werden, dass die Items des Instruments weiter angepasst werden. Vor allem sollten die beiden schwächeren Blöcke überarbeitet werden, die im Vergleich deutlich niedrigere Faktorladungen aufweisen. Bei der Annäherungsskala sollten zudem die einseitigen Utilities berücksichtigt werden und Items so umformuliert werden, dass Items der Wertesysteme ***Geborgenheit***^{A} und ***Nachhaltigkeit***^{A} leichter zu bevorzugen werden und Items von ***Gleichheit***^{A} und ***Verstehen***^{A} schwieriger werden. Andererseits ist zu berücksichtigen, dass es sich bei der Stichprobe um ein Convenience Sample handelt und diese deshalb nicht repräsentativ ist. Es ist möglich, dass die Utilities die durchschnittlichen Ausprägungen der Wertesysteme in dieser Stichprobe widerspiegeln und sich die Utilities verschieben, wenn zusätzlich Bevölkerungsschichten in der Stichprobe enthalten wären, die vermeintlich höhere Ausprägungen der Wertesysteme ***Geborgenheit***^{A} und ***Nachhaltigkeit***^{A} sowie niedrigere Ausprägungen der Wertesysteme ***Gleichheit***^{A} und ***Verstehen***^{A} hätten.

6.2.3 Plausibilitätsanalyse der TIRT-Scores

In diesem Abschnitt wird nun die Plausibilität der Scores anhand der eingangs formulierten Kriterien überprüft. Es handelt sich dabei gleichzeitig um einen Vergleich der Scoring-Methoden, wobei hier nur interessiert, ob die mittels `kcirt` modellierten und geschätzten (normativen) Scores plausibel sind.

Die mittleren Interkorrelationen der Wertesysteme beträgt bei der MVSQ^A-Skala $r = -.02$ und bei der MVSQ^V-Skala $r = -.03$. Beide Werte unterscheiden sich drastisch von der mittleren Korrelation der ipsativen Scores, die formatbedingt bei $r = -0.17$ liegt. Diese Unterschiede sprechen somit für die Plausibilität der normativen Scores, da sie von der Einschränkung der festgelegten durchschnittlichen Skaleninterkorrelation von -.17 befreit sind.

Zum Vergleich der normativen und ipsativen Profile wurden für jedes Profilpaar (TIRT-und KTT-Profil) sowohl Produkt-Moment-Korrelationen (r) als auch Spearmans Rangkorrelationskoeffizienten (ρ) berechnet. Letztere geben Auskunft darüber, inwiefern sich die Rangfolgen der Wertesytempräferenz zwischen den Scoring-Methoden unterscheiden.

Bzgl. der Annäherungsprofile reichen die Produkt-Moment-Korrelationen von $r = .63$ bis $r = 1$ (M = .97, Md = .98, SD = .04) und die Rangkorrelationen von $\rho = .21$ bis $\rho = 1$ (M = .93, Md = .96, SD = .08). In ähnlichen Größenordnungen bewegen sich die Korrelationen der Vermeidungsprofile: die Bravais-Pearson-Korrelationen schwanken zwischen $r = .47$ und

$r = 1$ (M $= .96$, Md $= .98$, SD $= .05$), die Spearman-Korrelationen zwischen $\rho = .32$ und $\rho = 1$ (M $= .92$, Md $= .96$, SD $= .09$).

Abbildung 7 zeigt die Häufigkeitsverteilungen der z-standardisierten Profilmittelwerte der beiden Subskalen je Scoring-Methode. Es ist deutlich zu sehen, dass die mittleren Profilausprägungen der KTT-Scores sehr nahe bei Null liegen (Mittelwerte der Annäherungsprofile reichen von -0.13 bis 0.1, SD $= 0.04$; Vermeidungsprofile von -0.07 bis 0.08, SD $= 0.03$). Wären die ipsativen Scores nicht z-standardisiert, würde der Mittelwert exakt bei 0 liegen. Die TIRT-Scores verteilen sich hingegen deutlich breiter (Mittelwerte der Annäherungsprofile streuen von -1.01 bis 1.21, SD $= 0.36$; Vermeidungsprofile von -1.24 bis 0.91, SD $= 0.35$).

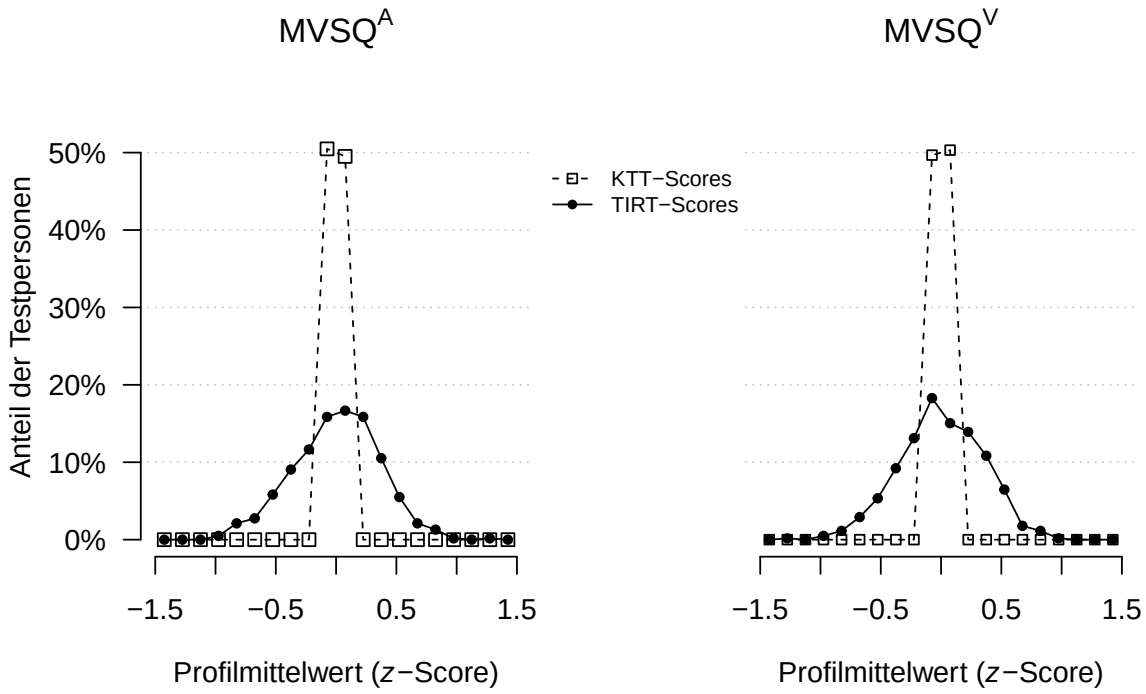

Abbildung 7. Verteilungen der Profilmittelwerte.

Auf den Abbildungen 8 und 9 ist zu sehen, dass es in beiden normativ gescorten Subskalen sowohl „all-high"-, als auch „all-low"-Profile gab. Gut zu erkennen ist auch, dass es sich bei den KTT-Profilen um ipsative Profile handelt, da Werte eines Wertesystems über 0 automatisch andere Wertesysteme innerhalb des Profils erzwingen, die unter 0 liegen. Bei den TIRT-Profilen ist diese Limitierung aufgehoben.

Alle untersuchten Kriterien zeigen deutliche Unterschiede zwischen den beiden Scoring-Methoden und die im methodischen Teil dieses Kapitels formulierten erwarteten Eigenschaften der TIRT-Scores konnten bestätigt werden. Auf der allgemeinsten Ebene zeigen die mittleren Merkmalskorrelationen der TIRT-Scores, dass diese von den Restriktionen der Ipsativität befreit

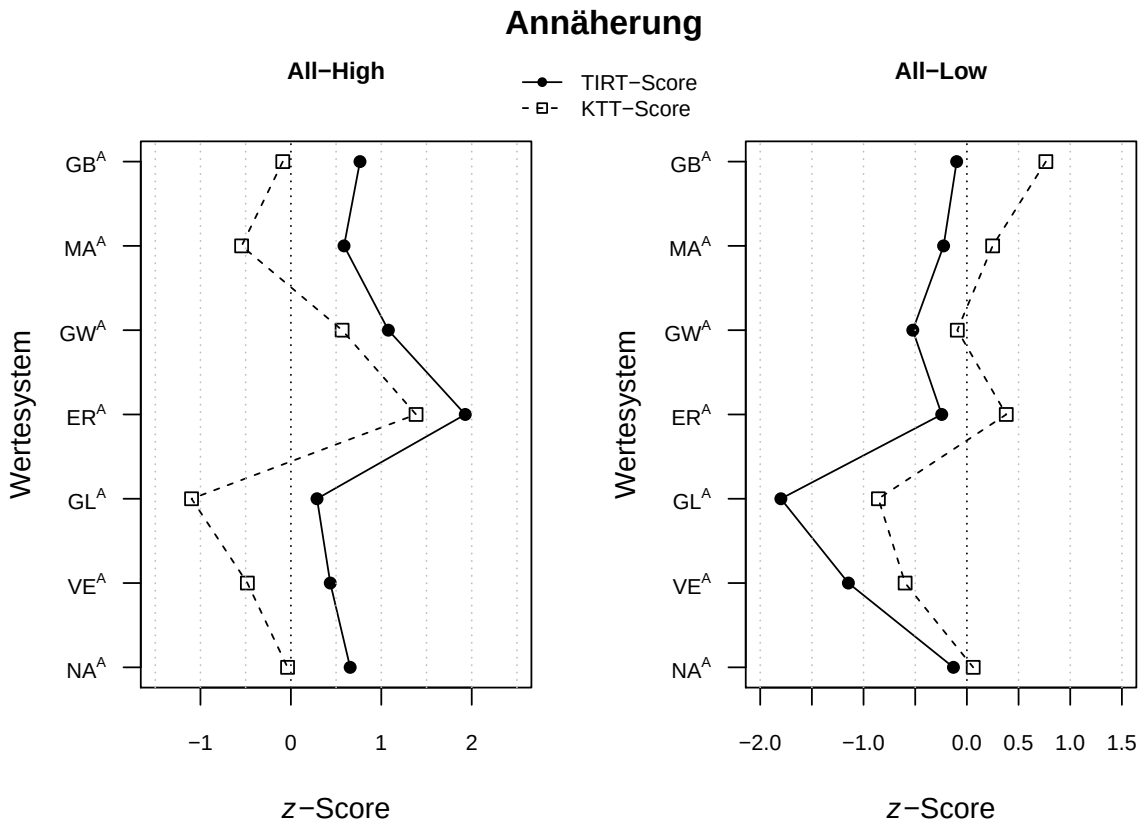

Abbildung 8. 'All-high' und 'all-low' Profile der Annäherungs-Skala.

sind, also nicht mehr zwangsweise durch das Fragebogendesign festgesetzt sind, sondern durch die TIRT-Modellierung freier bestimmt werden können.

Des Weiteren zeigen sowohl die Pearson- als auch die Spearman-Korrelationen, dass sich die TIRT-Scores teils erheblich von den KTT-Scores unterscheiden. Die hier berichteten mittleren Korrelationen und ihre Standardabweichungen entsprechen in etwa den Korrelationswerten, die Brown und Maydeu-Olivares (2013) in einer ähnlichen Vergleichsstudie zwischen TIRT- und KTT-Scores feststellen konnten. Ebenso zeigen die Verteilungen der Profilmittelwerte ein sehr ähnliches Ausmaß der Dimensionen, wie die entsprechenden Verteilungen in Brown und Maydeu-Olivares's (2013) Studie.

Ferner wurde gezeigt, dass es in der vorhandenen Stichprobe TIRT-Profile gibt, in denen alle Werte über bzw. alle Werte unter Null liegen, also alle Wertesysteme entweder hoch oder niedrig ausgeprägt sind. Zwar folgen die Höhen und Reihenfolgen der Scores bei beiden Scoring-Methoden tendenziell ähnlichen Mustern, dennoch haben die zuvor berechneten Rangkorrelationen gezeigt, dass es einige sehr stark abweichende Profile gibt.

Die hier berichteten Merkmale der TIRT-Scores weisen im Vergleich mit den KTT-Scores darauf hin, dass das TIRT-Scoring mittels `kcirt` zu plausiblen Ergebnissen führt. In Summe erscheinen die Befunde ausreichend, um daraus zu schließen, dass die mittels `kcirt` ermittel-

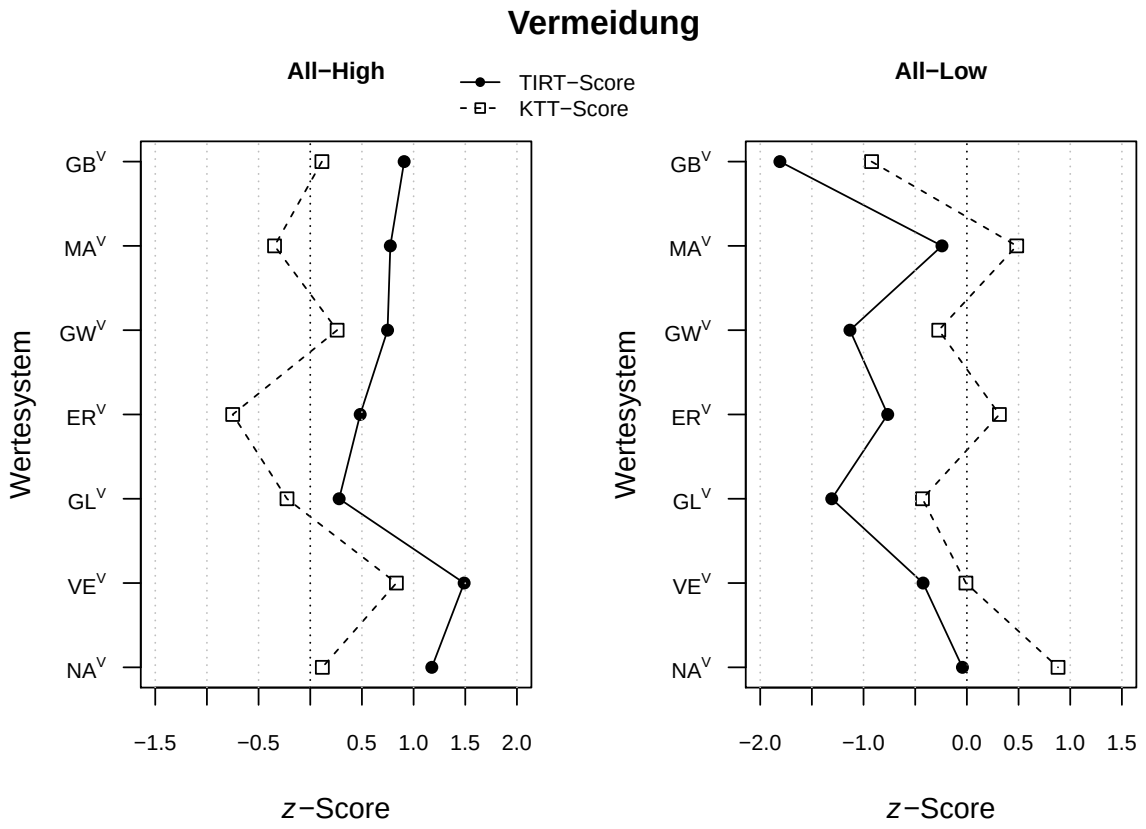

Abbildung 9. 'All-high' und 'all-low' Profile der Vermeidungs-Skala.

ten TIRT-Scores von den Restriktionen der Ipsativität befreit sind und dass die Verwendung der TIRT-Scores deshalb für einen validen interindividuellen Vergleich der Merkmalsausprägungen zulässig ist.

6.3 Diskussion

In diesem Kapitel wurden drei Untersuchungen durchgeführt. Zuerst wurden adäquate Einstellungen des Algorithmus' für die Schätzungen der TIRT-Modelle ermittelt. Es konnte eine Einstellung gefunden werden, die für beide Subskalen akzeptable Anpassungsgüten versprach. Mithilfe dieser Einstellung wurden dann jeweils ein TIRT-Modell an die beiden Subskalen des Fragebogens angepasst. Zum Abschluss wurden die normativen TIRT-Scores mit den ipsativen KTT-Scores verglichen und deren Plausibilität festgestellt.

Insgesamt kann die Vorgehensweise zur Anpassung von TIRT-Modellen mithilfe des `kcirt`-Pakets aufgrund der erforderlichen manuellen Suche nach einer geeigneten Einstellung des Schätzalgorithmus als sehr aufwändig bezeichnet werden. Allerdings konnten die so ermittelten Scores von den der Ipsativität geschuldeten Restriktionen befreit werden und sind

damit aus testtheoretischer Sicht plausibel, um für Untersuchungen der Validität verwendet zu werden.

Die Suche nach geeigneten Parameterkombinationen hat sich als sehr rechenintensive Vorgehensweise erwiesen, da insgesamt mehr als 600 Schätzvorgänge notwendig waren, um die finale TP-Kombination festzulegen. Da ein Schätzvorgang, bei Verwendung von zwei Rechenkernen, zwischen 18 und 24 Stunden reine Rechenzeit erfordert, kann diese Herangehensweise nur dann einigermaßen schnell durchgeführt werden, wenn Zugriff auf ein High Performance Computing Cluster besteht und darin Schätzungen parallel durchgeführt werden können.[5] Dies kann mit hohen Kosten verbunden sein und erfordert in der Regel zusätzliches Expertenwissen zur Bedienung eines solchen Systems.

Da in beiden Subskalen jeweils zwei tendenziell schwächere Blöcke enthalten waren, ergibt sich die Überlegung, jeweils ein reduziertes TIRT-Modell ohne diese Blöcke zu schätzen. Daraus könnte abgeleitet werden, ob ein Fragebogen mit 56 Items (je 7 Items in 8 Blöcken) genügend Informationen bereitstellt, um ein TIRT-Modell mit `kcirt` zu schätzen. Dies könnte in Zukunft untersucht werden.

Außerdem könnte zukünftig erforscht werden, welche Dateneigenschaften welchen Einfluss auf die Passung von Tuning-Parametern haben, um den Aufwand für andere Anwender zu reduzieren. Gerade unter Berücksichtigung der Bedeutung des TIRT-Ansatzes, als einzige existierende Vorgehensweise mit der dominant formulierte FC-Daten modelliert werden können, sowie der Limitierung der Umsetzung mittels M*plus*, erscheint dies sinnvoll.

[5] Die reine Rechenzeit der hier durchgeführten Modellschätzungen beträgt insgesamt mehr als 500 Tage.

Kapitel 7

Reliabilität

In diesem Kapitel wird die Messgenauigkeit des MVSQ anhand der empirischen Reliabilität, sowie der Test-Retest-Reliabilität bestimmt (vgl. hierzu Kapitel 3.6.2). Die entsprechenden Koeffizienten werden für beide Subskalen berechnet. Außerdem wird mit einer weiteren Simulationsstudie bestimmt, ob und in welchem Ausmaß die Schätzungen der empirischen Reliabilitäten verzerrt sind. Dies ist deshalb angebracht, da Brown und Maydeu-Olivares (2011) davon berichten, dass in FC-Fragebögen mit mehr als zwei Items pro Block Verzerrungen der empirischen Reliabilitäten auftreten, obwohl diese Verzerrungen durch die Shrinkage-Technik vermieden werden sollten. Abgesehen davon lohnt diese Analyse, da es bisher wenig Erfahrungswerte im Umgang mit dem TIRT-Ansatz gibt.

7.1 Methode

Der Berechnung der empirischen Reliabilitäten (ρ) werden die TIRT-Modelle aus Kapitel 6.2.2 zugrunde gelegt, die an die Daten aus Stichprobe II (Fragebogenversion 2) angepasst wurden. Zur Berechnung der Test-Retest-Reliabilitäten wird die Teilstichprobe IVa herangezogen (siehe für Beschreibung beider Stichproben Kapitel 3.7).

Die empirischen Reliabilitäten werden gemäß der in Kapitel 3.6.2.1 vorgestellten Prozedur berechnet. Da es sich dabei um ein simulationsbasiertes Verfahren handelt, in dem an mehreren Stellen zufallsgenerierte Daten verwendet werden, wird, wie schon bei der Bestimmung der Anpassungsgüte, das Prinzip der Replikation angewandt. A. Brown (persönliche Kommunikation, 30.10.2013) schlägt hierfür zehn Replikationen und die Berechnung von Median und Spanne der Reliabilitätswerte vor.

Um die Reproduzierbarkeit sicherzustellen, wurde zudem die deterministische Zufallsgenerierung angewandt, d.h. es wurde bei jeder Replikation einer der Werte von 1 bis 10 als sogenannter *seed* (Startwert der Zufallsgenerierung) gewählt. In Replikation 1 kam somit als *seed* die Zahl 1 zum Einsatz.

Zudem wurde in Anlehnung an Brown und Maydeu-Olivares (2011) ein simulationsbasierter Ansatz verfolgt, um die Verzerrung der Reliabilitätswerte abzuschätzen. Dies ist erforderlich, um zu überprüfen, ob die metaheuristische stochastische Suche im `kcirt`-Paket tatsächlich die von Brown und Maydeu-Olivares (2011) beschriebene Problematik der Überschätzung der Testinformation in Fragebögen mit Blockgrößen von mehr als vier Items löst. Dieser Ansatz besteht in der Erweiterung der Prozedur der Reliabilitätsschätzung, die im Folgenden beschrieben und in Abbildung 10 illustriert wird.

Von Schritt (1) bis Schritt (6) handelt es sich dabei um die schon in Kapitel 3.6.2.1 erläuterte Bestimmung der empirischen Reliabilität. Die so berechnete empirische Reliabilität wird als *wahr* angenommen. Auf Basis des (in der Abbildung) mittleren TIRT-Modells werden, analog zu den Schritten (2) bis (6), in den Schritten (7) bis (11) wieder empirische Reliabilitäten bestimmt, die dann als *geschätzte* Reliabilität bezeichnet werden können. Die Abweichung (12) der beiden so bestimmten empirischen Reliabilitäten stellt dann die Verzerrung bei der Reliabilitätsbestimmung dar und kann in Prozent spezifiziert werden. Positive Werte drücken demnach eine Über- und negative Werte eine Unterschätzung aus. Im Endeffekt hängt die Genauigkeit der Reliabilitätsbestimmung eng mit der Anpassungsgüte des TIRT-Modells zusammen, die wiederum von den Tuning-Parameter im Schätzalgorithmus abhängt. Der Mehrwert dieser Prozedur liegt vor allem darin, eine prozentuale Angabe zur Genauigkeit der Reliabilitätsschätzung angeben zu können und diese mit bei Brown und Maydeu-Olivares (2011) berichteten Werten zu vergleichen.

Die Test-Retest-Reliabilitätsuntersuchung wurde als Follow-up im Rahmen eines Laborexperiments durchgeführt, dessen Ergebnisse im weiteren Verlauf der Arbeit vorgestellt werden. Alle Teilnehmer wurden zum selben Zeitpunkt – zehn Wochen nach der letzten Testung im Rahmen der Experimentalstudie – dazu eingeladen, den MVSQ erneut zu bearbeiten. Eine Erinnerung zur zweiten Bearbeitung des MVSQ erfolgte nach 16 Wochen. Die Zeitintervalle zwischen erster und zweiter Testung betrugen im Durchschnitt 15 Wochen und variierten zwischen 11 und 23 Wochen (SD$= 2.6$). Die Rücklaufquote lag bei 37.3%. Die Durchführungsbedingungen waren nicht kontrolliert, da die Teilnehmer den Fragebogen zum Zeitpunkt und am Ort eigener Wahl durchführen konnten. Zudem haben alle Teilnehmer Feedback zu ihren MVSQ-Profilen der ersten Zeitpunkte erhalten. Die Test-Retest-Reliabilitäten (r_{tt}) wurden als Korrelation der TIRT-Scores an den zwei Messzeitpunkten berechnet.

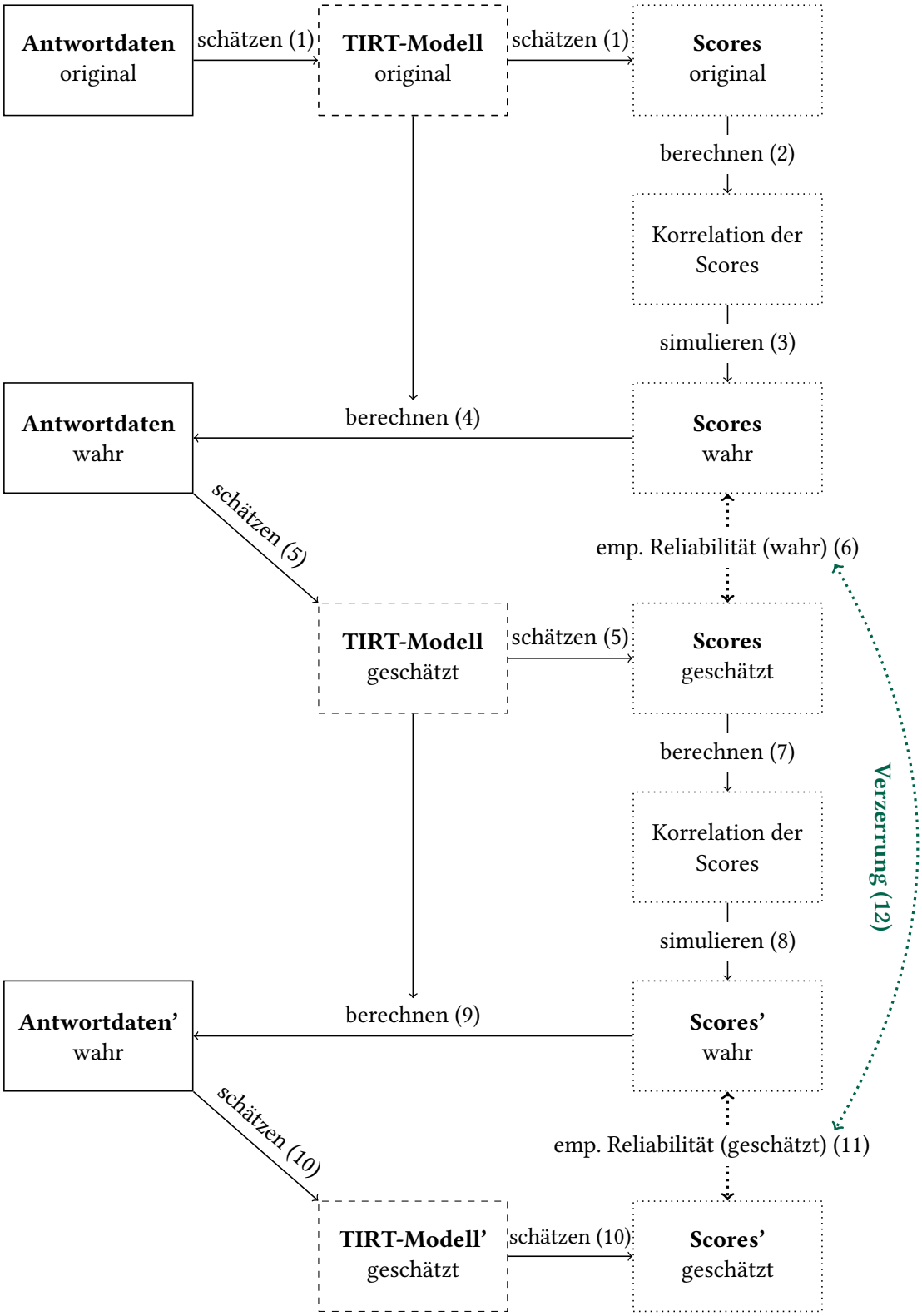

Abbildung 10. Vorgehen zur Bestimmung der Verzerrung der Schätzung der empirischen Reliabilität.

7.2 Ergebnisse

Tabelle 24 zeigt die ermittelten Reliabilitätswerte, wobei die Mediane der empirischen Reliabilitäten der Annäherungsskala zwischen $\rho_{md} = .62$ und $\rho_{md} = .75$ (M = .69, SD = .05) liegen. Die Test-Retest-Reliabilitäten derselben Skala zeigen Werte zwischen $r_{tt} = .68$ und $r_{tt} = .87$ bei einem Mittelwert von M = .78 (SD = .06).

Tabelle 24. Empirische und Test-Retest-Reliabilitäten.

	MVSQ^A				MVSQ^V			
Wertesystem	ρ_{md}	Spanne ρ		r_{tt}	ρ_{md}	Spanne ρ		r_{tt}
GB	.67	.63	.71	.78	.65	.60	.67	.70
MA	.70	.68	.72	.68	.71	.70	.73	.70
GW	.75	.73	.79	.85	.77	.72	.79	.81
ER	.72	.66	.75	.77	.70	.67	.73	.75
GL	.62	.59	.64	.75	.65	.63	.68	.65
VE	.66	.59	.71	.75	.66	.62	.69	.81
NA	.74	.70	.78	.87	.60	.57	.66	.73

Anmerkung. ρ_{md} = Median empirischen Reliabilität der zehn Replikationen; Spanne ρ = der zehn Replikationen; r_{tt} = Test-Retest-Reliabilität; GB = Geborgenheit; MA = Macht; GW = Gewissheit; ER = Erfolg; GL = Gleichheit; VE = Verstehen; NA = Nachhaltigkeit; A = Annäherung; V = Vermeidung.

Auf der Vermeidungsskala liegt die mittlere empirische Reliabilität bei M = .68 (SD = .05) und die Koeffizienten schwanken zwischen $\rho_{md} = .60$ und $\rho_{md} = .77$. Der Mittelwert der Test-Retest-Reliabilitäten beträgt M = .74 (SD = .06). Diese rangieren zwischen $r_{tt} = .65$ und $r_{tt} = .81$.

In der Gesamtschau liegen die Test-Retest-Reliabilitäten in fast allen Fällen über den empirischen Reliabilitäten derselben Wertesysteme, und zwar im Schnitt um .08 (Annäherung) und .06 (Vermeidung). Außer für die Wertesysteme ***Macht***^{A} und ***Gleichheit***^{V} gilt diese Aussage auch für jedes Konstrukt einzeln.

Die Analyse der Verzerrungen der empirischen Reliabilitäten ergab, dass die Reliabilitätsschätzungen insgesamt in einem für die praktische Anwendung vernachlässigbaren Bereich von durchschnittlich unter 2% liegen. Bei der MVSQ^A-Skala wurden die Reliabilitäten im Durchschnitt über zehn Replikationen nur um 1.41% überschätzt und liegen zwischen 0.79% und 2.76%. Bei der MVSQ^V-Skala liegt der Wert der Verzerrung bei 1.99% und bewegt sich zwischen 0.81% und 2.86%.

7.3 Diskussion

In diesem Kapitel wurden die empirischen Reliabilitäten sowie der Grad der Genauigkeit der Berechnung ebendieser bestimmt. Zudem wurde anhand einer kleinen Stichprobe die Test-Retest-Reliabilität berechnet. Von den 14 empirischen Reliabilitäten liegen die Hälfte bei mindestens .70, fünf Werte zwischen .65 und .70 und zwei Werte unter .65. Die Verzerrungen der Reliabilitätsschätzungen betragen im Schnitt über alle Replikationen weniger als 2% und liegen somit in einer vernachlässigbaren Größenordnung. Die Test-Retest-Reliabilitäten liegen deutlich über den empirischen Reliabilitäten mit vier Werten zwischen .80 und .90, acht Werten zwischen .70 und .80 und nur zwei Werten zwischen .65 und .70.

Ruft man sich die in Kapitel 3.6.2.2 dargestellten typischen Reliabilitäten von Werte- und Persönlichkeitsfragebögen in Erinnerung, kann gesagt werden, dass die Reliabilitäten des MVSQ im durchschnittlichen Bereich liegen, wenngleich unter Berücksichtigung der Tatsache, dass die Ergebnisse des MVSQ für Individualuntersuchungen verwendet werden, höhere Reliabilitäten als .80 wünschenswert wären (Groth-Marnat, 2003). Eine weitere Revision des Instruments mit dem Ziel, die empirischen Reliabilitäten zu erhöhen, ist deshalb zu empfehlen.

Die Höhe von Test-Retest-Reliabilitäten hängt in großem Maße vom verwendeten Zeitintervall zwischen den beiden Testungen ab. Die vorliegenden Werte können für den durchschnittlichen Zeitraum von mehr als zehn Wochen als durchaus hoch eingestuft werden (Amelang & Schmidt-Atzert, 2006) und die vom MVSQ gemessenen Wertesysteme als zumindest mittelfristig stabil gesehen werden. Als nächster Schritt sollte eine Test-Retest-Untersuchung über längere Zeiträume durchgeführt werden, um die langfristige Stabilität der Wertesysteme und die zeitliche Messgenauigkeit des MVSQ feststellen zu können.

Zur Schätzung der empirischen Reliabilitäten unter Verwendung des `kcirt`-Pakets kann gesagt werden, dass sich dieses als verlässlich gezeigt hat und sich der in `kcirt` implementierte Schätzer, bei richtiger Einstellung der Tuning-Parameter, als sehr nützliche Methode der Reliabilitätsbestimmung, auch für Fragebögen mit sieben Items pro Block erwiesen hat.

Zur Berechnung der empirischen Reliabilitäten ist zu sagen, dass das Format des MVSQ durch die unidimensionale Formulierung der Items innerhalb der Blöcke suboptimal ist, um die Iteminformationen zu maximieren und hohe Messgenauigkeiten zu erzielen (Brown & Maydeu-Olivares, 2013). Es könnte demnach sein, dass die ermittelten empirischen Reliabilitäten nicht die wahre Qualität des Instruments widerspiegeln, sondern aufgrund methodischer Effekte niedriger ausfallen.

Bei der Untersuchung der Test-Retest-Reliabilität ist einschränkend hinzuzufügen, dass die Stichprobe zum einen relativ klein ausfiel und die Teilnehmer zum anderen unmittelbar nach der ersten Testung Feedback zu ihren Ergebnissen erhielten. Obgleich aufgrund der langen Zeitintervalle von mindestens zehn Wochen Erinnerungseffekte nahezu ausgeschlossen werden können, ist es denkbar, dass das Wissen über das Wertemodell einen Einfluss auf die zweite

Messung hatte. In einer Folgeuntersuchung sollten deshalb zum einen größere Stichproben verwendet werden und zum anderen die Retestung erfolgen, ohne dass die Teilnehmer das Wertemodell vorgestellt bekommen.

Beide Arten der Reliabilität führen zu dem Schluss, dass die Messgenauigkeit von ***Gleichheit***^{V} (ρ und r_{tt} = .65) verhältnismäßig niedrig ausfällt. Die Items dieses Wertesystems sollten deshalb einer Überprüfung hinsichtlich ihrer Formulierungen unterzogen werden. Außerdem waren die empirischen Reliabilitäten insbesondere bei ***Gleichheit***^{A} ($\rho = .62$) und ***Nachhaltigkeit***^{V} ($\rho = .60$) so niedrig, dass bei einer weiteren Revision des Instruments auch ein besonderes Augenmerk auf die Messgenauigkeit dieser beiden Wertesysteme gelegt werden sollte, obgleich die Retest-Reliabilitäten für beide Wertesysteme über .70 lagen.

Insgesamt kann geschlussfolgert werden, dass die Bedingung der Messgenauigkeit für Validitätsuntersuchungen hinreichend erfüllt ist, da die ermittelten Reliabilitäten zumindest für Untersuchungen auf Gruppenebene in einem annehmbaren Bereich liegen. Allerdings sind die Reliabilitäten in einer Größenordnung, in der nicht ausgeschlossen werden kann, dass sie diminuierende Auswirkungen auf folgende Validitätsuntersuchungen haben können. Für die Anwendung auf Einzelfälle ist eine Verbesserung der Reliabilitäten hingegen dringend empfohlen (Groth-Marnat, 2003).

Kapitel 8

Vergleich der Fragebogenversionen

Zum Zeitpunkt als die Überarbeitung der Items der ersten Version erfolgte, war `kcirt` noch nicht publiziert und die Anpassung von TIRT-Modellen an die Daten des MVSQ nicht möglich. Deshalb wurde eine klassische Itemanalyse durchgeführt (vgl. Kapitel 5) und darauf basierend Items zu Überarbeitung vorgeschlagen. In diesem Kapitel werden nun TIRT-Modelle an die Daten der ersten Fragebogenversion angepasst. Daraus ergeben sich mehrere weitere Analysen. Erstens kann überprüft werden, wie sinnvoll die Überarbeitungsempfehlungen waren, die von der klassischen Itemanalyse abgeleitet wurden (Kapitel 5.2.2). Zweitens kann überprüft werden, inwiefern sich die Überarbeitung der Items auf die psychometrischen Eigenschaften des MVSQ ausgewirkt hat.

Dazu werden zuerst die Modellparameter der TIRT-Modelle der ersten Fragebogenversion berichtet und analysiert. Danach erfolgt der Vergleich mit den Ergebnissen der klassischen Itemanalyse, die ebenso auf Version 1 des MVSQ beruhen. Im Anschluss werden die TIRT-Modellparameter der beiden Fragebogenversionen miteinander verglichen und zum Abschluss werden, als weiteres Kriterium der Beurteilung der Qualität der Überarbeitung, die empirischen Reliabilitäten von Version 1 bestimmt und mit denen der überarbeiteten Version kontrastiert.

8.1 Methode

Im ersten Schritt wurden zwei TIRT-Modelle an die Daten der ersten Version angepasst. Dazu wurden die in Kapitel 6 ermittelten Tuning-Parameter verwendet, da damit die Modellparameter im selben Ausmaß (wie bei Version 2) geschrumpft werden und dadurch mit denen von Version 2 vergleichbar sind. Ebenso wurden die empirischen Reliabilitäten und die Verzerrung bei deren Schätzung mit den bereits zuvor angewandten Prozeduren bestimmt (vgl. Kapitel 7). Bei den Stichproben, die dieser Analyse zugrunde liegen, handelte es sich um Stichprobe I und Stichprobe II. Diese sind in Kapitel 3.7 beschrieben.

Da es sich weder bei den Utilities noch den Faktorladungen um standardisierte Maße handelt, können die klassischen Beurteilungskriterien dafür nicht verwendet werden. Deshalb werden

im Folgenden die Mittelwerte und Standardabweichungen der beiden Fragebogenversionen jeweils für Utilities und Faktorladungen berechnet, um sie als Orientierung bei der Bewertung zu verwenden.

Tabelle 25. Kennwerte zur Beurteilung der TIRT-Utilities und Faktorladungen.

	Version 1				Version 2			
Parameter	M	SD	Spanne		M	SD	Spanne	
Utilities	0.00	0.60	-1.51	1.26	0.01	0.60	-1.63	1.55
Faktorladungen	1.31	0.72	-0.66	2.80	1.40	0.63	-0.50	2.86

Zur Beurteilung der Überarbeitungsempfehlungen werden die TIRT-Utilities und Faktorladungen herangezogen, um zu überprüfen, ob diese Kennwerte dieselben Items wie die Kennwerte der Itemanalyse (vgl. Kapitel 5) als überarbeitsungsbedürftig identifizieren. Die Utilities werden dabei mit den Itemschwierigkeiten, die Faktorladungen mit den Trennschärfen verglichen. Des Weiteren werden zwischen den beiden Fragebogenversionen je die Utilities, Faktorladungen und die Skaleninterkorrelationen (nach Fisher's (1925) z-Transformation) miteinander verglichen. Bei den Skaleninterkorrelationen werden zudem Konfidenzintervalle nach Zou (2007) bestimmt, da sie auf unabhängigen Stichproben basieren. Zur Berechnung der Konfidenzintervalle (inklusive z-Transformation) wurde das R Paket `cocor` (Diedenhofen et al., 2015) verwendet.

8.2 Ergebnisse

Der Ergebnisteil dieser Analyse setzt sich aus drei Abschnitten zusammen. Zuerst werden die TIRT-Modelle von Version 1 des MVSQ berichtet. Als zweites werden diese mit den in der Itemanalyse (Kapitel 5) ermittelten klassischen Itemkennwerten verglichen. Daran kann beurteilt werden, ob die abgegebenen Empfehlungen zur Überarbeitung des Fragebogens plausibel waren. Im dritten Abschnitt werden die TIRT-Modellparameter der beiden Fragebogenversionen verglichen. Dadurch kann einerseits abgeschätzt werden, wie ähnlich sich die beiden Versionen sind und andererseits bestimmt werden, ob die Überarbeitung zu besseren Modellparametern geführt hat. Außerdem können die Korrelationen der TIRT-Merkmalsausprägungen der beiden Versionen als Indikator dafür gesehen werden, wie ähnlich die beiden Versionen des Instruments messen. Zudem werden die empirischen Reliabilitäten der Version 1 berichtet und diese als weiterer Indikator für die Güte der Überarbeitung herangezogen.

8.2.1 TIRT-Modelle der ersten Version

Die Modellschätzungen waren erfolgreich, d.h. alle Parameterschätzungen sind wie bei Version 2 innerhalb von 50 Iterationen konvergiert, die entsprechenden Diagramme wurden aus Platzgründen jedoch nicht aufgeführt. Der RMSE betrug beim Annäherungsmodell 0.083 und beim Vermeidungsmodell 0.176. Diese Werte fielen etwas schlechter als bei Version 2 aus, insbesondere beim MVSQ^V-Modell. Die verwendeten Tuning-Parameter waren folglich weniger gut geeignet, um die TIRT-Modelle an diese Daten anzupassen. Um die Vergleichbarkeit der absoluten Ausprägungen der Modellparameter zu gewährleisten, wurden diese Schätzungen dennoch beibehalten.

Die Tabellen 26 und 27 zeigen die Utilities der beiden Subskalen und Tabellen 28 und 29 enthalten die Faktorladungen. Dabei kann festgestellt werden, dass zwei Wertesysteme auf den gesamten Test bezogen leichter zu bevorzugen waren, da sie tendenziell hohe Utility-Werte aufwiesen: ***Gleichheit***^{A} mit der mittleren Utility von M $= 0.76$ und ***Verstehen***^{A} mit M $= 0.67$. Verhältnismäßig schwierig waren die Wertesysteme ***Geborgenheit***^{A} (M $= -0.67$) und ***Nachhaltigkeit***^{A} $M_u = -0.55$ zu bevorzugen. Die mittleren Utilities der Vermeidungswertesysteme schwankten zwischen M $= -0.24$ und M $= 0.32$ und waren damit deutlich gleichmäßiger verteilt.

Tabelle 26. Utilities der MVSQ^A-Skala der ersten Version des Fragebogens.

	Block										
	1	2	3	4	5	6	7	8	9	10	M
GB^A	-1.30	-0.79	-0.13	-0.91	-0.52	0.08	-1.50	-0.42	-0.69	-0.54	-0.67
MA^A	-0.16	-0.54	0.10	-0.93	0.06	-0.48	-0.75	-0.40	-0.28	-0.39	-0.38
GW^A	-0.03	-0.24	0.31	-0.45	-0.70	-0.22	0.35	0.74	-0.66	-0.38	-0.13
ER^A	0.78	0.00	-0.74	0.78	0.32	0.40	0.75	0.03	0.65	-0.86	0.21
GL^A	0.87	0.80	0.72	0.32	1.18	0.63	0.17	0.94	0.82	1.15	0.76
VE^A	0.53	0.73	0.73	0.52	0.28	1.06	1.04	0.09	0.51	1.26	0.67
NA^A	-0.87	-0.20	-0.90	0.64	-0.65	-1.51	-0.23	-1.30	-0.27	-0.24	-0.55

Anmerkung. GB = Geborgenheit; MA = Macht; GW = Gewissheit; ER = Erfolg; GL = Gleichheit; VE = Verstehen; NA = Nachhaltigkeit; A = Annäherung.

Bei den Faktorladungen der MVSQ^A-Skala der ersten Version zeigten sich zwei Muster. Erstens waren die Ladungen des ***Erfolg***^{A}-Wertesystems deutlich niedriger als die übrigen und zweitens stach Block 6 als besonders schlecht hervor, mit zwei Ladungen unter Null. Abgesehen davon fiel Block 9 auf, der vergleichsweise hohe Faktorladungen aufwies. Bei den MVSQ^V-

Tabelle 27. Utilities der MVSQ^{V}-Skala der ersten Version des Fragebogens.

	Block										
	1	2	3	4	5	6	7	8	9	10	M
GB^{V}	0.81	0.53	0.03	-0.22	-0.05	0.61	-0.07	-0.61	-0.60	0.38	0.08
MA^{V}	-0.30	0.68	0.65	0.20	0.42	-0.63	0.25	0.71	0.41	-0.12	0.23
GW^{V}	0.45	0.18	0.88	-0.28	0.31	0.04	-0.19	0.64	0.98	0.17	0.32
ER^{V}	0.46	0.54	-0.06	0.85	-0.81	-0.66	0.25	0.07	-0.23	0.24	0.07
GL^{V}	-0.24	-0.76	-0.46	-0.26	0.55	0.28	-0.06	-1.13	-0.06	0.32	-0.18
VE^{V}	-0.66	-1.10	-0.30	0.09	-0.08	0.12	-0.34	0.93	-0.45	-0.55	-0.24
NA^{V}	-0.14	-0.20	-0.55	-0.41	-0.25	0.14	0.26	-0.18	-0.05	-0.34	-0.17

Anmerkung. GB = Geborgenheit; MA = Macht; GW = Gewissheit; ER = Erfolg; GL = Gleichheit; VE = Verstehen; NA = Nachhaltigkeit; V = Vermeidung.

Tabelle 28. Faktorladungen der MVSQ^{A}-Skala der ersten Version des Fragebogens.

	Block										
	1	2	3	4	5	6	7	8	9	10	M
GB^{A}	1.69	1.49	2.28	1.16	1.52	0.46	1.97	2.04	1.89	1.29	1.58
MA^{A}	1.55	0.88	2.06	0.88	1.79	0.92	1.33	0.85	2.21	1.56	1.40
GW^{A}	2.51	1.80	0.39	1.73	2.10	0.35	0.70	1.80	2.38	0.73	1.45
ER^{A}	-0.63	0.80	0.16	1.71	-0.44	-0.26	0.94	1.03	1.96	2.48	0.78
GL^{A}	1.55	1.98	1.40	1.37	2.12	0.83	1.72	1.89	1.77	1.37	1.60
VE^{A}	1.62	2.04	1.59	1.86	1.55	0.59	1.76	1.49	0.91	0.14	1.35
NA^{A}	0.86	0.71	1.54	0.77	1.58	-0.04	2.57	1.91	2.80	0.96	1.37
M	1.31	1.39	1.35	1.35	1.46	0.41	1.57	1.57	1.99	1.22	1.36

Anmerkung. GB = Geborgenheit; MA = Macht; GW = Gewissheit; ER = Erfolg; GL = Gleichheit; VE = Verstehen; NA = Nachhaltigkeit; A = Annäherung.

Ladungen waren die Unterschiede der mittleren Faktorladungen zwischen den Wertesystemen weniger deutlich. Die niedrigsten Ladungen traten bei den Items des ***Verstehen***^{V}-Wertesystems auf. Auch bezogen auf die Blöcke war die Volatilität der Ausprägungen weniger stark als bei den MVSQ^{A}-Ladungen. Dennoch gab es zwei Blöcke (9 und 10) mit merklich schlechteren Faktorladungen. Besonders hohe Ladungen waren in Block 1 zu beobachten.

Tabelle 29. Faktorladungen der MVSQ^V-Skala der ersten Version des Fragebogens.

	Block										
	1	2	3	4	5	6	7	8	9	10	M
GB^V	2.24	1.30	1.09	0.98	1.46	2.44	1.35	-0.06	0.22	2.15	1.32
MA^V	1.95	1.50	1.40	1.31	1.89	1.45	1.74	0.87	1.61	0.74	1.44
GW^V	1.75	1.13	1.49	1.14	2.42	1.90	1.09	1.05	1.29	2.43	1.57
ER^V	2.37	0.51	1.71	1.34	1.10	0.11	0.30	2.56	0.67	0.12	1.08
GL^V	2.42	1.44	1.60	0.75	1.88	2.13	0.78	0.48	0.56	1.19	1.32
VE^V	1.87	1.21	1.52	1.24	-0.66	0.77	2.02	0.52	1.10	0.36	1.00
NA^V	1.29	1.43	1.16	1.49	2.28	1.18	1.52	0.17	-0.22	0.44	1.07
M	1.98	1.22	1.42	1.18	1.48	1.43	1.26	0.80	0.75	1.06	1.26

Anmerkung. GB = Geborgenheit; MA = Macht; GW = Gewissheit; ER = Erfolg; GL = Gleichheit; VE = Verstehen; NA = Nachhaltigkeit; V = Vermeidung.

8.2.2 Übereinstimmung der Überarbeitungsempfehlungen

Als nächster Schritt werden die Überarbeitungsempfehlungen sowohl einzelner Items als auch auf Block- und Konstruktebene verglichen. In der klassischen Analyse basierten diese ausschließlich auf den Itemschwierigkeiten, da diese dem geringsten Einfluss der Ipsativität unterliegen (vgl. Kapitel 5.1). Bei der Analyse der TIRT-Kennwerte wird nicht mehr auf die paarweisen Schwierigkeiten zurückgegriffen, sondern es stehen Faktorladungen als aussagekräftigstes Kriterium der Qualität einzelner Items zur Verfügung. Die Utilities sind die fast identische Entsprechung der klassischen Itemschwierigkeiten, wie ein Vergleich dieser beiden Maße zeigte. Produkt-Moment-Korrelationen zwischen den klassischen Itemschwierigkeiten und den TIRT-Utilities lagen bei Annäherung und Vermeidung gleichermaßen fast bei $r = 1$.

In Kapitel 5.2.2 wurde Block 6 der MVSQ^A-Skala als besonders überarbeitungsbedürftig identifiziert. Dieser Block weist auch die mit Abstand niedrigsten Faktorladungen auf (im Mittel 0.41, vgl. Tabelle 28). Alle anderen Blöcke haben im Durchschnitt Ladungen zwischen 1.22 und 1.99. Im Unterschied zu den klassischen Trennschärfen, die auch Block 10 als schlechter identifizieren, gilt dies bei den TIRT-Faktorladungen nur noch begrenzt. Zwar hat Block 10 im Schnitt etwas niedrigere Faktorladungen, diese erscheinen jedoch nicht bedenklich, da sie nur minimal unter dem allgemeinen Durchschnitt liegen. Ein Blick auf die einzelnen Ladungen dieses Blocks zeigt zudem, dass es lediglich ein wirklich problematisches Item in diesem Block gibt, nämlich Item ***Verstehen*** mit einer Ladungen von 0.14.

Als problematisch auf Basis der TIRT-Kennwerte wurden Items bezeichnet, die entweder Faktorladungen aufwiesen, die mehr als eine Standardabweichung von der allgemeinen mittleren Faktorladung abwichen oder extreme Utilities, die mehr als zwei Standardabweichungen vom Mittel divergierten. Dass bei den Utilities zwei Standardabweichungen statt einer gewählt wurden, liegt darin begründet, dass bei ihnen eine gewisse Streuung der Ausprägungen wünschenswert ist, da dadurch das Merkmal an unterschiedlichen Stellen des Kontinuums gemessen wird (Embretson & Reise, 2000). Demnach zu urteilen, waren in der MVSQ^A-Skala die Items ER_1^A, ER_2^A, ER_3^A, ER_5^A, ER_6^A, GW_3^A, GB_6^A, GW_6^A, NA_6^A und VE_{10}^A auf Basis der Faktorladungen und Items GB_1^A, NA_6^A, GB_7^A und NA_8^A auf Basis der Utilities problematisch. Damit bestätigen die TIRT-Parameter die Empfehlung zur Überarbeitung der Items GB_1^A, NA_6^A, GB_7^A, NA_8^A und VE_{10}^A und damit exakt die Hälfte der insgesamt empfohlenen Items. Von den acht weiteren Items, die laut TIRT Revisionsbedarf hatten, fanden sich drei in der Liste der fragwürdigen Items (Tabelle 14) wieder. Die verbleibenden fünf Items fielen laut klassischer Kriterien nicht auf.

Bei der MVSQ^V-Skala war, den TIRT-Ladungen (Tabelle 29) nach zu urteilen, neben Block 8 auch Block 9 von minderer Qualität. Die mittleren Faktorladungen lagen bei 0.8 (Block 8) und 0.75 (Block 9) und damit knapp eine Standardabweichung unter dem allgemeinen Durchschnitt. Die übrigen Ladungen lagen im Mittel zwischen 1.06 und 1.98. Auch auf Basis der klassischen Trennschärfen wurden diese beiden Blöcke als schlechter eingestuft. Faktorladungen und Trennschärfen stimmen also überein. Interessant ist aber der folgende Vergleich der Überarbeitungsempfehlungen der einzelnen Items.

Die klassische Analyse identifizierte bei der MVSQ^V-Skala dabei nur drei problematische Items: VE_2^V, GB_8^V und GL_8^V. Den Utilities zufolge fiel keines dieser Items in den festgelegten Extrembereich mit mehr als zwei Standardabweichungen Abstand zur Mitte. Den Faktorladungen nach zu urteilen, gab es jedoch einige revisionswürdige Items, die alle mehr als eine Standardabweichung unter dem Durchschnitt lagen: GB_8^V, GB_9^V, ER_2^V, ER_6^V, ER_7^V, ER_{10}^V, GL_8^V, GL_9^V, VE_5^V, VE_8^V, VE_{10}^V, NA_8^V, NA_9^V und NA_{10}^V. Selbst wenn man die nach klassischen Kriterien als fragwürdig eingestuften Items mit einbezieht (eine Übereinstimmung), macht dieser Vergleich deutlich, dass die TIRT-Parameter zu deutlich unterschiedlichen und wesentlich detaillierteren Überarbeitunsempfehlungen führen.

8.2.3 Vergleich der Modellparameter, Skaleninterkorrelationen und Reliabilitäten

Im folgenden Abschnitt werden zuerst die Utilities und Faktorladungen der beiden Fragebogenversionen verglichen. Als zweites werden die Skaleninterkorrelationen und schließlich die empirischen Reliabilitäten der beiden Versionen gegenübergestellt.

Vergleich der Utilities und Faktorladungen

Ob die Überarbeitung global zu einer Verbesserung geführt hat, kann anhand der Utilities und Faktorladungen beurteilt werden. Bei den Utilities wären zwei Tendenzen wünschenswert. Erstens spiegelt die mittlere Varianz der Utilities pro Wertesystem wider, wie breit die Abdeckung des Merkmalskontinuums ist. Größere Werte sind hier wünschenswert. Zweitens zeigt die Varianz der mittleren Utilities pro Wertesystem an, wie vergleichbar die Wertesysteme hinsichtlich der Schwierigkeit sind. Hier wäre ein niedrigerer Wert besser.

Bei erstem Kriterium hat die Überarbeitung bei der MVSQ^A-Skala zu einem leichten positiven Effekt geführt, da die Varianz gemittelt über die Wertesysteme angestiegen ist ($\Delta \mathrm{V} = 0.07$). Bei der MVSQ^V-Skala war die Entwicklung gegenteilig. Die mittlere Varianz pro Wertesystem sank um 0.05. Die Varianzen der mittleren Utilities pro Wertesystem sind bei beiden Skalen nahezu unverändert geblieben: $\Delta \mathrm{V} = -0.02$ bei MVSQ^A und $\Delta \mathrm{V} = 0$ bei MVSQ^V.

Bei den Faktorladungen ist wünschenswert, dass sich diese erhöhen. Dies ist in beiden Skalen der Fall gewesen. Die MVSQ^A-Faktorladungen haben sich im Schnitt um 0.07, die der MVSQ^V-Skala um 0.12 verbessert. Global hat die Überarbeitung also zu einer Verbesserung geführt, wenngleich nur in geringem Ausmaß.

Beim Vergleich der beiden Fragebogenversionen beträgt die Korrelation zwischen den Utilities der Annäherungsdimension $r = .68$, den Vermeidungs-Utilities $r = .56$, den Annäherungs-Faktorladungen $r = .42$ und den Vermeidungs-Faktorladungen $r = .56$. Diese Korrelationen zeigen, dass die Überarbeitung zu beträchtlichen Veränderungen der TIRT-Parameter hinsichtlich ihrer Struktur geführt hat. Die Tabellen 30 und 31 beinhalten die Korrelationen zwischen Utilities und Faktorladungen der beiden Fragebogenversionen, jeweils aufgeschlüsselt nach Blöcken bzw. Wertesystemen. Sie zeigen, inwiefern sich die Modellparameter durch die Überarbeitung verändert haben.

Beim Vergleich der Blöcke (Tabelle 30) unterscheiden sich die Annäherungs-Utilities vergleichsweise deutlich in den Blöcken 6 ($r = .36$), 7 ($r = .42$) und 8 ($r = .41$), sowie die Vermeidungs-Utilities in den Blöcken 4 ($r = -.55$), 5 ($r = .44$), 8 ($r = .50$) und 9 ($r = .21$). Die übrigen Werte liegen im hohen Bereich zwischen $r = .60$ und $r = .96$.

Die Faktorladungen differieren in der Annäherungsskala besonders stark in den Blöcken 1 ($r = .14$), 7 ($r = -.27$), 8 ($r = .12$), sowie relativ deutlich in den Blöcken 6 ($r = .34$) und 10 ($r = .32$), in der Vermeidungsskala in den Blöcken 1 ($r = .23$) und 8 ($r = -.39$). Die restlichen Faktorladungen liegen, auf beide Skalen bezogen, im mittleren bis hohen Bereich (zwischen $r = .41$ und $.89$).

Bezogen auf die Wertesysteme (Tabelle 31) hat die Überarbeitung zu deutlich unterschiedlichen Utilities der Annäherungswertesysteme ***Geborgenheit*** ($r = -.21$) , ***Gleichheit*** ($r = .24$) und ***Nachhaltigkeit*** ($r = -.06$) geführt. In Kapitel 6.2.2 wurde allerdings gezeigt, dass auch in der überarbeiteten Version ***Geborgenheit***^{A} und ***Nachhaltigkeit***^{A} besonders schwer zu bevorzugen waren und auch die leicht zu bevorzugenden Annäherungswertesysteme (***Gleichheit***^{A}

Tabelle 30. Korrelationen der TIRT-Utilities und Faktorladungen zwischen den Fragebogenversionen pro Block.

	1	2	3	4	5	6	7	8	9	10
Utilities A	.91	.89	.90	.60	.91	.36	.42	.41	.92	.82
Utilities V	.88	.82	.92	-.55	.44	.96	.62	.50	.21	.95
Faktorladungen A	.14	.78	.43	.41	.53	.34	-.27	.12	.46	.32
Faktorladungen V	.23	.83	.52	.82	.53	.89	.81	-.39	.74	.72

Anmerkung. A=Annäherung; V=Vermeidung.

und ***Verstehen***^{A}) unverändert geblieben sind. Bei den Vermeidungswertesystemen fiel lediglich die Korrelation der ***Erfolg***^{V}-Utilities ($r = .11$) auf, die vergleichsweise niedrig war. Absolut gesehen ist der Mittelwert der Utilities von ***Erfolg***^{V} mit M = 0.21 in Version 1 und M = 0.08 in der überarbeiteten Version allerdings in einem ähnlichen Bereich geblieben.

Tabelle 31. Korrelationen der TIRT-Utilities und Faktorladungen zwischen den Fragebogenversionen pro Wertesystem.

	GB	MA	GW	ER	GL	VE	NA
Utilities A	-.21	.43	.84	.73	.24	.77	-.06
Utilities V	.34	.90	.73	.11	.74	.52	.71
Faktorladungen A	.21	.77	.51	.50	.27	.42	.68
Faktorladungen V	.75	.67	.81	.49	.93	.50	.28

Anmerkung. Wertesysteme: GB = Geborgenheit; MA = Macht; GW = Gewissheit; ER = Erfolg; GL = Gleichheit; VE = Verstehen; NA = Nachhaltigkeit; A=Annäherung; V=Vermeidung.

Bei den Faktorladungen pro Wertesystem waren ***Geborgenheit***^{A} und ***Gleichheit***^{A} ($r = .21$ bzw. $r = .27$) sowie ***Nachhaltigkeit***^{V} ($r = .28$) relativ niedrig. Alle anderen Korrelationen lagen im mittleren bis hohen Bereich zwischen $r = .42$ und $r = .93$.

Des Weiteren gibt Tabelle 32 einen Überblick über die Veränderungen der Faktorladungen der Items, deren Überarbeitung in Kapitel 5 empfohlenen wurde. Darin kann gesehen werden, dass sich die problematischen Faktorladungen nur in der Vermeidungsskala verbessert haben, in der Annäherungsskala hingegen schlechter geworden sind. Bei den fragwürdigen Items waren die Veränderungen geringer, die der MVSQ^{A}-Skala haben sich leicht verbessert und die der MVSQ^{V}-Skala geringfügig verschlechtert.

Tabelle 32. Veränderung der Faktorladungen der zur Überarbeitung empfohlenen Items.

Items	MVSQ^A	MVSQ^V
problematisch	-0.31	0.41
fragwürdig	0.19	-0.16

Vergleich der TIRT-Skaleninterkorrelationen

Zum Vergleich der Skaleninterkorrelationen zwischen den beiden Versionen wurden Fisher's z und 95%-Konfidenzintervalle nach Zou (2007) berechnet. Die Tabellen 33 und 34 enthalten für die Annäherungs- bzw. Vermeidungsskalen neben diesen beiden Werten die entsprechenden Korrelationen und p-Werte. Von Korrelationen der Annäherungsskalen unterschieden sich 12 der 21 Korrelationen signifikant voneinander und bei den Vermeidungswertesystemen wurden 8 der 21 Unterschiede signifikant.

Auf die einzelnen Wertesysteme bezogen, kann festgestellt werden, dass jedes Wertesystem von einer signifikant unterschiedlichen Korrelation betroffen war. Bzgl. der Wertesysteme ***Geborgenheit***^{A}, ***Gewissheit***^{A}, ***Erfolg***^{A} und ***Nachhaltigkeit***^{A} haben sich je vier Korrelationen zu anderen Wertesystemen signifikant verändert, bei ***Gleichheit***^{A} und ***Verstehen***^{A} waren drei Korrelationen und bei ***Macht***^{A} eine Korrelation bei Version 2 anders als bei Version 1. Bei der Vermeidungsskala unterschieden sich vier Korrelationen signifikant, die ***Geborgenheit***^{V} beinhalteten, drei Korrelationen die ***Gewissheit***^{V} und ***Nachhaltigkeit***^{V} betreffen, sowie zwei Korrelationen, von denen ***Macht***^{V}, ***Erfolg***^{V} und ***Gleichheit***^{V} Bestandteil sind. Keine unterschiedlichen Korrelationen ergaben sich bzgl. ***Verstehen***^{V}. In Summe lässt sich feststellen, dass die Überarbeitung zu beträchtlichen Veränderungen der Zusammenhänge der Wertesysteme untereinander geführt hat, wobei bei der Annäherungsskala größere Veränderungen auftraten als bei der Vermeidungsskala.

Vergleich der empirischen Reliabilitäten

Die empirischen Reliabilitäten von Version 1 wurden auf die gleiche Art bestimmt wie bei Version 2 (vgl. Kapitel 7). Ebenso wurden dieselben Tuning-Parameter eingesetzt und auch die Verzerrung der Reliabilitätsschätzung wurde wie in Kapitel 7 durch Simulation geschätzt.

Tabelle 35 zeigt die empirischen Reliabilitäten der Version 1 des Fragebogens. Für die Annäherungsskala ergab sich eine mittlere empirische Reliabilität von $\rho_{md} = .68$ und eine Bandbreite der Werte von $\rho_{md} = .53$ bis $\rho_{md} = .73$. Bei der Vermeidungsskala lagen die Werte zwischen $\rho_{md} = .54$ und $\rho_{md} = .76$ (M $= .65$, SD $= .08$).

Im Vergleich zu Version 2 waren die empirischen Reliabilitäten zum einen im Schnitt etwas niedriger (ΔM $= .02$ bei der Annäherungsskala und ΔM $= .13$ bei Vermeidung), zum anderen

Tabelle 33. Vergleich der Fragebogenversionen: Merkmalsinterkorrelationen der MVSQ^A-Skala.

Wertesystem		Version						
		r_1	r_2	z	p	KI$_z$ (95%)		
GB	MA	-.06	-.23	3.16	<.01	.07	-	.28
GB	GW	.57	.46	2.73	<.01	.03	-	.19
GB	ER	.09	-.13	3.97	<.001	.11	-	.33
GB	GL	.46	.42	0.96	.34	-.04	-	.13
GB	VE	-.16	-.22	1.16	.25	-.04	-	.17
GB	NA	.18	.05	2.27	<.05	.02	-	.23
MA	GW	-.15	-.12	-0.56	.57	-.14	-	.08
MA	ER	.18	.27	-1.84	.07	-.20	-	.01
MA	GL	-.14	-.23	1.58	.11	-.02	-	.19
MA	VE	.00	.01	-0.23	.82	-.12	-	.10
MA	NA	-.08	-.20	2.16	<.05	.01	-	.22
GW	ER	.15	.00	2.60	<.01	.04	-	.25
GW	GL	.30	.17	2.42	<.05	.02	-	.23
GW	VE	-.36	-.17	-3.67	<.001	-.29	-	-.09
GW	NA	-.07	-.14	1.32	.19	-.04	-	.18
ER	GL	-.07	-.37	5.69	<.001	.20	-	.40
ER	VE	-.05	.06	-1.91	.06	-.21	-	.00
ER	NA	-.22	-.33	2.04	<.05	.00	-	.20
GL	VE	.14	-.07	3.82	<.001	.10	-	.32
GL	NA	.27	.28	-0.09	.93	-.11	-	.10
VE	NA	.22	.09	2.40	<.05	.02	-	.24

Anmerkung. r_1 = Korrelation von Version 1; r_2 = Korrelation Version 2; z = Fisher's z; KI_z = Zou's Konfidenzintervall.

streuten die Werte mehr. Die Standardabweichungen der empirischen Reliabilitäten lagen in Version 1 bei beiden Skalen um .02 über den Werten von Version 2.

Bezogen auf die empirischen Reliabilitäten kann somit festgehalten werden, dass die Überarbeitung insgesamt positive Auswirkungen hatte. Wobei zu sagen ist, dass diese Verbesserung vor allem daran liegt, dass die besonders niedrigen Reliabilitäten deutlich gestiegen sind.

Tabelle 34. Vergleich der Fragebogenversionen: Merkmalsinterkorrelationen der MVSQ^V-Skala.

Wertesystem		Korrelation / Version r_1	r_2	z	p	KI_z (95%)
GB	MA	-.17	-.30	2.54	<.05	.03 - .23
GB	GW	.53	.44	2.02	<.05	.00 - .17
GB	ER	-.01	-.28	5.03	<.001	.17 - .37
GB	GL	.21	.30	-1.77	.08	-.19 - .01
GB	VE	-.24	-.21	-0.44	.66	-.13 - .08
GB	NA	-.12	.04	-3.01	<.01	-.27 - -.06
MA	GW	-.10	-.19	1.64	.10	-.02 - .19
MA	ER	.45	.40	1.23	.22	-.03 - .15
MA	GL	-.41	-.30	-2.26	<.05	-.20 - -.01
MA	VE	.02	-.04	1.09	.27	-.05 - .17
MA	NA	-.01	-.10	1.58	.11	-.02 - .19
GW	ER	.19	-.03	4.00	<.001	.11 - .33
GW	GL	.10	.12	-0.28	.78	-.12 - .09
GW	VE	-.15	-.06	-1.55	.12	-.19 - .02
GW	NA	-.28	-.16	-2.21	<.05	-.22 - -.01
ER	GL	-.25	-.35	1.84	.07	-.01 - .19
ER	VE	.05	-.02	1.35	.18	-.03 - .18
ER	NA	-.16	-.18	0.40	.69	-.08 - .13
GL	VE	-.06	-.07	0.28	.78	-.09 - .12
GL	NA	.01	.16	-2.85	<.01	-.26 - -.05
VE	NA	.24	.30	-1.18	.24	-.16 - .04

Anmerkung. r_1 =Korrelation von Version 1; r_2 = Korrelation Version 2; z = Fisher's z; KI_z = Zou's Konfidenzintervall.

Die auf Basis einer Simulation ermittelten Verzerrungen der Reliabilitätsschätzungen lagen mit durchschnittlich 1.95% für MVSQ^A und 4.86% für die MVSQ^V-Reliabilitäten zwar höher als die Verzerrungen bei Version 2, jedoch noch in einem akzeptablen Bereich. Diese größere Verzerrung der Reliabilitätsschätzung kann so gedeutet werden, dass die Tuning-Parameter

Tabelle 35. Empirische Reliabilitäten von Version 1.

	MVSQ^A				MVSQ^V			
	ρ_{md}	Spanne ρ			ρ_{md}	Spanne ρ		
GB	.72	.70	-	.73	.70	.67	-	.73
MA	.68	.65	-	.72	.69	.66	-	.73
GW	.73	.70	-	.75	.76	.70	-	.78
ER	.53	.51	-	.56	.61	.56	-	.67
GL	.71	.68	-	.74	.68	.64	-	.70
VE	.67	.62	-	.68	.54	.47	-	.56
NA	.71	.69	-	.74	.56	.50	-	.60

Anmerkung. ρ_{md} = Median empirischen Reliabilität über zehn Replikationen; Spanne ρ = der zehn Replikationen; GB = Geborgenheit; MA = Macht; GW = Gewissheit; ER = Erfolg; GL = Gleichheit; VE = Verstehen; NA = Nachhaltigkeit; A = Annäherung; V = Vermeidung.

weniger gut zu den Daten von Version 1 passten, die Beibehaltung der Tuning-Parameter jedoch auf Grund der besseren Vergleichbarkeit zwischen den Version gerechtfertigt werden kann.

8.3 Diskussion

In dieser Untersuchung wurden zunächst zwei TIRT-Modelle an die Daten der ersten Version des Fragebogens angepasst. Auf Basis der Kenndaten dieser Modelle wurden die Überarbeitungsempfehlungen aus Kapitel 5 überprüft. Dabei wurden festgestellt, dass zwar Itemschwierigkeiten und Utilities zu den selben Überarbeitungsempfehlungen führen, Trennschärfen und Faktorladungen allerdings zu deutlich unterschiedlichen Empfehlungen führen. Ferner wurden die Auswirkungen der Revision untersucht. Dabei hat sich gezeigt, dass die Utilities relativ unverändert geblieben sind. Die Faktorladungen haben sich zwar in Summe leicht verbessert haben, jedoch nur teilweise bezogen auf die Überarbeitungsempfehlungen. Bezüglich der Skaleninterkorrelationen haben sich eine Vielzahl der Werte signifikant in ihren Höhen verändert. Dies zeigt, dass die Überarbeitung der Items einen merklichen Einfluss gezeigt hat, wenngleich aufgrund fehlender Hypothesen über die Höhen der Skaleninterkorrelationen die Qualität dieser Veränderungen nicht beurteilt werden kann. Bei den empirischen Reliabilitäten hat sich die Mehrheit der empirischen Reliabilitäten von Version 1 auf Version 2 verbessert. Besonders sind an dieser Stelle die Verbesserungen von ***Erfolg***^{A} (um .19), ***Erfolg***^{V} (um .09) und ***Verstehen***^{V} (um .12), sowie die Verschlechterung von ***Gleichheit***^{A} (um .09) hervorzuheben.

Zum Zusammenhang zwischen Faktorladungen und klassischen Trennschärfen kann gesagt werden, dass diese zwar tendenziell ähnlich waren, jedoch teilweise auch gravierende Unterschiede zeigten. Zum Beispiel wären nach den Trennschärfen die Items ER_4^A und ER_{10}^A problematisch, die TIRT-Faktorladungen bescheinigen diesen Items jedoch eine hohe Aussagekraft. Da es noch weitere Unterschiede zwischen dem KTT und IRT-Kennwerten gab, kann somit im Nachhinein die Vorgehensweise in Kapitel 5 gerechtfertigt werden, Trennschärfen nicht auf Itemebene als Kriterium heranzuziehen. Andererseits zeigen die Ergebnisse dieser Analyse auch, dass die Überarbeitung nur auf Basis von Itemschwierigkeiten suboptimal ist, da so Informationen über den Zusammenhang von Items, die denselben latenten Trait messen, fehlen. Insgesamt kann aus diesem Kapitel vor allem eine Schlussfolgerungen gezogen werden. Es hat sich gezeigt, dass die Überarbeitung von Forced-Choice-Items auf Basis einer klassischen Itemanalyse zu unvorhergesehen Veränderungen der TIRT-Kennwerte geführt hat. Damit hat sich bestätigt, dass die Analyse von ipsativen Items durch klassische Methoden unvorteilhaft sein kann (Brown & Maydeu-Olivares, 2011). Allerdings haben sich die Eigenschaften des Fragebogen dennoch insgesamt tendenziell verbessert, was zeigt, dass auch aus den klassischen Itemkennwerten zumindest teilweise nützliche Informationen über die Qualität der Items gewonnen werden konnten. Außerdem kann auch gesagt werden, dass sich insgesamt sehr viele Itemkennwerte verändert haben und es sich deshalb für eine weitere Revision empfiehlt, weniger Items zu überarbeiten, um dann auch gezielter feststellen zu können, welche Auswirkung die Überarbeitung einzelner Items hat.

Kapitel 9

Konstruktvalidität

Dieses Kapitel widmet sich der Untersuchung der Konstruktvalidität des MVSQ. Dazu werden die faktorielle Validität, sowie die konvergente und divergente Validität in mehreren Analysen begutachtet.

9.1 Faktorielle Validität

Zur Analyse der faktoriellen Validität eines Instruments wird die Struktur der gemessenen Merkmale inferenzstatistisch auf ihre Hypothesenkonsistenz überprüft (Bühner, 2011). Für den MVSQ können dazu zwei Hypothesen formuliert werden: Erstens gilt im Grundsatz die Annahme, dass die Wertesysteme unabhängig voneinander sind. Zwar ist in der Psychologie nicht davon auszugehen, vollständige Unabhängigkeit zwischen Merkmalen zu erreichen (Fabrigar et al., 1999), dennoch sollten die Wertesysteme weitestgehend unabhängig voneinander sein. Andererseits unterteilte Graves die Wertesysteme in zwei Gruppen (vgl. Kapitel 2.2): *express-self* - und *sacrifice-self* -Wertesysteme. Beide Gruppierungen von Wertesystemen teilen je eine zentrale Eigenschaft, nämlich den Fokus auf das Ausdrücken des eigenen Selbst (*express-self*) bzw. die Orientierung zur Aufopferung des Selbst (*sacrifice-self*). Aufgrund dieser Kategorisierung liegt die Annahme nahe, dass Wertesysteme je innerhalb der Gruppen stärker kovariieren als als zwischen den Gruppen. Diese Fragestellung wird anhand der Skaleninterkorrelationen des MVSQ sowie einer Faktorenanalyse der Scores untersucht.

Die zweite Hypothese betrifft die Anzahl der Merkmale mit der zentralen Frage, ob es sich bei Annäherungs- und Vermeidungsdimension um zwei entgegengesetzte Pole desselben Konstrukts handelt oder ob diese voneinander unabhängig sind (vgl. Kapitel 3.1.2). Diese Hypothese behandelt folglich die Thematik der Orthogonalität versus Bipolarität der Dimensionen. Zur Annäherung daran können einerseits die Merkmalskorrelationen zwischen den Wertesystemen beider Dimensionen und auch die Ergebnisse einer Faktorenanalyse der Scores verwendet werden. Außerdem wird ein TIRT-Modell geschätzt, indem die Items beider Skalen

zusammengefasst als Indikatoren von sieben bipolaren Wertesystemen spezifiziert werden. Die so geschätzten Scores können wiederum in Bezug zu den orthogonalen Scores gesetzt werden.

9.1.1 Methode

Zunächst sei gesagt, dass es sich bei der zugrunde liegenden Stichprobe um Stichprobe II handelt, die bereits in Kapitel 3.7 beschrieben wurde. Zur Überprüfung der aufgeführten Fragestellungen werden in diesem Kapitel zuerst die Interkorrelationen zwischen den beiden MVSQ-Subskalen begutachtet. Danach erfolgt die Durchführung einer explorativen Faktorenanalyse (EFA). Dabei sei erwähnt, dass die EFA auf Grund der Ipsativität nicht unmittelbar auf Itemebene durchführbar ist (vgl. Kapitel 3.3.2), was zur Folge hat, dass die Faktorenstruktur nicht in einer konfirmatorischen Faktorenanalyse bestätigt werden kann. Stattdessen wird eine explorative Faktorenanalyse auf der Ebene der Scores durchgeführt und darin untersucht, ob sich sinnvoll interpretierbare Faktoren zweiter Ordnung ergeben. Die EFA und dies betreffenden Koeffizienten (z.B. KMO und MSA) werden mit dem R-Paket `psych` (Revelle, 2015) berechnet. Nach der EFA erfolgt die Schätzung eines TIRT-Modells mit sieben latenten Traits (bipolaren Wertesystemen), das aus 20 Blöcken mit je sieben Items besteht. Zur Schätzung dieses Modells wird dasselbe Verfahren wie in Kapitel 6 angewandt. Das heißt, es werden dieselben Tuning-Parameter verwendet und sowohl die Güte der Modellanpassung als auch die empirische Reliabilität mittels Simulation bestimmt. Zur Analyse der faktoriellen Validität werden dann bivariate und semipartielle Korrelationen (Kim, 2012a, 2015, R-Paket `ppcor`) der bipolaren Scores mit den MVSQ^A bzw. MVSQ^V Scores berechnet.

Im Folgenden werden die hypothetisierten Zusammenhänge zwischen den bipolaren und orthogonalen Scores erläutert. Dabei wird vereinfachend die Annahme getroffen, dass die einzelnen Messungen der Wertesysteme (Annäherungs-, Vermeidungs-, und bipolare Wertesysteme) ohne Messfehler erfolgen (vgl. Russell & Carroll, 1999). Annäherung wird durch den Buchstaben *A*, Vermeidung durch *V* und bipolar durch *b* indiziert.

Im Falle von Bipolarität der Annäherungs- und Vermeidungsdimension gilt:

- Die Korrelationen zwischen Annäherungs- und bipolaren Scores desselben Wertesystems sind perfekt positiv: $r_{Ab} = 1$
- Aufgrund der negativen Kodierung der Items der Vermeidungswertesysteme gilt das Gegenteil für den Zusammenhang von Vermeidungs- und bipolaren Scores: $r_{Vb} = -1$.
- Die semipartielle Korrelation zwischen Annäherungs- und bipolaren Scores eines Wertesystems, aus der die Vermeidungs-Scores auspartialisiert wurden, verringern sich im Vergleich zu den bivariaten Scores: $r_{Ab.V} < r_{Ab}$. Analoges gilt für die semipartiellen Korrelationen zwischen Vermeidungs- und bipolaren Scores: $r_{Vb.A} < r_{Vb}$.

Im Falle von Orthogonalität gilt:

- Die Korrelationen zwischen Annäherungs- und bipolaren Scores desselben Wertesystems liegen bei $r_{Ab} = .71$.
- Wieder aufgrund der Kodierung liegen die Korrelationen zwischen Vermeidungs- und bipolaren Scores desselben Wertesystems bei $r_{Vb} = -.71$.
- Die semipartiellen Korrelationen verändern sich nicht, wenn Annäherungs- und Vermeidungsdimension unabhängig sind: $r_{Ab.V} = r_{Ab} = .71$ bzw. $r_{Vb.A} = r_{Vb} = -.71$

Die Korrelation von .71 erklärt sich dadurch, dass dieser Wert einer Varianzaufklärung (R^2) von .50 bzw. exakt 50% entspricht. Eben diese wäre zu erwarten, wenn Annäherungs- und Vermeidungs-Scores desselben Wertesystems vollständig unabhängig wären.

9.1.2 Ergebnisse

Im Ergebnisteil werden nun zuerst die Skaleninterkorrelationen berichtet. Als zweites wird die explorative Faktorenanalyse durchgeführt und als drittes die Kenndaten des bipolaren TIRT-Modells sowie der Vergleich der bipolaren Scores mit den Annäherungs- und Vermeidungs-Scores dargestellt.

9.1.2.1 Skaleninterkorrelationen

Die Skaleninterkorrelationen (Tabelle 36) können innerhalb der Subskalen (oberes bzw. unteres Dreieck) als niedrig bis moderat bezeichnet werden, da keine Interkorrelation über .46 bzw. unter -.37 liegt. Die durchschnittliche Interkorrelation der Annäherungswertesysteme beträgt dabei -.02. Die Interkorrelationen der Vermeidungswertesysteme liegen zwischen -.35 und .44 mit M = -.03. Allgemein ist in beiden Subskalen ein ähnliches Muster zu erkennen, indem alle Koeffizienten ähnliche Größen und entgegengesetzte Richtungen aufwiesen. Auch das eingangs hypothetisierte Muster bestätigt sich somit tendenziell. Bei den Annäherungswertesystemen liegt die mittlere Korrelation zwischen den *express-self*-Systemen nach Fisher's z-Transformation bei $r_{express} = .12$ und den *sacrifice-self*-Systemen bei $r_{sacrifice} = .22$. Die Korrelation zwischen den Wertesystemgruppen beträgt $r_{zwischen} = -.17$. Die Korrelationen der Vermeidungswertesysteme gestalteten sich wie folgt: $r_{express} = .12$, $r_{sacrifice} = .16$ und $r_{zwischen} = -.20$.

Auch bei den Korrelationen zwischen den Skalen konnten ähnliche Muster festgestellt werden. Die Wertesysteme ***Geborgenheit***, ***Gewissheit*** und ***Gleichheit*** korrelieren jeweils negativ mit den Wertesystemen der anderen Skala und auch für ***Macht*** und ***Erfolg*** gilt diese Aussage. Zwischen diesen beiden Gruppen sind die Korrelationen jedoch positiv zwischen den Skalen. Interessant sind bei den Korrelationen zwischen den Skalen vor allem die Korrelationen der zwei Seiten (Annäherung und Vermeidung) „derselben" Wertesysteme. Diese sind durchwegs negativ im mittleren bis hohen Bereich (zwischen -.46 und -.68).

Tabelle 36. Skaleninterkorrelationen.

	GB^A	MA^A	GW^A	ER^A	GL^A	VE^A	NA^A	GB^V	MA^V	GW^V	ER^V	GL^V	VE^V
GB^A													
MA^A	-.23												
GW^A	.46	-.12											
ER^A	-.13	.27	.00										
GL^A	.42	-.23	.17	-.37									
VE^A	-.22	.01	-.17	.06	-.07								
NA^A	.05	-.20	-.14	-.33	.28	.09							
GB^V	-.55	.30	-.37	.25	-.36	.29	.01						
MA^V	.34	-.67	.25	-.16	.34	.04	.17	-.30					
GW^V	-.32	.20	-.68	.07	-.07	.20	.17	.44	-.19				
ER^V	.33	-.27	.10	-.46	.54	.02	.33	-.28	.40	-.03			
GL^V	-.27	.33	-.13	.29	-.50	.14	-.16	.30	-.30	.12	-.35		
VE^V	.16	.05	.07	.04	.10	-.66	-.04	-.21	-.04	-.06	-.02	-.07	
NA^V	.08	.13	.19	.23	-.17	-.23	-.48	.04	-.10	-.16	-.18	.16	.30

Anmerkung. Interkorrelationen der $MVSQ^A$ Skala über der Diagonalen, $MVSQ^V$-Interkorrelationen unter der Diagonalen; GB = Geborgenheit; MA = Macht; GW = Gewissheit; ER = Erfolg; GL = Gleichheit; VE = Verstehen; NA = Nachhaltigkeit; A = Annäherung; V = Vermeidung.

Die mittlere Korrelation zwischen den *express-self*-Wertesystemen geht bei beiden Subskalen jedoch ausschließlich auf die Korrelation zwischen den Wertesystemen ***Macht*** und ***Erfolg*** zurück. ***Verstehen*** ist fast unkorreliert zu ***Macht*** und ***Erfolg***, aber negativ korreliert zu den *sacrifice-self*-Wertesystemen, vor allem zu ***Geborgenheit*** und ***Gewissheit***. Ein ähnliches Muster liegt bei den *sacrifice-self*-Wertesystemen vor, denn ***Nachhaltigkeit*** korreliert negativ mit ***Gewissheit*** und fast nicht mit ***Geborgenheit***. Auch hier gilt diese Beobachtung für beide Subskalen. Die Wertesysteme ***Verstehen*** und ***Nachhaltigkeit*** bilden demzufolge eine dritte Gruppe positiv korrelierter Wertesysteme (Annäherung $r = 0.09$; Vermeidung $r = 0.09$), die zudem in negativem Zusammenhang mit allen anderen Wertesystemen mit Ausnahme von ***Gleichheit*** steht. Die durchschnittlichen Korrelationen dieser beiden Wertesysteme mit den Wertesystemen ***Geborgenheit***, ***Macht***, ***Gewissheit*** und ***Erfolg*** betragen nach z-Transformation auf der Annäherungsdimension $r = -.12$ und auf der Vermeidungsdimension $r = -.09$).

Die Korrelationen zwischen der Annäherungs- und Vermeidungsdimension derselben Wertesysteme (z.B. ***Geborgenheit***^{A} mit ***Geborgenheit***^{V}) liegen zwischen $r = -.46$ und $-.68$ und sind damit zwar absolut gesehen merklich höher als Korrelationen zwischen unterschiedlichen Wertesystemen, dennoch „nur“ dem moderat (negativen) Bereich zuzuordnen. Bei perfekter Messgenauigkeit und absoluter Orthogonalität wären hier Korrelationen von 0 zu erwarten. Davon weichen diese Korrelationen deutlich ab. Bei Bipolarität wären hier Korrelationen von -1 zu erwarten. Von den sieben entsprechenden Korrelationen liegen vier leicht über der Mitte von $-.50$ und zwei leicht darunter und eine betrug exakt $-.50$. Aus diesen Koeffizienten kann somit keine klare Aussage für oder gegen die Orthogonalitätshypothese abgeleitet werden.

9.1.2.2 Explorative Faktorenanalyse

Mit der explorativen Faktorenanalyse wird untersucht werden, wie viele Faktoren zweiter Ordnung benötigt werden, um die 14 Dimensionen des MVSQ zu erklären. Dadurch sollen zudem Rückschlüsse zur Frage bzgl. Orthogonalität versus Bipolarität gezogen werden.

Zunächst wurden die Voraussetzungen zur Durchführung einer EFA auf die 14 Wertesysteme geprüft. Sowohl der Kaiser-Meyer-Olkin (KMO)-Koeffizient mit .72, als auch die Measure of Sample Adequacy-Koeffizienten der einzelnen Variablen (zwischen .54 und .84) können nach Bühner (2011, S. 347) als ausreichend hoch eingestuft werden. Die Daten sind nach den Richtlinien bei Bühner „mittelmäßig“ geeignet, um eine Faktorenanalyse durchzuführen. Der Bartlett-Test auf Sphärizität wurde signifikant, $\chi^2(91) = 3104; p \leq .001$. Alle drei Kriterien sprechen also dafür, dass es lineare Zusammenhänge zwischen den Variablen gibt und Faktoren extrahiert werden können.

Zur Bestimmung der Anzahl der zu extrahierenden Faktoren wurden mehrere Richtlinien betrachtet. Scree-Plot und Kaiser-Kriterium (Eigenwerte > 1 Kaiser, 1960) sprachen beide für vier zu extrahierende Faktoren. Die Parallelanalyse nach Horn (1965) schlug fünf Faktoren vor und nach dem Jolliffe-Kriterium (Eigenwerte > .7; Jolliffe, 1986) wären sieben Faktoren zu extrahieren gewesen. Da diese Empfehlungen deutlich variierten, wurde wie von Eid et al. (2015) vorgeschlagen, die Anzahl der Faktoren schrittweise um eins erhöht und die Modelle miteinander verglichen.

Da der χ^2-Test bei großen Stichproben leicht dazu führen kann, dass eigentlich passenden Modelle verworfen werden (Eid et al., 2015), wurde als Kriterium der Anpassungsgüte zusätzlich der *Root Mean Square Error of Approximation* (RMSEA) berechnet und berichtet. Außerdem wurden das *Bayesian Information Criterion* (BIC) und der *Tucker Lewis Index of factoring reliability* (TLI) herangezogen, um die Modelle relativ zueinander zu vergleichen (Eid et al., 2015). In Tabelle 37 sind die entsprechenden Modellgütekoeffizienten aufgeführt.

Da bei psychologischen Konstrukten grundsätzlich nicht von Unabhängigkeit ausgegangen werden kann (Fabrigar et al., 1999) und im MVSQ teils mittlere Skaleninterkorrelationen vorliegen, wäre oblique Rotation der orthogonalen Rotation vorzuziehen. Damit traten bei der

Faktorextraktion je nach Schätzer jedoch bereits bei vier (OLS-Schätzer) bzw. fünf (ML-Schätzer) Faktoren Heywood-Fälle auf. Als Heywood-Fall wird bezeichnet, wenn die Residualvarianz unzulässigerweise einen negativen Wert annimmt (Heywood, 1931). Sie treten häufig dann auf, wenn nicht genügend Input-Variablen vorhanden sind (McDonald, 1985), was im vorliegenden Fall mit nur 14 Wertesystemen wahrscheinlich die Ursache darstellte. Auch mit der Methode der generalisierten kleinsten Quadrate (GLS), die speziell dafür entwickelt wurde, effektiver mit diesen Fällen umzugehen (Joreskog & Goldberger, 1972), traten Heywood-Fälle auf (bei sechs Faktoren). Da keines der Modelle eine zufriedenstellende Anpassungsgüte aufwies, wurde letztlich die orthogonale Rotationsmethode *Varimax* gewählt.

Tabelle 37. Anpassungsgüte der EFA-Modelle.

Anzahl Faktoren	Chi-Quadrat χ^2 (df), p	RMSEA (KI)	BIC	TLI
1	1839 (77), <.001	.19 (.18 - .20)	1343.97	.31
2	1134 (64), <.001	.17 (.16 - .17)	722.21	.49
3	694 (52), <.001	.14 (.13 - .15)	360.11	.63
4	390 (41), <.001	.12 (.11 - .13)	126.12	.74
5	263 (31), <.001	.11 (.10 - .12)	64.23	.77
6	161 (22), <.001	.10 (.09 - .12)	19.42	.81
7	76 (14), <.001	.09 (.07 - .10)	-13.48	.86
8	23 (7), <.01	.06 (.03 - .09)	-22.11	.93
9	4 (1), <.05	.07 (.01 - .15)	-2.52	.91

Anmerkung. RMSEA = Root Mean Square Error of Approximation; BIC = Bayesian Information Criterion; TLI = Tucker-Lewis Index.

Ab zehn Faktoren sind die Modelle unteridentifiziert (df< 0) und konnten deshalb nicht mehr geschätzt werden. Tabelle 37 zeigt die Modellgütekoeffizienten der erfolgreich geschätzten Modelle. Diese werden im Folgenden gemäß der drei Kriterien der Anpassungsgüte, Sparsamkeit und Interpretierbarkeit bewertet (Eid et al., 2015). Insgesamt lässt sich sagen, dass die χ^2-Tests (alle $p < .05$) bei allen Modellen ungenügende Modellanpassung suggerieren. Dies ist jedoch bei großen Stichproben wenig verwunderlich, da beim χ^2-Test bei großen Stichproben bereits geringe Modellabweichungen zur Ablehnung der Nullhypothese führen (Eid et al., 2015). Nach den RMSEAs hingegen stellt das Modell mit acht Faktoren einen zufriedenstellenden Fit dar. Es ist das einzige, dessen RMSEA nicht über der akzeptablen Grenze von .06 (Hu & Bentler, 1999) liegt. Obwohl der TLI des Modells mit acht Faktoren knapp unter der empfohlenen Grenze von .95 (Hu & Bentler, 1999) ist, kann gesagt werden, dass abgesehen von diesem Modell, keines gut auf die Daten passt.

Beim Vergleich der Modelle untereinander mittels BIC und TLI kann die Tendenz beobachtet werden, dass mit zunehmender Anzahl Faktoren (bis acht Faktoren) die Anpassungsgüte der Modelle zunimmt (BIC sinkt mit steigender Zahl Faktoren, TLI nimmt zu). Erst ab dem neunten Faktor wird diese wieder geringer. Auch im relativen Vergleich weist Modell 8 somit die besten Gütekoeffizienten auf. Die acht Faktoren dieses Modells erklärten zusammen 72% der Gesamtvarianz.

Tabelle 38. Standardisierte Faktorladungsmatrix des Modells mit acht Faktoren.

	F1	F2	F3	F4	F5	F6	F7	F8
GB^A	.46	-.33		.35				
GB^V	-.23	.88		-.29				
MA^A			.92					
MA^V	.36		-.68					
GW^A				.97				
GW^V		.24		-.68				
ER^A	-.27				.84			
ER^V	.66				-.26			
GL^A	.65					-.32		
GL^V	-.30					.74		
VE^A							-.66	
VE^V							.99	
NA^A	.29							-.52
NA^V							.23	.86

Anmerkung. F1 bis F8 bezeichnen die extrahierten Faktoren. Ladungen < .2 wurden ausgeblendet.

Für die Beurteilung hinsichtlich des Kriteriums der Interpretierbarkeit werden die Faktorladungen des Modells mit acht Faktoren in Tabelle 38 aufgeführt. Dabei kann gesehen werden, dass nur die Faktoren F3, F4 und F7 nach Bortz und Schuster (2010, S. 422) hinreichend große standardisierte Faktorladungen aufweisen (> .60), um interpretierbar zu sein, wenngleich die Mindestzahl von vier Ladungen hier nicht erreicht wird. Lässt man dieses Kriterium außer Acht, dann zeigt sich jedoch ein relativ klares Muster, wonach die Faktoren F2 bis F8 jeweils der Annäherungs- und Vermeidungsdimension desselben Wertesystems zugeordnet werden können und Faktor F1 durch relativ viele Facetten gekennzeichnet ist. Von den Ladungen aus Faktor F1 können die Ladungen von ***Geborgenheit***^{A}, ***Macht***^{V}, ***Erfolg***^{V} und ***Gleichheit***^{A}

interpretiert werden (> .32 Tabachnick & Fidell, 2007, S. 649). Unter Berücksichtigung derer Vorzeichen erinnert dieser Faktor an die Einteilung in *sacrifice/express-self*-Wertesysteme.

Insgesamt können aus dieser EFA zwei Schlussfolgerungen gezogen werden. Die eine lautet, dass trotz der ungenügend großen Faktorladungen zumindest die Tendenz ausgemacht werden kann, dass die Wertesysteme der Annäherungs- und Vermeidungsdimension zusammen je einen Faktor beschrieben. Diese Tendenz spricht eher für Bipolarität und gegen Orthogonalität. Abgesehen davon kann festgestellt werden, dass die sieben Faktoren, die die Wertesysteme repräsentieren, relativ deutlich voneinander unterscheidbar sind, da sie als andeutungsweise eigenständige Faktoren identifiziert wurden. Dies spricht somit für die faktorielle Validität von zumindest sieben Wertesystemen.

9.1.2.3 Bipolare TIRT-Modelle

Für den dritten Teil dieser Untersuchung wurde ein TIRT-Modell spezifiziert, in dem sieben Items in jedem der 20 Blöcke auf sieben bipolar konzeptualisierte Wertesysteme laden. Dieses Modell wird als *bipolares* Modell bezeichnet und auch die damit gemessenen Merkmalsausprägungen als *bipolare* Wertesystemausprägungen bzw- Scores.

Der Schätzalgorithmus hat erfolgreich konvergiert und der RMSE liegt mit 0.11 zwischen den RMSEs des Annäherungs- und Vermeidungsmodells (vgl. Kapitel 6.2), also im akzeptablen Bereich. Die empirischen Reliabilitäten des bipolaren Modells (Tabelle 39) sind deutlich höher als die der beiden Subskalen mit Werten zwischen $\rho_{md} = .74$ und $.84$. Da die Input-Daten des bipolaren Modells exakt dieselben der beiden orthogonalen TIRT-Modelle sind, liegt der Verdacht nahe, dass Veränderungen der empirischen Reliabilitäten eher methodischen als empirischen Ursprungs sind. Laut Brown und Maydeu-Olivares (2011, 2013) ist eine größere Anzahl Items gleichbedeutend mit mehr Information, weshalb die Modellparameter und Scores präziser geschätzt werden können. Dies könnte eine Erklärung für die hier auftretenden höheren empirischen Reliabilitäten darstellen.

Um die Eingangs formulierten Hypothesen bzgl. der Orthogonalitätshypothese zu überprüfen, wurden bivariate Produkt-Moment-Korrelationen und semipartielle Korrelationen zwischen den bipolaren und den orthogonalen Wertesystemen berechnet. Diese sind in Tabelle 40 zusammengefasst. Aus der semipartiellen Korrelation zwischen Annäherungs- und bipolaren Scores wurden die Vermeidungswertesystem-Scores auspartialisiert und aus den Korrelationen zwischen Vermeidungs- und bipolaren Scores logischerweise die Annäherungswertesystem-Scores.

Zur Interpretation der bivariaten Korrelationen ist zu sagen, dass diese in den meisten Fällen näher an der Mitte als an einer dem beiden hypothetisierten Werte liegen. $r = \pm.86$ stellt diese Mitte zwischen den eingangs erläuterten hypothetischen Korrelationen dar. Damit lagen nur die bivariaten Korrelationen für ***Gleichheit***^{A} ($r = .78$) und ***Nachhaltigkeit***^{V} ($r = -.70$) klar darunter sowie für ***Nachhaltigkeit***^{A} ($r = .95$) sowie ***Macht***^{V} und ***Gewissheit***^{V} (beide

Tabelle 39. Empirische Reliabilitäten des bipolaren TIRT-Modells.

Wertesystem	ρ_{md}	Spanne ρ		
Geborgenheit	.73	.71	-	.75
Macht	.80	.78	-	.83
Gewissheit	.83	.80	-	.84
Erfolg	.77	.73	-	.80
Gleichheit	.70	.68	-	.74
Verstehen	.76	.73	-	.78
Nachhaltigkeit	.76	.72	-	.79

Anmerkung. ρ_{md} = Median empirischen Reliabilität über zehn Replikationen; Spanne ρ = Spanne über zehn Replikationen; GB = Geborgenheit; MA = Macht; GW = Gewissheit; ER = Erfolg; GL = Gleichheit; VE = Verstehen; NA = Nachhaltigkeit.

Tabelle 40. Bivariate und semipartielle Korrelationen.

	GB	MA	GW	ER	GL	VE	NA
r_{Ab}	.87	.89	.90	.83	.78	.90	.95
$r_{Ab.V}$	.80	.70	.70	.84	.78	.70	.86
r_{Vb}	-.88	-.92	-.92	-.85	-.91	-.91	-.70
$r_{Vb.A}$	-.80	-.71	-.71	-.85	-.83	-.71	-.77

Anmerkung. r_{AB} = Korrelationen der Annäherungs- mit bipolaren Wertesystemen; $r_{AB.V}$ = Semipartielle Korrelationen der Annäherungs- mit bipolaren Wertesystemen; r_{VB} = Korrelationen der Vermeidungs- mit bipolaren Wertesystemen; $r_{VB.A}$ = Semipartielle Korrelationen der Vermeidungs- mit bipolaren Wertesystemen.

$r = -.92$) deutlich darüber. Alle anderen Korrelationen lagen in einem Bereich $.86 \pm .05$ Wobei insgesamt zumindest eine leichte Tendenz hin zur Bipolarität gesehen werden kann, da mehr Korrelationen betragsmäßig über als unter diesem Grenzwert liegen.

Bei den semipartiellen Korrelationen ist das Bild etwas klarer. Hier kann gesagt werden, dass die Wertesysteme beider Dimensionen von ***Geborgenheit***, ***Macht***, ***Gewissheit*** und ***Verstehen*** eher bipolar zueinander standen, da die semipartiellen Korrelationen deutlich niedriger als die bivariaten Gegenstücke sind. Die Korrelationen bei ***Erfolg*** und ***Gleichheit*** sprechen eher für Orthogonalität, da sie jeweils entweder ähnlich groß oder betragsmäßig kleiner zu den entsprechenden bivariaten Korrelationen sind. Bei diesen sechs Wertesystemen gelten die Resultate für beide Dimensionen, einzig bei ***Nachhaltigkeit*** unterscheiden sich die Ergebnisse in den Dimensionen. ***Nachhaltigkeit***^{A} schien eher bipolar, da sich die semipartielle Korrelation verringerte und ***Nachhaltigkeit***^{V} eher orthogonal.

9.1.3 Diskussion

In der Gesamtschau kann aus den Ergebnissen nicht nur gefolgert werden, dass der MVSQ faktorielle Validität besitzt, sondern auch das Wertemodell von Graves konzeptuelle Gültigkeit aufweist, da die Befunde darauf hinweisen, dass es sieben Wertesysteme gibt, die durch das Instrument gemessen werden. Des Weiteren verdeutlicht die unklare Lage zur Orthogonalitätsfrage, dass eine Erhöhung der Messgenauigkeit des MVSQ auch deshalb wünschenswert wäre, um diese Fragestellung mit höherer Präzision untersuchen zu können. Bei den aktuell vorliegenden Reliabilitäten kann nicht gesagt werden, ob bzw. welchen Einfluss die Messungenauigkeit auf die Untersuchungen zur Orthogonalität hat. Außerdem stellt sich die Frage, ob konzeptuell ein Mittelweg zwischen Orthogonalität und Bipolarität vorstellbar ist. Auch zur tiefer gehenden Untersuchung dieser Frage auf empirischer Basis ist ein Instrument mit höheren Reliabilitäten notwendig.

Des Weiteren ist kritisch zu sehen, dass sich in der Faktorenanalyse ein Modell als passend erwiesen hat, in dem zwar sieben Faktoren die bipolaren Wertesysteme repräsentierten, jedoch darüber hinaus ein achter Faktor gefunden wurde. Zwar könnte dieser Faktor andeutungsweise als Indikator für die *express/sacrifice-self* -Dimension interpretiert werden, denkbar ist jedoch auch, dass sich dieser Faktor aus einer unzureichenden Messgenauigkeit ergibt und somit als weiterer Anlass zur Überarbeitung des Instruments gesehen werden kann.

Zudem sei angemerkt, dass es sich bei der Stichprobe um dieselbe Stichprobe handelte, die bereits zur Entwicklung und Schätzung der TIRT-Modelle verwendet wurde. Dies kann insofern kritisiert werden, dass es die Generalisierbarkeit der Ergebnisse einschränkt. Eine Ausdehnung der Untersuchung auf andere Stichproben mit bevölkerungsrepräsentativeren Zusammensetzungen wäre an dieser Stelle wünschenswert.

9.2 Studie zur konvergenten Validität

Konvergente Validität eines Konstrukts liegt bekanntlich dann vor, wenn die Messungen des Konstrukts mit mehreren Methoden übereinstimmen, d.h. wenn verschiedene Operationalisierungen desselben Konstrukts zum selben Messergebnis führen (Bortz & Döring, 2006). Als Königsweg der Überprüfung der konvergenten (und auch divergenten) Validität gilt die von Campbell und Fiske (1959) vorgeschlagene Multitrait-Multimethod-Analyse (Amelang & Schmidt-Atzert, 2006). Dazu werden die untersuchten Konstrukte mit unterschiedlichen Methoden (z.B. Ranking und Rating oder Selbstauskunft und Fremdbeobachtung) gemessen, die dann in einer sog. Multitrait-Multimethod-Matrix zusammengefasst dargestellt werden. Im vorliegenden Fall handelt es sich nicht um eine MTMM-Analyse im eigentlichen Sinn, da keine parallelen Instrumente gefunden werden konnten, die Wertesysteme nach Graves messen würden. Stattdessen wurde der Schwartz Value Survey (SVS) herangezogen, der – wie nachfolgend gezeigt wird – zumindest konzeptuell ähnliche Konstrukte misst.

9.2.1 Methode

Ziel dieser Analyse war die Untersuchung des MVSQ auf seine konvergente Validität. Dazu wurden die MVSQ-Scores mit denen des weit verbreiteten und häufig eingesetzten Schwartz Value Surveys (SVS) kontrastiert. Beide Instrumente wurden dazu in Stichprobe IIIa (N = 104, vgl. Kapitel 3.7)) erhoben. Die Studierenden dieser Stichprobe wurden in verschiedenen Veranstaltungen rekrutiert und bearbeiteten nacheinander sowohl den MVSQ als auch den SVS. Bei der Anwerbung wurden die potenziellen Teilnehmer über den Hintergrund der Studie aufgeklärt. Alle Teilnehmer erhielten als Gegenleistung für ihre Teilnahme Rückmeldungen in Form von schriftlichen Berichten zu beiden untersuchten Fragebögen.

Die Auswahl des „parallelen“ Fragebogens wurde von zwei Kriterien geleitet. Erstens sollten die erhobenen Wertedimensionen mit denen des MVSQ vergleichbar sein und zweitens sollte es sich um ein valides Instrument handeln. Beide Bedingungen sind beim SVS erfüllt, wenngleich klar ist, dass es sich nicht um ein wirklich paralleles Instrument handelt. Unter Berücksichtigung des Fehlens eines echt parallelen Fragebogens, kann der SVS jedoch als ausreichend eingestuft werden, um zur initialen Überprüfung der konvergenten Validität herangezogen werden zu können. Tabelle 41 zeigt eine Gegenüberstellung der beiden den Instrumenten zugrunde liegenden Konzeptionen der Wertesysteme. Um einfacher zwischen den Theorien und Konstrukten unterscheiden zu können, wird im folgenden bei den Konstrukten im SVS von Werte-Typen und bei denen des MVSQ von Wertesystemen gesprochen, wenngleich diese Bezeichnungen austauschbar sind (Schwartz, 1996).

Bei einigen Wertesystem-Werte-Typ-Kombinationen wird deren Ähnlichkeit bereits durch die Bezeichnungen erkennbar, z.B. gibt es in beiden Skalen *Macht* als Kategorie. Beruft man sich nun auf die in Kapitel 2.2 dargestellten Beschreibungen der Wertesysteme, können folgende

Tabelle 41. Vergleich der Wertekonstrukte nach Graves und Schwartz.

Graves	Schwartz	Beschreibung nach Schmidt et al. (2007)
Geborgenheit	Tradition	Respekt vor, Verbundenheit mit und Akzeptanz von Gebräuchen und Ideen, die traditionelle Kulturen und Religionen für ihre Mitglieder entwickelt haben
Macht	Macht	Sozialer Status und Prestige, Kontrolle oder Dominanz über Menschen und Ressourcen
Gewissheit	Konformität	Beschränkung von Handlungen und Impulsen, die andere beleidigen oder verletzten könnten oder gegen soziale Erwartungen und Normen verstoßen
Erfolg	Leistung	Persönlicher Erfolg durch die Demonstration von Kompetenz bezüglich sozialer Standards
Gleichheit	Benevolenz	Bewahrung und Erhöhung des Wohlergehens der Menschen, zu denen man häufigen Kontakt hat
Verstehen	Selbstbestimmung	Unabhängiges Denken und Handeln, schöpferisch Tätigsein, erforschen
Nachhaltigkeit	Universalismus	Verständnis, Wertschätzung, Toleranz und Schutz des Wohlergehens aller Menschen und der Natur
	Hedonismus	Vergnügen und sinnliche Belohnungen für einen selbst
	Stimulation	Aufregung, Neuheit und Herausforderungen im Leben
(Gewissheit)	Sicherheit	Sicherheit, Harmonie und Stabilität der Gesellschaft, von Beziehungen und des Selbsts

Anmerkung. Die Reihenfolge der Schwartz-Werte-Typen wurden entsprechend der hypothetisierten Kongruenz zu den Gravesschen Wertesystemen angepasst.

Schlussfolgerungen zur Parallelität bzw. Ähnlichkeit der Werte-Konzeptualisierungen gezogen werden:

***Geborgenheit*-Tradition.** Diese Konzepte stimmen von der Definition her zumindest in einem Bereich überein. In beiden Konzeptualisierungen ist von Traditionen, Gebräuchen und Verbundenheit die Rede. Beim MVSQ-Wertesystem liegt der Fokus jedoch mehr auf dem Verbundenheitsaspekt und die Orientierung zur engen Gruppe hin. Der SVS Werte-Typ stellt Tradition in den Vordergrund.

***Macht*-Macht.** Obgleich der Titel des Wertesystems bzw. Werte-Typs gleich ist, stimmen die Beschreibungen nur teils überein. Ähnlich ist bei beiden die Orientierung hin zu sozialer

Dominanz über Menschen. Der Aspekt des „sozialen Status“ und „Prestige“ im SVS-Macht-Typ ist in der MVSQ-Konzeptualisierung eher dem Wertesystem Erfolg zuzuordnen.

Gewissheit*-Konformität.** Übereinstimmend kann hier die Beschränkung von Handlungen durch die Einhaltung von Regeln und Normen gesehen werden. Der Aspekt, dass die Nichteinhaltung von Regeln und Normen mit der Beleidigung oder Verletzung anderer einhergeht, ist im Wertesystem ***Gleichheit nicht enthalten. Dafür geht es darin mehr um Klarheit, Voraussehbarkeit und das Halten an Vorgaben.

Erfolg*-Leistung.** ***Erfolg im MVSQ enthält über die Konzeptualisierung von Leistung hinaus des Weiteren die Orientierung zum Wettbewerb und dem damit verbunden möglichen Status- und Prestige-Gewinn.

Gleichheit*-Benevolenz.** Gemeinsam ist den beiden Konzepten die Orientierung hin zu Menschen und deren Wohlergehen im ihren sozialen Umfeldern. Abgesehen davon enthält das ***Gleichheit-Wertesystem noch die Werte Kommunikation und Netzwerke, die so nicht unbedingt zu Benevolenz passen müssen.

Verstehen*-Selbstbestimmung.** Beide Konstrukte stimmen überein in der Orientierung hin zur Unabhängigkeit, schöpferischem Tätig-sein und erforschen. Im ***Verstehen-Wertesystem wird das als *Verstehen wollen* beschrieben.

Nachhaltigkeit*-Universalismus.** Diese Konstrukte beziehen sich im Unterschied zu allen übrigen Werte-Typisierungen auf alle Menschen gleichermaßen und teilen den Bezug zum Globalen. Das Wertesystem ***Nachhaltigkeit beinhaltet darüber hinaus mehr Fokus auf Zukunftsorientierung und ökologische Themen.

Bei diesen sieben Wertesystem-Werte-Typ-Paaren kann eine gewisse Parallelität festgestellt werden, wenngleich in allen Vergleichen auch Unterschiede ausgemacht wurden.

Der SVS erhebt des Weiteren die Werte-Typen Hedonismus, Stimulation und Sicherheit. Bei Hedonismus stellt sich die Frage, ob es sich dabei um einen Werte-Typen handelt, der im MVSQ entweder schlicht nicht enthalten ist, also *keinem* Wertesystem zuordenbar ist oder ob es sich dabei um ein Konstrukt handeln könnte, das *allen* Wertesystemen zugeordnet werden kann. Zweiteres wäre dann schlüssig, wenn Hedonismus nicht ein Werte-, sondern ein übergeordnetes Konstrukt wäre. Da Hedonismus als Prinzip der Lustvermehrung und Unlustvermeidung definiert ist, kann es als solches als übergeordnetes Ziel menschlichen Handelns verstanden werden (Brandstätter et al., 2013). Je nach Wertesystempräferenz führen dann unterschiedliche Situationen bzw. Handlungen zu Lust oder Unlust. Eine Person mit einer hohen Präferenz des Wertesystems ***Verstehen*** wird die theoretische Erforschung mit

einem komplexen Thema vermutlich als Lust wahrnehmen, wohingegen eine Person mit eine Präferenz von ***Macht*** die theoretische Auseinandersetzung eher als Unlust interpretieren dürfte und stattdessen vermutlich eine praktische Lösung suchen würde. Auch beim Werte-Typ Stimulation ist denkbar, dass er als übergeordnete Dimension verstanden werden kann. In jedem Fall kann dieser Werte-Typ keinem Wertesystem so zugeordnet werden, wie das bei den anderen Werte-Typen der Fall war. Zu guter Letzt der Werte-Typ Sicherheit. Dieser könnte eine Gemeinsamkeit mit dem Wertesystem ***Gewissheit*** haben und zwar die Aspekte von Sicherheit und Stabilität in der Gesellschaft. Diese Werte passen am ehesten zu ***Gewissheit***, da Sicherheit und Stabilität eng mit der Einhaltung von Regeln, Gesetzen und Normen zusammenhängen und dem Wunsch nach Gewissheit und Vorhersehbarkeit zuträglich sind.

Multitrait-Multimethod-Analyse (MTMM)

Bei der MTMM beschreiben die *Monotrait-Heteromethod*-Korrelationen die Zusammenhänge derselben Konstrukte, die mit unterschiedlichen Methoden gemessen wurden (MVSQ und SVS, siehe Tabelle 41). Diese Korrelationen liegen auf der sogenannten Validitäts-Diagonale (Campbell & Fiske, 1959). Werden diese Korrelationen signifikant, spricht dies für konvergente Validität, wobei gilt, je höher die Korrelation, umso größer die Übereinstimmung der Messung und desto höher die konvergente Validität (Schmitt & Stults, 1986). Außerdem sollten diese Korrelationen höher als die *Heterotrait-Heteromethod*-Korrelationen sein. Des Weiteren sollten die Werte in der Validitäts-Diagonale auch größer sein als die übrigen Skaleninterkorrelationen (Schmitt & Stults, 1986), die man auch als *Heterotrait-Monomethod*-Korrelationen bezeichnen könnte. Es fehlt noch zu erwähnen, dass in MTMM-Matrizen normalerweise Reliabilitäten berichtet werden (Campbell & Fiske, 1959). Diese wurden für den MVSQ bereits in Kapitel 7 berichtet und werden für den SVS separat in einer Voruntersuchung aufgeführt.

Materialien

Neben dem MVSQ wurde der Schwartz Value Survey (SVS) ausgegeben. Das Instrument setzt sich aus zwei Wertelisten (instrumentelle und terminale Werte) zusammen, die aus insgesamt 57 Items bestehen und die zehn Werte-Typen nach Schwartz messen. Die Anleitung und Itemformulierungen wurden aus Glöckner-Rist (2012) entnommen und in einer Online-Form umgesetzt. Jeder Einzelwert wird dabei auf einer neunstufigen Rating-Skala mit den Stufen -1 „entgegengesetzt“, 0 „nicht wichtig“, 3 „wichtig“, 6 „sehr wichtig“ und 7 „äußerst wichtig“ bewertet. Die Stufen 1, 2, 4 und 5 erhielten wie vorgegeben keine Bezeichnung. Die Validität des Instruments wurde in zahlreichen Studien untersucht und dabei wiederholt die Struktur der Werte-Typen bestätigt (Schmidt et al., 2007). Auch zur Reliabilität des Instruments gibt es Befunde, die den Werte-Typen annehmbare Reliabilitäten (> .7) bescheinigen (Vgl. Calogero

et al., 2009; Schwartz, 2005), wobei die Reliabilitäten der deutschsprachigen Version in einer Untersuchung von Sagiv und Schwartz (2000) mehrheitlich Werte unter .7 aufwiesen (M = .64).

Die beiden Bedingungen zur Auswahl des „parallelen“ Fragebogens können somit als ausreichend gegeben angesehen werden, wenngleich nicht von echter Parallelität ausgegangen werden kann, da nicht zu erwarten ist, dass sowohl gleiche wahre Werte und gleiche Fehlervarianzen (Schermelleh-Engel & Werner, 2012) bei den Messungen vorliegen. Es ist an dieser Stelle deshalb angebrachter, von Ähnlichkeit anstatt von Parallelität zu sprechen.

Insbesondere die Tatsache jedoch, dass die sieben Wertesysteme der Gravesschen Theorie eigenständig im Schwartzschen Wertemodell wiedergefunden werden können, spricht für die Verwendung des SVS in dieser Untersuchung. Mit anderen möglichen Instrumenten ist diese Bedingung nicht erfüllt: In Hofstede's (1993) Modell z.B. gibt es nur fünf Wertedimensionen und von Rokeach's (1973) Value Survey existieren (vermutlich aufgrund seiner Ipsativität) nur wenige Daten zu dessen Validität. Der SVS hingegen, wurde laut Schmidt et al. (2007) in über 200 Studien in mehr als 60 Ländern eingesetzt und dessen Validität kann auch für die deutschsprachige Version des Instruments als gesichert angesehen werden.

9.2.2 Ergebnisse Voruntersuchung

In einer Voruntersuchung wurden die internen Konsistenzen des SVS in der vorliegenden Stichprobe berechnet. Dazu wurde das R-Paket `psychometric` verwendet (Fletcher, 2010). Diese befanden sich zwischen $\alpha = .55$ und .79 (M = .67), wobei sich die internen Konsistenzen der einzelnen Werte-Typen wie folgt aufschlüsseln: $\alpha_{TR} = .56$, $\alpha_{MA} = .73$, $\alpha_{KO} = .68$, $\alpha_{LE} = .74$, $\alpha_{BE} = .65$, $\alpha_{SB} = .55$, $\alpha_{UN} = .79$, $\alpha_{HE} = .68$, $\alpha_{ST} = .72$ und $\alpha_{SI} = .57$. Zwar bewegen sich diese internen Konsistenzen in der Größenordnung, die auch von Sagiv und Schwartz (2000) für das deutschsprachige Instrument berichtet hat, aber insbesondere die Messgenauigkeiten der Werte-Typen Tradition (TR), Selbstbestimmung (SB) und Sicherheit (SI) können als problematisch eingestuft werden. In einer Meta-Analyse über 60 Studien weisen neben Selbstbestimmung (SB) genau diese drei Werte-Typen unterdurchschnittliche αs auf (Parks-Leduc et al., 2015), was andererseits wiederum für die Gültigkeit der hier vorliegenden Ergebnisse spricht.

9.2.3 Ergebnisse Hauptuntersuchung

Anders als von Campbell und Fiske (1959) vorgeschlagen, wurden hier nicht alle Korrelationen in *einer* Matrix zusammengefasst, sondern aufgrund der großen Anzahl an MVSQ-Wertesystemen in drei Matrizen aufgeteilt: Je eine Matrix zeigt die Skaleninterkorrelationen des SVS (Tabelle 42) bzw. MVSQ (Tabelle 43) und eine weitere Matrix zeigt die Heteromethod-Korrelationen (Tabelle 44).

Tabelle 42. Skaleninterkorrelationen der SVS Werte-Typen.

	TR	MA	KO	LE	BE	SB	UN	HE	ST
MA	.17								
KO	.59	.45							
LE	.18	.71	.47						
BE	.55	-.05	.43	-.02					
SB	.22	.09	.05	.25	.32				
UN	.27	-.15	.00	-.10	.52	.44			
HE	.03	.24	.08	.15	.07	.12	.16		
ST	-.15	.17	-.09	.21	-.02	.52	.34	.28	
SI	.52	.31	.67	.25	.39	.09	.27	.32	-.04

Anmerkung. SVS Werte-Typen: TR = Tradition, MA = Macht, KO = Konformität, LE = Leistung, BE = Benevolenz, SB = Selbstbestimmung, UN = Universalismus, HE = Hedonismus, ST = Stimulation, SI = Sicherheit.

Bei den Skaleninterkorrelationen fällt beim SVS vor allem eine sehr hohe Korrelation auf: Macht (MA) und Leistung (LE) korrelieren mit $r = .71$, sodass sich die Frage stellt, wie unabhängig diese Werte-Typen jeweils voneinander sind. Die übrigen Korrelationen befinden sich in einem Bereich zwischen $r = -.15$ und $r = .67$. Beim MVSQ liegen die Korrelationen innerhalb der Skalen zwischen $r = -.55$ und $r = .60$. Einzig bei den Korrelationen zwischen den Skalen weisen die entgegengesetzten Wertesysteme je einen starken negativen Zusammenhang auf.

Tabelle 44 zeigt die Korrelationen zwischen den MVSQ-Wertesystemen und den SVS-Werte-Typen. Die Korrelationen befinden sich im niedrigen bis moderaten Bereich (betragsmäßig maximales $r = .55$). Nachfolgend wird nun jedes Wertesystem anhand der in einer MTMM-Analyse verwendeten Kriterien untersucht. Diese Kriterien lauten zusammengefasst:

1. Werte auf der Validitäts-Diagonale werden signifikant
2. Werte auf der Validitäts-Diagonale > Heterotrait-Heteromethod-Korrelationen (HTHM)
3. Werte auf der Validitäts-Diagonale > Skaleninterkorrelationen

Das erste Kriterium trifft für alle Wertesystem-Paare außer ***Geborgenheit***-Tradition zu. Zwar sind die Korrelationen im Ausmaß nur im niedrigen bis moderaten Bereich, dennoch spricht dieser Befund allgemein gesehen für die konvergente Validität. Die Zusammenhänge zeigen sich dabei auf der MVSQ^A-Skala in positiver Richtung und auf der MVSQ^V-Skala in negativer Richtung und folgen einem ähnlichen (entgegengesetzten) Muster. Im Folgenden werden die Wertesysteme nacheinander behandelt:

Tabelle 43. Skaleninterkorrelationen der MVSQ Wertesysteme.

	GB^A	MA^A	GW^A	ER^A	GL^A	VE^A	NA^A	GB^V	MA^V	GW^V	ER^V	GL^V	VE^V
MA^A	-.12												
GW^A	.60	-.15											
ER^A	-.08	.17	.03										
GL^A	.37	-.35	.19	-.37									
VE^A	-.19	-.01	-.16	-.11	.04								
NA^A	.06	-.09	-.22	-.30	.28	.02							
GB^V	-.56	.18	-.55	.11	-.31	.23	.05						
MA^V	.34	-.68	.27	-.26	.47	-.01	.17	-.29					
GW^V	-.40	.20	-.78	.01	-.09	.25	.20	.58	-.21				
ER^V	.19	-.19	.10	-.61	.45	.22	.25	-.15	.31	.02			
GL^V	-.24	.37	-.18	.35	-.68	.05	-.31	.36	-.40	.14	-.42		
VE^V	.27	.08	.26	.19	.11	-.59	-.03	-.30	-.03	-.22	-.08	-.03	
NA^V	.08	.05	.26	.33	-.06	-.22	-.66	-.06	-.10	-.29	-.21	.25	.37

Anmerkung. MVSQ Wertesysteme: GB = Geborgenheit; MA = Macht; GW = Gewissheit; ER = Erfolg; GL = Gleichheit; VE = Verstehen; NA = Nachhaltigkeit; A = Annäherung; V = Vermeidung.

Geborgenheit. Weder als Annäherungs- noch als Vermeidungswertesytem wird die Korrelation mit SVS-Tradition signifikant. Obgleich dieses Kriterium nicht erfüllt ist, kann Kriterium 2 bestätigt werden, da keine der anderen Heterotrait-Heteromethod-Korrelationen größer ist als die Korrelation zwischen SVS-Tradition und ***Geborgenheit***. Diese Aussage gilt für $MVSQ^A$ und $MVSQ^V$ gleichermaßen. Stattdessen werden einige Korrelationen mit SVS-Werte-Typen in entgegengesetzter Richtung signifikant (Macht, Leistung, Selbstbestimmung und Stimulation). Dies kann so gedeutet werden, dass keiner der Werte-Typen besonders ähnlich zum ***Geborgenheit***-Wertesystem ist, sondern sich das Wertesystem als gewissermaßen Anti-Konstrukt mehrerer Werte-Typen beschreiben lässt. Beim dritten Kriterium ist die Lage klar: ***Geborgenheit*** ist ***Gewissheit*** und ***Gleichheit*** ($r = 0.6$ bzw. $r = 0.37$ für Annäherung sowie $r = 0.58$ und $r = 0.36$ für Vermeidung) wesentlich ähnlicher als irgendeinem anderen in dieser Studie enthaltenen Konstrukt.

Macht. Auf beiden MVSQ-Skalen korreliert das Wertesystem ***Macht*** signifikant mit SVS-Macht (***Macht***^{A}: $r = .22$ und ***Macht***^{V}: $r = -.39$). Auch die zweite Bedingung ist erfüllt, da kein anderer SVS-Werte-Typ höher mit ***Macht***^{A} bzw. niedriger mit ***Macht***^{V} korreliert und auch

Tabelle 44. Korrelationen zwischen SVS Werte-Typen und MVSQ Wertesystemen.

	SVS									
	TR	MA	KO	LE	BE	SB	UN	HE	ST	SI
GB^A	.13	-.25**	.03	-.39***	.06	-.49***	-.04	.00	-.41***	.07
MA^A	-.11	.22*	-.11	.11	-.18	.03	-.06	.02	.13	-.06
GW^A	.24*	-.21*	.26**	-.22*	.08	-.32***	-.20*	-.10	-.40***	.18
ER^A	-.12	.49***	.20*	.46***	-.35***	-.21*	-.40***	.01	.11	.02
GL^A	.03	-.40***	-.12	-.42***	.25**	-.19	.21*	.04	-.21*	.04
VE^A	-.02	-.20*	-.17	-.03	-.09	.34***	.00	-.13	.00	-.32***
NA^A	-.10	-.30**	-.22*	-.32***	.23*	.05	.55***	-.09	.11	-.06
GB^V	-.15	.27**	-.10	.30**	-.04	.35***	.00	.00	.21*	-.14
MA^V	.09	-.39***	-.02	-.33***	.22*	-.18	.09	-.02	-.30**	.09
GW^V	-.28**	.19*	-.25**	.13	-.04	.23*	.12	.18	.30**	-.18
ER^V	.06	-.38***	-.14	-.43***	.22*	.05	.33***	.05	-.05	-.04
GL^V	-.11	.34***	.04	.36***	-.38***	.13	-.33***	.17	.12	.02
VE^V	.00	.10	.30**	.03	.03	-.38***	-.08	.18	-.05	.32**
NA^V	.07	.18	.24*	.14	-.18	-.26**	-.50***	.12	-.19	.17

Anmerkung. Werte-Typen nach Schwartz (in Spalten): TR = Tradition, MA = Macht, KO = Konformität, LE = Leistung, BE = Benevolenz, SB = Selbstbestimmung, UN = Universalismus, HE = Hedonismus, ST = Stimulation, SI = Sicherheit; Wertesysteme nach Graves (in Zeilen): GB = Geborgenheit; MA = Macht; GW = Gewissheit; ER = Erfolg; GL = Gleichheit; VE = Verstehen; NA = Nachhaltigkeit; A = Annäherung; V = Vermeidung; *** p < .001, ** p < .01, * p < .05.

im Vergleiche mit den MVSQ-Skaleninterkorrelationen sind die beiden Validitäts-Korrelationen je mit einer Ausnahme höher (für ***Macht***^{A}) bzw. niedriger (für ***Macht***^{V}). Die Ausnahmen betreffen in beiden fällen das Wertesystem ***Gleichheit***^{V}, das mit $r = .37$ mit ***Macht***^{A} und mit $r = -.40$ mit ***Macht***^{V} korreliert. Demzufolge könnte die Ausprägung auf ***Gleichheit***^{V} mehr Aufschluss über die ***Macht***-Orientierung geben als die SVS-Macht-Ausprägung. Abgesehen davon sprechen die Befunde für die konvergente Validität des ***Macht***-Konstrukts und zwar in beiden Richtungen. Und auch die Zusammenhänge mit ***Gleichheit***^{V} sind theoretisch nachvollziehbar und sollten nicht weiter problematisch sein.

Gewissheit. Die Korrelationen mit dem „parallelen" Werte-Typ Konformität betragen $r = .26$ (Annäherung) und $r = -.25$ (Vermeidung) und werden beide signifikant. Kriterium 1 ist damit erfüllt. Die HTHM-Korrelationen mit ***Gewissheit***^{A} liegen alle unter $r = .26$, wobei die

Korrelation mit dem Werte-Typ Tradition mit $r = .24$ fast gleich hoch und auch signifikant wird. Mit ***Gewissheit***^{V} ist die Korrelation mit Tradition ($r = -.28$) sogar ausgeprägter. Berücksichtigt man die hohen Skaleninterkorrelationen von Tradition mit Konformität ($r = .59$) sind diese Korrelationen allerdings wenig überraschend. Tradition und Konformität hängen somit relativ stark miteinander zusammen und können deshalb hingenommen werden, ohne dass die konvergente Validität von ***Gewissheit*** deshalb als schlecht bezeichnet werden muss, denn für alle anderen HTHM-Korrelationen ist die zweite Bedingung erfüllt. Bei den Skaleninterkorrelationen des MVSQ zeigt sich ein ähnliches Muster. Mit Ausnahme der Korrelationen mit ***Geborgenheit*** sind alle Interkorrelationen betragsmäßig niedriger als die Korrelation der Validitäts-Diagonale. Auch ***Gewissheit*** kann somit als konvergent valide bezeichnet werden.

Erfolg. Die Korrelationen der Validitäts-Diagonale sind signifikant zum .01-Niveau und liegen bei $r = .46$ für Annäherung und $r = -.43$ für Vermeidung. Im Vergleich mit den HTHM-Korrelationen sind alle Korrelationen mit einer Ausnahme betragsmäßig höher. ***Erfolg***^{A} korreliert noch stärker mit SVS-Macht ($r = .49$) als mit SVS-Leistung. Auch auf der Vermeidungsskala korreliert ***Erfolg*** signifikant mit SVS-Macht, allerdings nicht so stark. Ähnlich wie bei ***Gewissheit*** kann diese Konfundierung in der relativ hohen Interkorrelation der Werte-Typen Macht und Leistung mit $r = .71$ liegen. Abgesehen davon wurde diese Wertesystem-Werte-Typ-Kombination bei der Analyse der Parallelität der Instrumente als problematisch eingestuft, da die Beschreibung von SVS-Macht einige Elemente des Wertesystems ***Erfolg*** enthielt. Bedingung 3 ist vollständig erfüllt, wobei anzumerken ist, dass ***Gleichheit***^{V} (wie schon mit ***Macht***) relativ hoch mit den ***Erfolg***-Wertesystemen korreliert, wenngleich niedriger als die Validitäts-Korrelation.

Gleichheit. Auch bei ***Gleichheit*** sind die Kriterien der konvergenten Validität weitestgehend erfüllt. Erstens werden die Korrelationen auf der Validitäts-Diagonalen signifikant ($r = .25$ für ***Gleichheit***^{A} und $r = -.38$ für ***Gleichheit***^{V}). Zweitens fallen alle HTHM-Korrelationen in der jeweils selben Richtung betragsmäßig niedriger aus. Es zeigt sich jedoch ein Muster, dass sich schon bei der Analyse von ***Macht*** und ***Erfolg*** angedeutet hat. Die Korrelationen von ***Gleichheit*** mit den diesen beiden Wertesystemen parallelen Werte-Typen sind betragsmäßig größer, allerdings in entgegengesetzter Richtung. Das bedeutet, dass SVS-Macht und SVS-Erfolg anders gepolt das Wertesystem ***Gleichheit*** jeweils besser beschreiben als der parallele Werte-Typ Benevolenz. Auch die Skaleninterkorrelationen des MVSQ von ***Gleichheit*** mit ***Macht*** und ***Erfolg*** (in Annäherungs- und Vermeidungsskala gleichermaßen) sind entgegengesetzter Richtung und betragsmäßig höher als die Validitäts-Korrelation. Die Befunde deuten in Summe auf eine gewisse Dualität der Wertesysteme ***Macht***/***Erfolg*** versus ***Gleichheit***. Nichtsdestotrotz sprechen die Ergebnisse für die konvergente Validität der ***Gleichheit***-Wertesysteme.

Verstehen. Die Korrelationen auf der Validitäts-Diagonalen betragen $r = .34$ für ***Verstehen***^{A} sowie $r = -.38$ für ***Verstehen***^{V} und werden beide hoch signifikant (p <.001). Abgesehen davon gibt es keine positive Korrelation mit einem anderen Werte-Typ und auch die Korrelationen mit anderen Wertesystemen liegen deutlich unter den genannten Werten. Alle drei Kriterien zur konvergenten Validität sind somit erfüllt. Erwähnenswert ist noch eine entgegengesetzte Tendenz zu den Werte-Typen Konformität und Sicherheit mit jeweils signifikanten Korrelationen (negativen für MVSQ^A und positiven für MVSQ^V).

Nachhaltigkeit. Dieses Wertesystem weist die höchsten Werte auf der Validitäts-Diagonale auf. ***Nachhaltigkeit***^{A} korreliert mit $r = .55$ sowie ***Nachhaltigkeit***^{V} mit $r = -.50$ mit Universalismus. Keine der HTHM-Korrelationen ist annähernd so hoch und auch die Skaleninterkorrelationen des MVSQ liegen jeweils deutlich unter diesen Werten.

9.2.4 Diskussion

In dieser Studie wurde die konvergente Validität des MVSQ anhand einer Multitrait-Multimethod-Analyse überprüft. Der Vorteil bei der Anwendung der MTMM-Technik liegt darin, dass auch mögliche Methodeneffekte berücksichtigt werden (Amelang & Schmidt-Atzert, 2006), denn wenn sich Zusammenhänge zwischen ähnlichen Konstrukten zeigen, obwohl diese mit unterschiedlichen Methoden erhoben wurden, spricht dies für die Validität des untersuchten Instruments. Obwohl aufgrund des Fehlens eines echten parallelen Instruments ein Fragebogen verwendet wurde, der „nur“ ähnliche Konstrukte misst, konnte die MTMM-Analyse erfolgreich angewandt werden. Die konvergente Validität der Wertesysteme wurde anhand der drei klassischen Kriterien der MTMM-Analyse (Schmitt & Stults, 1986) beurteilt und es kann zusammenfassend gesagt werden, dass diese Bedingungen bei allen Merkmalen größtenteils erfüllt sind und für die konvergente Validität des MVSQ sprechen. Abgesehen davon konnten drei systematische Muster beobachtet werden. Die Wertesysteme ***Geborgenheit*** und ***Gewissheit*** sind sich relativ ähnlich, sowohl die Skaleninterkorrelationen im MVSQ als auch die Korrelationen mit den SVS-Werte-Typen führen zu dieser Schlussfolgerung, die für Annäherung und Vermeidung gleichermaßen gilt. Eine erneute Begutachtung und gegebenenfalls Revision der Konzeptualisierungen dieser beiden Wertesysteme scheint deshalb angebracht. Das zweite Muster betrifft die Wertesysteme ***Macht***, ***Erfolg*** und ***Gleichheit***. Hierbei gilt, dass die ersteren beiden signifikant negativ mit ***Gleichheit*** zusammenhängen. Da sich dieses Muster sowohl bei den Skaleninterkorrelationen zeigt, als auch bei den Heteromethod-Korrelationen, die „parallelen“ Werte-Typen jedoch unkorreliert sind, stellt sich die Frage, ob diese Konstrukte trennscharf genug voneinander konzeptualisiert worden sind. Als dritte Tendenz wurde festgestellt, dass die Korrelationen der Werte-Typen mit den Annäherungswertesystemen spiegelbildlich, d.h. mit umgekehrten Vorzeichen, zu den Korrelationen der Werte-Typen mit den Vermeidungswertesystemen verlaufen. Dass dieses Schema für alle Heteromethod-Korrelationen gleichermaßen gilt,

kann als Indiz gewertet werden, das für Bipolarität und gegen Orthogonalität der Wertesysteme spricht.

Als Limitierung dieser Studie müssen folgende Punkte aufgeführt werden. Zunächst einmal muss bedacht werden, dass es sich beim SVS nicht um ein echt paralleles Instrument zum MVSQ handelt. Ein solches wäre für eine Untersuchung der konvergenten Validität prinzipiell besser geeignet und eine zukünftige Entwicklung anderer Methoden der Messung von Wertesystemen ist anzustreben. Nichtsdestotrotz zeigen die Ergebnisse, dass im SVS ähnliche Konstrukte gemessen werden, was ex post den Einsatz dieses Instruments rechtfertigt.

Bei der zugrunde liegenden Stichprobe handelt es sich um ein *Convenience Sample* von Regensburger Studierenden, weswegen die Generalisierbarkeit der Ergebnisse eingeschränkt ist. Daraus kann die Empfehlung abgeleitet werden, weitere derartige Untersuchungen mit anders zusammengesetzten Stichproben durchzuführen. Als initiale Konstruktvalidierung erfüllt diese Untersuchung jedoch den angestrebten Zweck.

Des Weiteren ist anzufügen, dass neben den verbesserungswürdigen Reliabilitäten des MVSQ, die bereits berichtet wurden (vgl. Kapitel 7), der SVS nur über mäßige bis akzeptable Reliabilitäten verfügte und diese Messungenauigkeiten die Ergebnisse verfälschen können. In Anbetracht der Tatsache, dass es sich dabei ohnehin nicht um eine paralleles Instrument handelt, liegt die Entwicklung eines parallelen Messverfahrens der Wertesysteme nach Graves näher als die Überarbeitung des SVS. Denkbar sind hier z.B. die Entwicklung von Richtlinien zur Fremdbeurteilung oder – obgleich als konzeptuell unpassend eingestuft (Kapitel 3.2.1) – die Messung mit einem Rating-Fragebogen.

Ungeachtet dessen, dass es sich bei den Wertesystemen und Werte-Typen um inhaltlich jeweils nicht exakt dieselben Konstrukte handelt (dafür sind die Korrelationen zu niedrig), stellt sich die Frage, ob eines der beiden Instrumente dem anderen hinsichtlich der Erfassung von Wertesystemen überlegen ist. Um das zu beurteilen, müssten beide Instrumente gleichermaßen in Untersuchungen der Kriteriumsvalidität eingesetzt werden, um dadurch zu überprüfen, ob eines der Instrumente mehr zu erklären fähig ist als das andere. Diese Fragestellung fällt jedoch nicht in den Zielbereich dieser Arbeit und kann deshalb in weiterführenden Studien untersucht werden.

9.3 Untersuchungen zur divergenten Validität

In dieser Untersuchung wird die divergente Validität der Wertesysteme durch den Vergleich mit divergierenden Konstrukten (Big Five und Intelligenz) untersucht. Divergente Validität liegt vor, wenn Merkmale, die theoretisch nicht miteinander zusammen hängen, unabhängig von der verwendeten Messmethode nicht oder nur gering miteinander korrelieren (Eid et al., 2015). In einer Voruntersuchung wurden zudem die psychometrischen Eigenschaften der verwendeten Skalen untersucht.

9.3.1 Hintergrund

Die Eigenschaftstheorie der Persönlichkeit von Cattell (1973) unterteilt Persönlichkeitsmerkmale in drei weitestgehend unabhängige Kategorien. Wertesysteme können dabei der Kategorie der motivationalen Dispositionen zugeordnet werden (vgl. Kapitel 2.5). Es liegt deshalb nahe, zur Untersuchung der divergenten Validität Konstrukte aus den beiden verbleibenden Dimensionen zu wählen. Nachdem die Big Five einerseits als Temperamentsdispositionen gesehen werden können (Scheffer & Heckhausen, 2010) und andererseits zu den etabliertesten Konstrukten in der Persönlichkeitstheorie gehören (Viswesvaran & Ones, 2000), liegt es auf der Hand, die Big Five mit den Wertesystemen zu vergleichen. Abgesehen davon wurden bereits Studien zum Zusammenhang von Werten und den Big Five durchgeführt, die herangezogen werden können, um Hypothesen abzuleiten. Eine Meta-Analyse von 60 Studien zum Zusammenhang von Werte-Typen nach Schwartz und dem Fünf-Faktoren-Modell der Persönlichkeit (Parks-Leduc et al., 2015) kommt zu dem Schluss, dass es sich bei beiden Konstruktarten um weitestgehend unabhängige Konstrukte handelt, es jedoch konsistente und theoretisch sinnvolle Zusammenhänge zwischen beiden Konstruktarten gibt, obgleich diese Zusammenhänge mit wenigen Ausnahmen eher klein ausfallen. Bei den größten Korrelationen handelt es sich demnach um mittelhohe Korrelationen (absolute Werte $< .61$) und diese liegen für die Faktoren Offenheit und Verträglichkeit vor. Die Faktoren Extraversion und Gewissenhaftigkeit zeigen geringe bis keine signifikanten Zusammenhänge mit Werte-Typen. Für Emotionale Stabilität (Neurotizismus) konnten überhaupt keine signifikanten Korrelationen festgestellt werden. Für die Wertesysteme lassen sich demzufolge folgende Annahmen auf einer allgemeineren Ebene ableiten:

- Es treten signifikante moderate Korrelationen zwischen MVSQ-Wertesystemen und Offenheit sowie Verträglichkeit auf.
- Es treten signifikante niedrige Korrelationen zwischen MVSQ-Wertesystemen und Extraversion sowie Gewissenhaftigkeit auf.
- Wertesysteme und Neurotizismus korrelieren nicht signifikant miteinander.

Unter Berücksichtigung der signifikanten Zusammenhänge zwischen den MVSQ-Wertesystemen und den SVS Werte-Typen (Kapitel 9.2, Tabelle 44) sowie der Ergebnisse von Parks-Leduc et al. (2015) lassen sich spezifischere Hypothesen ableiten:

- ***Verstehen*** und ***Nachhaltigkeit*** korrelieren positiv und ***Gewissheit*** negativ mit Offenheit.
- ***Gleichheit***, ***Nachhaltigkeit*** und ***Gewissheit*** korrelieren positiv und ***Macht*** negativ mit Verträglichkeit.
- ***Macht*** und ***Erfolg*** korrelieren positiv mit Extraversion.
- ***Gewissheit*** und ***Erfolg*** korrelieren positiv mit Gewissenhaftigkeit.

Aufgrund der bereits beobachteten gegensätzlichen Tendenz der Annäherungs- und Vermeidungssysteme (siehe z.B. Korrelationen zwischen den Skalen, Kapitel 9.1.2.1) gelten diese Hypothesen wie formuliert für die MVSQ^{A}-Wertesysteme und in entgegengesetzter Richtung für MVSQ^{V}-Wertesysteme.

Auch beim Zusammenhang zwischen Wertesystemen und Intelligenz lässt sich Cattell's Eigenschaftstheorie heranziehen. Demnach ist Intelligenz den kognitiven Dispositionen zuzuordnen (Cattell, 1973) und folglich theoretisch unabhängig von Wertesystemen. Diese Annahme hat Graves in einer eingeschränkten Form untersucht (Graves, 1971c), indem er Personen, die er persönlich in vier Gruppen (je eine Gruppe mit Präferenz eines der vier Wertesysteme ***Gewissheit***, ***Erfolg***, ***Gleichheit*** und ***Verstehen***) eingeteilt hat, die Wechsler Adult Intelligence Scale bearbeiten ließ und keine signifikanten Mittelwertsunterschiede zwischen diesen vier Gruppen fand. Aus den Aufzeichnungen von Graves (1971c) geht jedoch weder die Stichprobengröße noch die Berechnungsmethode hervor. Auch ist fraglich, wie treffend die Einteilung der Versuchspersonen in bevorzugte Wertesysteme nach seiner persönlichen Einschätzung war.

Es sei noch hinzuzufügen, dass das Konstrukt Intelligenz häufig in zwei Bereiche aufgeteilt wird: fluide Intelligenz und kristalline Intelligenz (Schmitt & Platzer, 2010). Rein logisch ist die Unabhängigkeitshypothese zwischen Werten und Intelligenz nur für fluide Intelligenz sinnvoll, demjenigen Teil der Intelligenz, der vererbt wurde und ein Leben lang gleich bleibt. Bei kristalliner Intelligenz hingegen, die im Laufe des Lebens erlernte Fähigkeiten repräsentiert, ist nur schwer vorstellbar, dass diese nicht motivationalen Einflüssen ausgesetzt wäre. Dies wird umso deutlicher, wenn man sich die Teilbereiche der kristallinen Intelligenz anschaut: Verbales, figurales und numerisches Wissen. Welches Wissen sich jemand aneignet, hängt logischerweise davon ab, was diese Person interessiert (z.B. Sprache, Kunst oder Mathematik) und ist demnach motivationalen Einflüssen unterworfen. Ob sich dieser Zusammenhang jedoch in den Korrelationen zwischen Wertesystemen und den Kennwerten der kristallinen Intelligenz zeigt, kann an dieser Stelle nicht gesagt werden. Diese Frage soll exploratorisch überprüft werden.

9.3.2 Methode

Die Zusammenhänge zwischen Wertesystemen und divergierenden Konstrukten werden wie bei Parks-Leduc et al. (2015) anhand von Produkt-Moment- und Rangkorrelationen untersucht. Da es sich beim BFI-10 um ein sehr kurze Skala handelt, wurde in einer Voruntersuchung zusätzlich zur Berechnung der internen Konsistenzen eine Faktorenanalyse durchgeführt, um zu überprüfen, ob die fünf Faktoren reproduziert werden konnten.

Für diese Studie wurden zwei unterschiedliche Stichproben herangezogen. Die Analyse der Zusammenhänge der Wertesysteme mit den Big Five beruht auf Stichprobe IIIb (N = 166), die im Rahmen einer Laborstudie erhoben wurde. Um die kognitive Belastung der Versuchspersonen gering und die Motivation, die Skala sorgfältig zu bearbeiten, hoch zu halten, wurde zur Messung der Big Five eine Kurzskala eingesetzt. Die Stichprobe zur Untersuchung des Zusammenhangs der Wertesysteme mit IQ wurde im Rahmen einer Studie in Kooperation mit der Personalabteilung eines Versicherungsunternehmens mit Sitz in München erhoben. Es handelt sich dabei um Stichprobe Ia (N = 39). Beide Stichproben sind in Kapitel 3.7 beschrieben.

Big Five Inventory

Die großen fünf Faktoren wurden mit dem 10 Item Big Five Inventory (BFI-10) gemessen, das aus zehn Items besteht und als Kurzskala entwickelt wurde, um die Persönlichkeitsdimensionen Extraversion, Verträglichkeit, Gewissenhaftigkeit, Neurotizismus und Offenheit zu erheben (Rammstedt & John, 2007). Dabei wird jedes Konstrukt von zwei Items auf einer fünfstufigen Ratingskala erfasst. Die Stufen waren: 1 „trifft überhaupt nicht zu“, 2 „trifft eher nicht zu“, 3 „weder noch“, 4 „eher zutreffend“ und 5 „trifft voll und ganz zu“ (Rammstedt et al., 2012). Die Ausprägungen wurden als Mittelwerte der beiden Items pro Persönlichkeitsdimension berechnet.

Es liegen mehrere Studien zur psychometrischen Güte des Instruments vor (Rammstedt, 2007; Rammstedt & John, 2007; Rammstedt et al., 2010, 2012), denen zufolge Objektivität, Reliabilität und Validität als gesichert angesehen werden kann. In einer Untersuchung an zwei Stichproben von Rammstedt und John (2007) wurden Test-Retest-Reliabilitäten im ausreichenden bis guten Bereich festgestellt (r_{tt} zwischen .65 und .87). Für die Konstruktvalidität des Instruments spricht, dass sowohl hohe Korrelationen der fünf Faktoren mit umfangreicheren und etablierten Big Five Inventaren nachgewiesen werden konnten, als auch die Fünf-Faktoren-Struktur in Hauptkomponentenanalysen bestätigt wurde (Rammstedt et al., 2010, 2012). Dennoch wurde auch hier in einer Voruntersuchung die Reliabilität und Faktorenstruktur des Instruments überprüft.

IST 2000 R

Zur Erhebung der Intelligenz wurde der Intelligenz-Struktur-Test (IST) 2000 R verwendet (Liepmann et al., 2007). Dieses Instrument gehört zu den am häufigsten eingesetzten Testverfahren im deutschsprachigen Raum (Hagemeister et al., 2010). Es werden darin elf Intelligenz-Kennwerte erfasst: verbale Intelligenz, figural-räumliche Intelligenz, rechnerische Intelligenz, Merkfähigkeit, schlussfolgerndes Denken, verbales Wissen, figural-bildhaftes Wissen, numerisches Wissen und Wissen (Gesamt) sowie fluide und kristalline Intelligenz. Verbale, numerische und figural-räumliche Intelligenz bilden dabei als Summenwerte den Kennwert des Schlussfolgernden Denkens. Verbales, numerisches und figurales Wissen ergeben zusammen den Kennwert Wissen (Gesamt). Die beiden Generalfaktoren fluide und kristalline Intelligenz werden als gewichtete Summenscores der von Schlussfolgerndem Denken bzw. Wissen (Gesamt) berechnet (Liepmann et al., 2007). Das Instrument weist laut Manual (Liepmann et al., 2007) sehr hohe interne Konsistenzen (zwischen $\alpha = .87$ und $\alpha = .97$) und gute bis sehr gute Split-Half-Reliabilitäten (von $r = .88$ bis $r = .96$) auf. In der vorliegenden Stichprobe konnten diese Werte nicht überprüft werden, da die Itemwerte nicht zur Verfügung standen. Ferner deuten auch die Untersuchungen zur Validität, in denen das Instrument mit verschiedenen Tests kontrastiert wurde, auf eine hohe Gültigkeit hin (Liepmann et al., 2007).

9.3.3 Ergebnisse Voruntersuchungen

Zunächst wurde die Frage untersucht, ob sich fünf Faktoren extrahieren lassen, die die Big Five Faktoren repräsentieren. Zwar führten die Entwickler des Instruments zur Bestätigung der Faktorenstruktur jeweils eine Hauptkomponentenanalyse durch (Rammstedt et al., 2010, 2012), da es sich bei den Big Five um latente Konstrukte handelt, ist eine Faktorenanalyse – deren Ziel darin besteht, latente Faktoren zu extrahieren – verfahrenstechnisch sinnvoller (Costello & Osborne, 2005).

Die Eignung der Daten zur Durchführung einer Faktorenanalyse waren nach den Voraussetzungen von Bühner (2011, S. 347) mäßig bis schlecht, da der Kaiser-Meyer-Olkin-Koeffizient bei KMO $= .56$ lag und die Measure of Sample (MSA)-Koeffizienten der Items mit Werten zwischen $.45$ und $.66$ sehr niedrig ausfielen. Zwei der MSAs lagen unter dem empfohlenen Minimalwert von .50 (Bühner, 2011). Aufgrund der geringen Anzahl an Items wurden dennoch alle Items beibehalten. Des Weiteren war der Bartlett-Test signifikant ($p < .001$), was für die Existenz substanzieller Korrelationen zwischen den Items sprach. Zwar sind zwischen den Big Five Faktoren in der Regel (geringe) Skaleninterkorrelationen zu erwarten (Vgl. z.B. Goldberg, 1992; Lang et al., 2001; Musek, 2007), was eher für eine oblique Rotation sprechen würde. Diese produzierten allerdings ausnahmslos Heywood-Fälle (Heywood, 1931), weswegen eine orthogonale Rotation (Varimax) angewandt wurde. Obwohl logischerweise fünf Faktoren extrahiert werden sollen, wurde das Kaiser-Kriterium und eine Parallel-Analyse nach Horn

(1965) durchgeführt. Beide Methoden bestätigten die Anzahl der zu extrahierenden Faktoren von fünf.

Die Faktorenanalyse bestätigt die Fünf-Faktoren-Lösung. Der χ^2-Test wurde nicht signifikant, $\chi^2(5) = 4, p = .62$, und bestätigt somit die Beibehaltung des Modells mit fünf Faktoren. Die Faktorladungen (Tabelle 45) sind für vier der fünf Faktoren wie erwartet. Lediglich bei Verträglichkeit lädt nur ein Item („kritisieren“) einigermaßen adäquat auf den Faktor. Zudem gibt es eine problematische Querladung des Items „Gründlichkeit“ auf diesen Faktor. Ferner ist kritisch zu sehen, dass die fünf Faktoren nur 57.5% der Varianz erklären. Für die folgende Analyse ist der Faktor Verträglichkeit folglich nur eingeschränkt geeignet, um als divergentes Konstrukt herangezogen zu werden.

Tabelle 45. Ladungen des Faktorenmodells mit fünf Faktoren nach Varimax-Rotation.

	N	E	O	V	G
entspannt	-.52			-.29	.21
unsicher	.92				
reserviert		-.63			
gesellig		.98			
kuenstlerisch		-.29	-.44		
fantasievoll			.99		
vertrauen				-.26	
kritisieren				.47	
Faulheit					.99
gruendlich				.39	-.40

Anmerkung. N = Neurotizismus, E = Extraversion, O = Offenheit, V = Verträglichkeit, G = Gewissenhaftigkeit. Stichworte in der ersten Spalte indizieren die Items des BFI-10; Faktorladungen < |.2| wurden weggelassen.

Schließlich wurden die Cronbach's α des BFI-10 berechnet. Diese liegen im Mittel bei $.54$, wobei die Reliabilität der Verträglichkeitsdimension mit $\alpha_V = .17$ den Mittelwert nach unten zieht. Dieser niedrige Wert steht in Übereinstimmung mit den zuvor ermittelten schlechten Ladungen des Verträglichkeitsfaktors. Die Reliabilitäten der verbleibenden Dimensionen verteilen sich wie folgt: $\alpha_N = .60$, $\alpha_E = .77$, $\alpha_O = .64$, $\alpha_G = .51$. Für die Untersuchung der divergenten Validität bedeuteten diese Reliabilitäten, dass insbesondere die Zusammenhänge der Verträglichkeitsdimension mit Wertesystemen nicht interpretiert werden sollten. Auch die Reliabilitäten von Gewissenhaftigkeit, Neurotizismus und Offenheit lagen unter der Grenze von .70. Für den vorliegenden Fall sind die Reliabilitäten von Neurotizismus und Offenheit

jedoch akzeptabel, da auch eine ungenaue Messung valide sein kann (Schermelleh-Engel & Werner, 2012).

9.3.4 Ergebnisse Hauptuntersuchungen

9.3.4.1 Divergenz zu Big Five

Tabelle 46. Korrelationen der MVSQ-Wertesysteme mit den Big Five.

	Big Five									
	N		E		O		V		G	
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
GB^A	.25**	.23**	-.29***	-.28***	-.13	-.11	.15	.15	-.14	-.13
MA^A	.02	-.02	.07	.06	-.02	-.03	-.11	-.08	-.13	-.12
GW^A	.39***	.37***	-.33***	-.33***	-.18*	-.18*	.09	.10	-.04	-.05
ER^A	.03	.00	-.17*	-.19*	-.28***	-.24**	-.12	-.07	.02	-.01
GL^A	.07	.10	.08	.11	.14	.14	.35***	.31***	-.11	-.11
VE^A	-.15	-.11	-.03	-.05	.27***	.27***	-.07	-.04	.09	.09
NA^A	.02	.04	.12	.10	.22**	.21**	.13	.11	.00	.03
GB^V	-.18*	-.19*	.15	.15	.05	.04	-.19*	-.20**	.10	.07
MA^V	.05	.11	-.12	-.13	.06	.05	.13	.08	.06	.06
GW^V	-.31***	-.30***	.31***	.31***	.13	.13	.01	.00	.02	.04
ER^V	.07	.09	.07	.07	.30***	.29***	.13	.11	-.01	-.01
GL^V	-.06	-.10	.01	.01	-.15	-.12	-.33***	-.29***	.00	.00
VE^V	.11	.10	.12	.14	-.15	-.13	.20**	.20*	-.01	-.05
NA^V	.01	.04	-.16*	-.15	-.26***	-.23**	-.04	-.02	-.08	-.09

Anmerkung. N= Neurotizismus, E = Extraversion, O = Offenheit, V = Verträglichkeit, G = Gewissenhaftigkeit, MVSQ-B = Bipolare Skala, MVSQ-A = Annäherungsskala, MVSQ-V = Vermeidungsskala; r = Pearson's Produkt-Moment Korrelation, ρ = Spearman's Rangkorrelation, * $p < .05$; ** $p < .01$; *** $p < .001$.

Tabelle 46 zeigt die Korrelationen zwischen den MVSQ-Wertesystemen und den Big Five Scores. Als erstes fällt auf, dass sich die Produkt-Moment- und Rangkorrelationen zwar leicht in ihren Werten, aber mit zwei Ausnahmen nicht in ihren Signifikanzen unterscheiden. Des Weiteren kann festgestellt werden, dass die Korrelationen der Annäherungs-Wertesysteme mit den Big Five einem ähnlichen (reversen) aber nicht identischen Muster folgten, sowohl was

die Ausprägungen angeht, wie auch die Signifikanzen. Von den Vermeidungswertesystemen korrelieren einige weniger mit den Big Five.

Die allgemeinen Hypothesen, dass moderate und signifikante Korrelationen zwischen Offenheit mit den Wertesystemen auftreten, kann bestätigt werden. ***Gewissheit***^{A} und ***Erfolg***^{A} korrelieren (mit beiden Korrelationsmethoden) moderat negativ mit Offenheit (r und $\rho = -.18$ bzw. $r = -.28$, $\rho = -.24$) und ***Verstehen***^{A} und ***Nachhaltigkeit***^{A} korrelieren moderat positiv mit Offenheit (r und $\rho = .27$ bzw. $r = .22$, $\rho = .21$). Mit Ausnahme des Zusammenhangs von Offenheit mit ***Erfolg*** wurden diese Korrelationen in dieser Richtung hypothetisiert. Auf der Vermeidungsskala korrelieren ***Gewissheit***^{V} und ***Verstehen***^{V} nicht signifikant mit Offenheit. ***Erfolg***^{V} ($r = .30$, $\rho = .29$) und ***Nachhaltigkeit***^{V} ($r = -.26$, $\rho = -.23$) zeigen Korrelationen in derselben Größenordnung und in entgegengesetzter Richtung. Somit kann für die Vermeidungsdimension nur die Hypothese zum ***Nachhaltigkeit***-Wertesystem bestätigt werden.

Obgleich die Verträglichkeitsdimension eine nicht akzeptable interne Konsistenz aufwies und den Faktor nicht gut repräsentiert, soll an dieser Stelle erwähnt werden, dass ***Gleichheit*** hypothesenkonsistent mit dieser Dimension korreliert. Keine signifikanten Korrelationen liegen hingegen für die Wertesysteme ***Nachhaltigkeit*** und ***Gewissheit*** vor.

Bzgl. des Faktors Extraversion können ebenso zwei signifikante Zusammenhänge mit Wertesystemen festgestellt werden. Die allgemeine Hypothese kann damit bestätigt werden, dass leichte Korrelationen zwischen diesem Faktor und Wertesystemen bestehen. Im Spezifischen allerdings tritt keine Signifikanz der hypothetisierten Korrelationen auf. Stattdessen werden jedoch die Korrelationen der Wertesysteme ***Geborgenheit***^{A} und ***Gewissheit***^{A} negativ signifikant ($r = -.29$, $\rho = -.28$ bzw. $r = -.33$, $\rho = -.33$) und ***Gewissheit***^{V} positiv signifikant (r und $\rho = .31$). Diese Zusammenhänge waren so nicht hergeleitet.

Nicht konsistent mit der aus anderen Untersuchungen abgeleiteten Hypothese, dass es einen Zusammenhang zwischen Gewissenhaftigkeit und Wertesystemen (***Gewissheit*** und ***Erfolg***) gibt, stellen sich die Befunde hier dar. Kein Wertesystem korreliert signifikant mit Gewissenhaftigkeit.

Ebenso nicht Hypothesenkonsistent liegt der Fall bei den Neurotizismus-Korrelationen. Die Wertesysteme ***Geborgenheit*** ($r = .25$, $\rho = .23$ für Annäherung; $r = -.18$, $\rho = -.19$ für Vermeidung) und ***Gewissheit*** ($r = .39$, $\rho = .37$ für Annäherung; $r = -.31$, $\rho = -.30$ für Vermeidung) hängen signifikant mit Neurotizismus zusammen.

Insgesamt lässt sich sagen, dass die Mehrzahl der 140 untersuchten Korrelationen nicht signifikant wurden (51 der 70 Produkt-Moment- sowie 52 der 70 Rangkorrelationskoeffizienten) und diejenigen Korrelationen, die signifikant wurden, als niedrig eingestuft werden können (maximales $r \leq |.39|$, maximales $\rho \leq |.37|$). In der Gesamtschau spricht das Ergebnis somit dafür, dass die MVSQ-Wertesysteme weitestgehend von den Big Five Faktoren unabhängig sind. Es kann als Befund zu Gunsten der divergenten Validität gewertet werden.

9.3.4.2 Divergenz zu IST 2000 R

Bei kleinen Stichproben (hier N = 39) ist die Berechnung von Korrelationen nach Pearson nur dann empfohlen, wenn die involvierten Variablen multivariat normalverteilt sind (Bonett & Wright, 2000). In der vorliegenden Stichprobe sind allerdings 22 der 140 Kombinationen von Wertesystemen und IST-Dimensionen nicht multivariat normalverteilt. Dies wurde mit Mardia's Test auf multivariate Normalverteilung (Mardia, 1970) mithilfe des R-Pakets MVN (Korkmaz et al., 2014) berechnet. Wenn keine multivariate Normalverteilung vorliegt, empfehlen Bonett und Wright (2000) Spearman's ρ oder Kendall's τ als Alternative. In der folgenden Analyse wurden deshalb Rangkorrelationen nach Spearman herangezogen, um die Zusammenhänge zwischen Wertesystemen und IST-Werten zu berechnen. Die Ergebnisse sind in Tabelle 47 aufgeführt.

Tabelle 47. Rangkorrelationen der Wertesysteme mit IST 2000 R Intelligenz-Dimensionen.

	Verb. I.	Num. I.	Fig. I.	SD	Fl. I.	Verb. W.	Num. W.	Fig. W.	WG	Kr. I.	Merkf.
GB^A	-.37*	-.23	-.20	-.37*	-.29	-.35*	-.36*	-.43**	-.40*	-.38*	-.15
MA^A	-.17	.02	.08	-.04	.03	-.12	-.12	-.14	-.14	-.14	-.17
GW^A	-.53***	-.28	-.25	-.53***	-.43**	-.36*	-.51***	-.49**	-.48**	-.45**	-.27
ER^A	.05	.03	-.14	-.01	-.11	.00	-.07	.10	.01	.00	.03
GL^A	-.46**	-.24	-.13	-.43**	-.27	-.23	-.44**	-.49**	-.45**	-.41**	-.08
VE^A	.02	.00	-.03	-.06	-.15	-.15	-.16	-.03	-.13	-.14	.22
NA^A	.17	-.17	-.10	-.05	-.08	-.27	-.28	-.42**	-.41**	-.43**	-.03
GB^V	.57***	.19	.20	.39*	.30	.23	.33*	.30	.30	.28	.27
MA^V	.13	.12	.05	.06	.01	-.10	.02	-.01	-.02	-.02	.23
GW^V	.43**	.31	.16	.41**	.31	.10	.28	.28	.23	.21	.38*
ER^V	-.20	-.14	-.20	-.26	-.26	-.03	-.13	-.37*	-.19	-.15	-.22
GL^V	.26	.13	.06	.28	.18	.26	.21	.36*	.31	.27	-.06
VE^V	-.39*	-.16	-.18	-.31	-.34*	-.26	-.44**	-.41*	-.42**	-.37*	-.23
NA^V	-.34*	-.19	-.26	-.29	-.23	-.17	-.20	-.15	-.17	-.13	-.25

Anmerkung. Verb. I.= Verbale Intelligenz; Num. I. = Numerische Intelligenz; Fig. I. = Figurale Intelligenz; SD = Schlussfolgerndes Denken; Verb. W. = Verbales Wissen; Num. W. = Numerisches Wissen; Fig. W. = Figurales Wissen; WG = Wissen (Gesamt); Merkf. = Merkfähigkeit; Fl. I. = Fluide Intelligenz; Kr. I. = Kristalline Intelligenz; GB = Geborgenheit; MA = Macht; GW = Gewissheit; ER = Erfolg; GL = Gleichheit; VE = Verstehen; NA = Nachhaltigkeit; A = Annäherung; V = Vermeidung; * $p < .05$; ** $p < .01$; *** $p < .001$.

Im Folgenden werden angeborene und erworbene Intelligenz-Kennwerte getrennt betrachtet, da sich die Hypothesen zum Zusammenhang mit Wertesystemen unterscheiden. Kein Zusammenhang wurde zwischen Wertesystemen und den angeborenen (fluiden) Intelligenz-

Werten vermutet. Diese Hypothese kann weitestgehend bestätigt werden. Bei den drei Einzelwerten verbale, numerische und figurale Intelligenz wurden von den je 21 Korrelationen pro MVSQ-Skala drei mit Annäherungswertesystemen und vier mit Vermeidugswertesystemen signifikant. Auf beiden Skalen wurden die Korrelationen zwischen verbaler Intelligenz und ***Geborgenheit*** ($\rho = -.37$ bzw. $\rho = .57$) sowie ***Gewissheit*** ($\rho = -.53$ bzw. $\rho = .43$) signifikant. Abgesehen davon korrelierten ***Gleichheit***^{A} ($\rho = -.46$) negativ sowie ***Verstehen***^{V} ($\rho = .39$) und ***Nachhaltigkeit***^{V} ($\rho = -.34$) positiv mit verbaler Intelligenz. Numerische und figurale Intelligenz zeigen keinen signifikanten Zusammenhang mit den Wertesystemen. Beim Summenscore dieser drei Intelligenz-Werte, dem Schlussfolgernden Denken (SD) werden logischerweise einige der Korrelationen mit denselben Wertesystemen signifikant: ***Geborgenheit***^{A}, ***Gewissheit***^{A} sowie ***Gleichheit***^{A} korrelieren negativ ($\rho = -.37$, $\rho = -.53$ und $\rho = -.43$) und ***Geborgenheit***^{V} sowie ***Gewissheit***^{V} positiv ($\rho = .39$ und $\rho = .41$) mit SD. Fluide Intelligenz als gewichteter Score zeigt einen negativen Zusammenhang mit ***Gewissheit***^{A} ($\rho = -.43$) und ***Verstehen***^{V} ($\rho = .34$). Insgesamt kann somit die Hypothese der Unabhängigkeit für figurale und numerische Intelligenz vollständig und für verbale Intelligenz für sieben der 14 Wertesysteme bestätigt werden. Ähnlich liegen die Zusammenhänge der Wertesysteme mit Merkfähigkeit, wo mit einer Ausnahme (moderat positive signifikante Korrelation von $\rho = .38$ mit ***Gewissheit***^{V}) Unabhängigkeit vorliegt.

Mehrere signifikante Zusammenhänge gab es – passend zur eingangs formulierten Hypothese – zwischen Wertesystemen und den erlernten (kristallinen) Intelligenz-Kennwerten. Hier wurden von den Annäherungswertesystemen neun der 21 Korrelationen mit den drei Sub-Kennwerten verbales, numerisches und figurales Wissen signifikant. Die Vermeidungswertesysteme zeigten hier weniger Zusammenhang, von ihnen wurden nur fünf der 21 Korrelationen signifikant. Bei den übergeordneten Faktoren Wissen (Gesamt) und kristalline Intelligenz wurden von den MVSQ^A-Wertesystemen jeweils vier Korrelationen signifikant und von den MVSQ^V-Wertesystemen lediglich die Korrelationen mit ***Verstehen***^{V}. Auffallend ist hier, dass die *express-self*-Wertesysteme der Annäherungsdimension allesamt negativ mit den fluiden Intelligenz-Werten zusammenhingen. Darüber hinaus kann festgestellt werden, dass alle Korrelationen im moderaten bis niedrigen Bereich lagen (maximales $\rho \leq |.57|$).

Zusammenfassend kann – unter Vorbehalt, aufgrund des geringen Stichprobenumfangs – gesagt werden, dass volle Unabhängigkeit für vier Wertesysteme (***Macht***^{A}, ***Erfolg***^{A}, ***Verstehen***^{A} und ***Macht***^{V}) festgestellt werden konnte. Für die Wertesysteme ***Nachhaltigkeit***^{A}, ***Erfolg***^{V} und ***Gleichheit***^{V} gilt Unabhängigkeit für fluide Intelligenz. ***Geborgenheit***^{A}, ***Gewissheit***^{A}, ***Gleichheit***^{A}, ***Geborgenheit***^{V}, ***Gewissheit***^{V}, ***Verstehen***^{V} und ***Nachhaltigkeit***^{V} hängen von den drei Sub-Werten der fluiden Intelligenz nur mit verbaler Intelligenz signifikant zusammen.

9.3.5 Diskussion

In dieser Untersuchung wurde der Zusammenhang der Wertesysteme mit den Big Five sowie Intelligenz untersucht. Dazu wurden in zwei Stichproben zusätzlich zum MVSQ eine Big Five- und eine Intelligenz-Skala administriert. Die Ergebnisse sind weitestgehend hypothesenkonsistent, d.h. sie stützen die Konzeption von Wertesystemen als relativ unabhängige Konstrukte verglichen mit den Big Five und figuraler Intelligenz. Insgesamt sprechen die Befunde somit für die divergente Validität der im MVSQ gemessenen Wertesysteme. Im Vergleich zu den Befunden einer Meta-Analyse über den Zusammenhang zwischen Schwarz Werte-Typen und den Big Five (Parks-Leduc et al., 2015) zeigten die hier resultierenden Ergebnisse sogar schwächere Zusammenhänge.

Abgesehen von der divergenten Validität stehen diese Ergebnisse auch in Übereinstimmung mit der Konzeptualisierung von Wertesystemen als motivationale Dispositionen (vgl. Kapitel 2.5), die nach Cattell (1973) als unabhängig von kognitiven und Temperamentsdispositionen wie den Big Five und Intelligenz gelten.

Als Grenzen dieser Untersuchung sind folgende Punkte zu erwähnen: Bei der Untersuchung zum Zusammenhang mit den Big Five muss aufgeführt werden, dass die interne Konsistenz des Faktors Verträglichkeit extrem niedrig war und der Faktor in der Faktorenanalyse ungenügend von den beiden Items repräsentiert wurde. Als Folge müssen die (zwar hypothesenkonsistenten) Korrelationen von Verträglichkeit mit den Wertesystemen kritisch gesehen werden. Eine Untersuchung mit einer umfangreicheren Big Five-Skala erscheint an dieser Stelle nötig.

Außerdem muss hervorgehoben werden, dass die Stichprobe zur Untersuchung des Zusammenhangs mit Intelligenz erstens sehr klein war und zweitens aus dual Studierenden und Auszubildenden bestand, deren IQ-Werte ein Kriterium zur Einstellung darstellten. Die Stichprobe war demnach vorselektiert und kann deshalb nicht als bevölkerungsrepräsentativ gelten. Eine weitere Untersuchung mit einer größeren und repräsentativeren Stichprobe ist für eine Folgeuntersuchung anzustreben.

Abgesehen davon können die Ergebnisse als vielversprechend eingestuft werden und die These, dass Wertesysteme als motivationale Konstrukte gesehen werden, sollte durch weitere Vergleiche mit divergierenden (und nicht-motivationalen) Konstrukten wie z.B. Impulsivität oder Ungerechtigkeitssensibilität untersucht werden.

Kapitel 10

Kriteriumsvalidität

Die Kriteriumsvalidität eines Instruments kann wie vorgestellt (Kapitel 3.6.3) anhand der konkurrenten, prädiktiven und inkrementellen Validität untersucht werden. In den folgenden drei Kapiteln werden diese Aspekte der Kriteriumsvalidität behandelt.

10.1 Untersuchung zur konkurrenten Validität

In dieser Analyse wird die konkurrente Validität der MVSQ-Scores untersucht. Konkurrente Validität bezeichnet diejenige Kriteriumsvalidität, in der das Merkmal und das Außenkriterium zeitgleich gemessen werden (Groth-Marnat, 2003). Dabei sollte eine theoretisch begründbare Beziehung zwischen Merkmal und Außenkriterium vorliegen (Groth-Marnat, 2003). Beide Anforderungen sind hier erfüllt. Die Außenkriterien (Studiengänge, Studiengangsschwerpunkte, Aufgabenbereich und Hierarchieebene) werden im soziodemografischen Teil und somit zeitgleich zu den Wertesystemausprägungen erhoben und anhand einer Expertenbefragung wurden die hypothetisierten Beziehungen zwischen Außenkriterien und Wertesystemen hergestellt.

Die Zusammenhänge zwischen Wertesystemen und Studiengängen bzw. Studiengangsschwerpunkten sind insofern begründbar, als Wertesysteme als übergeordnete Konstrukte Ziele und Interessen beeinflussen können und somit konkrete Auswirkungen auf die Wahl des Studiengangs sowie des Studiengangsschwerpunkts haben können. Hat eine Person z.B. eine hohe Ausprägung auf dem Wertesystem ***Erfolg***, dann ist wahrscheinlich das Ziel, nach dem Studium einen gut bezahlten Beruf zu ergreifen wichtiger, als z.B. einen sozialen Beruf. Da Berufseinsteiger der Betriebswirtschaft (BW) bekanntlich höhere Einstiegsgehälter haben als Absolventen der sozialen Arbeit, ist unter BW-Absolventen eine überdurchschnittliche Ausprägung des Wertesystems ***Erfolg*** zu erwarten. Da allerdings nicht bei allen Studiengangs- und Wertesystem-Kombinationen so eingängige Zusammenhänge erkennbar sind, wurden zwei Experten danach befragt, welche Wertesysteme für welchen Studiengang und welchen Schwerpunkt charakteristisch sein könnten.

Die Hypothese, dass es einen Zusammenhang zwischen Werten und charakteristischen Merkmalen von Jobs oder Organisationen gibt, ist nicht neu. In einer Untersuchung über den Zusammenhang der Werte von Berufseinsteigern und den entsprechenden Organisationswerten, fand Chatman (1989) heraus, dass diejenigen Berufseinsteiger, deren Werte zu denen der Organisation passten, nicht nur zufriedener mit ihrem Job waren, sondern auch länger bei der Firma blieben. Ähnlich lauten die Ergebnisse einer Studie von Sagiv und Schwartz (2000): Die Kongruenz zwischen persönlichen Werten und denen der Umgebung fungieren als Moderator für subjektives Wohlbefinden. Geht man davon aus, dass subjektives Wohlbefinden eines der höchsten Güter des menschlichen Dasein ist (Diener, 1984), dann sollten sich Menschen bevorzugt und auch dauerhaft länger in Umgebungen (Berufen) befinden, die ihren eigenen Werten entsprechen.

Demzufolge ist einerseits zu erwarten, dass es dominante Wertesysteme je nach Aufgabenbereich gibt und die Existenz von dominanten Wertesystemen in einem bestimmten Beruf als Indiz der konkurrenten Validität interpretiert werden kann. Dies gilt zumindest dann, wenn man davon ausgeht, dass ähnliche Berufe über Organisationen hinweg stabile Merkmale teilen. Bei der Befragung der Experten zur Kongruenz von Wertesystemen und Studiengängen wurden auch die erwarteten Zusammenhänge von Wertesystemen mit Aufgabenbereichen und Hierarchieebenen abgefragt.

10.1.1 Methode

Die Hypothesen zu den charakteristischen Wertesystemen der Studiengänge, Schwerpunkte, Aufgabenbereiche und Hierarchieebenen wurden auf Basis einer Expertenbefragung formuliert. Zudem wurden Hypothesen zum Zusammenhang zwischen Wertesystemen und Alter bzw. Geschlecht aus vergleichbaren Untersuchungen abgeleitet. Alle Hypothesen wurden mit Daten analysiert, die ausnahmslos im allgemeinen Teil des MVSQ erhoben wurden. Die Zusammenhänge zwischen Alter und Wertesystemen wurden mit Produkt-Moment-Korrelationen untersucht. Für die Analyse der Zusammenhänge zwischen Wertesystemen und Geschlecht wurden t-Tests für unabhängige Stichproben und Effektstärken (Cohen's d mit gepoolter Varianz; Cohen, 1988) berechnet (R-Paket `lsr`; Navarro, 2015). Zur Untersuchung der Kongruenz von Wertesystemen mit Studiengängen, Aufgabenbereichen und Hierarchieebenen wurden einfaktorielle ANOVAS mit Post-hoc Tests gerechnet.

Zur Formulierung der Hypothesen wurden zwei Experten eine Liste der Abteilungen, Studiengängen und Hierarchieebenen vorgelegt. Sie hatten die freie Wahl, wie viele Annäherungs- und Vermeidungswertesysteme sie jeweils als charakteristisch pro Kategorie zuteilen. Für die Berechnung der Interrater-Reliabilität wurden Elemente ausgeschlossen, denen nur von einem Rater ein charakteristisches Wertesystem zugewiesen wurde. Die Interrater-Reliabilität lag für 40 Wertesystem-Zuordnungen, die von beiden Ratern vorlagen, bei $\kappa = .69$. Nach der

Beurteilungsrichtlinie zur Urteilerübereinstimmung von LeBreton und Senter (2007) liegt dieses κ an der oberen Grenze von „moderater Übereinstimmung“ und rechtfertigt somit die folgende Ableitung der Hypothesen.

10.1.1.1 Stichprobe

Um ein ausreichend großes N für Gruppentestungen (≥ 20) in den einzelnen Gruppen vor allem von Studiengängen und Aufgabenbereichen zu erreichen, wurden für diese Untersuchungen die Stichproben II und III, die beide auf der überarbeiteten Version des Instruments beruhten, zusammengelegt und anhand des Kriteriums studierend / berufstätig erneut aufgeteilt. Für die Beschreibung der einzelnen Stichproben sei auf Kapitel 3.7 verwiesen. Die wichtigsten Kennziffern der Aufteilung nach Studierenden und Berufstätigen folgt im nächsten Abschnitt, außerdem werden die Kennziffern der einzelnen Teilstichproben präsentiert. Die kombinierte Stichprobe wird herangezogen, um Zusammenhänge von Wertesystemen mit Geschlecht und Alter zu untersuchen.

Kumulierte Stichprobe der Studierenden

Insgesamt umfasste diese Stichprobe N = 449 Studierende, von denen zwei Drittel ($n = 299$) weiblich und dementsprechend ein Drittel männlich war ($n = 150$). Das Durchschnittsalter betrug M = 23.4 ($SD = 3.3$, Spanne von 16 bis 56 Jahre). Fast alle (98%) waren Deutsch und hatten Deutsch zur Muttersprache (96.4%).

Tabelle 48. Kennwerte der Studierendenteilstichproben.

		Frauen		Alter		Berufserfahrung	
Studiengang	N	n	%	M	SD	M	SD
Betriebswirtschaft	249	165	66.3	23.6	2.6	2.2	2.0
Informatik	25	14	56.0	24.3	3.9	1.5	2.5
IRM	21	17	81.0	22.4	2.5	1.2	1.9
Sonstige Studiengänge	154	103	66.9	22.9	4.1	1.2	1.8
Schwerpunkt des Studiengangs Betriebswirtschaft							
Personalmanagement	57	46	80.7	23.8	2.5	2.0	1.7
Projektmanagement	40	27	67.5	23.8	3.0	2.4	2.4
Finanzen	22	11	50.0	23.8	1.9	2.0	1.7
Sonstige Schwerpunkte	130	81	62.3	23.5	2.7	2.2	2.0

Die Berufserfahrung belief sich im Schnitt auf 1.8 Jahre ($SD = 2$). Für die Untersuchungen zu den Unterschieden der Wertesysteme zwischen Studierenden bestimmter Studiengänge wurden Studiengänge mit mehr als 20 Studierenden berücksichtigt. Außerdem gab es unter der Gruppe der BW-Studierenden mehrere Schwerpunkt-Fächer, die mehr als 20 Studierende gewählt hatten. Auch diese wurden als gesonderte Gruppen analysiert. Die Kennzahlen der Studierendengruppen zeigt Tabelle 48.

Kumulierte Stichprobe der Berufstätigen

Von den insgesamt N $= 867$ Berufstätigen waren mit $n = 313$ Personen (36.1%) etwas mehr als ein Drittel weiblich. Dementsprechend waren $n = 554$ Personen männlich (63.9%). Das Alter betrug im Mittel 36.9 Jahre ($SD = 10$) bei einer Spanne von 17 bis 67 Jahre. 799 Personen (92.2%) spezifizierten Deutsch als Nationalität und 787 Personen (90.8%) gaben Deutsch als Muttersprache an.

Die durchschnittliche Berufserfahrung belief sich auf 12.7 Jahre ($SD = 9.8$) und die Zugehörigkeiten der Berufstätigen zu den Hierarchieebenen gestalteten sich wie folgt: 33 Angestellte gehörten der Geschäftsführung (3.8%) an, 101 Personen arbeiteten eine Ebene unter der Geschäftsführung (5.2%), 168 zwei Ebenen (19.4%) und 484 arbeiteten drei und mehr Ebenen unter der Geschäftsführung (55.8%). 58 waren Selbstständige (18.3%) und von 23 Personen liegt keine Angabe vor (7.3%).

Für die Untersuchungen zur den Unterschieden der Wertesysteme zwischen Mitarbeitern derselben ähnlichen Aufgabenbereichen wurden nur solche Bereiche berücksichtigt, deren Gruppengröße bei $n \geq 20$ lag. Die Kennzahlen der entsprechenden Gruppenstichproben sind in Tabelle 49 aufgelistet.

10.1.1.2 Hypothesen

Im Folgenden werden die Hypothesen für die erwarteten Zusammenhänge zwischen Wertesystemen und Alter und Geschlecht, Studiengänge und BW-Schwerpunkte sowie Aufgabenbereich und Hierarchieebene dargestellt.

Alter und Geschlecht

Zum Zusammenhang von Wertesystemen mit beiden Variablen ist zunächst anzumerken, dass dazu keine Befunde von Graves vorliegen. Deshalb wurden zur Hypothesenformulierung Untersuchungen herangezogen, die auf anderen Wertetheorien beruhen. In mehreren Studien (Cherrington et al., 1979; Feather, 1977; Schwartz, 2003) zum Zusammenhang von Werten und Alter wurden übereinstimmend überwiegend niedrige Korrelationen ($\leq .30$) berichtet. Niedrige Korrelationen sind insofern plausibel, da Werte konzeptuell als zeitlich relativ stabil gelten (Rokeach, 1973), was Jin und Rounds (2012) auch empirisch in einer Langzeitstudie

Tabelle 49. Kennwerte der Teilstichproben aufgeschlüsselt nach Job bzw. Hierarchieebene.

		Frauen		Alter		Berufserfahrung	
Job	N	n	%	M	SD	M	SD
Vertrieb	140	41	29.3	38.9	9.8	15.1	10.0
Forschung & Entw.	122	23	18.9	35.9	9.0	10.5	8.7
Personal	70	51	72.9	33.3	11.6	9.2	9.9
IT	58	11	19.0	36.9	6.9	12.9	7.4
Produktion	47	6	12.8	38.2	8.9	15.2	9.6
Unternehmensführung	37	6	16.2	41.8	9.8	16.8	8.6
Logistik	33	7	21.2	39.1	9.2	14.8	9.5
Organisation & Verw.	32	19	59.4	34.6	9.7	11.0	9.7
Kundendienst	30	10	33.3	33.7	8.3	11.5	8.7
Marketing	28	17	60.7	33.4	9.8	8.0	8.5
Abteilungsleitung	28	8	28.6	42.8	9.6	18.6	8.9
Finanzen	25	14	56.0	36.6	10.0	15.2	12.4
Sonstige Abteilungen	217	100	46.1	36.4	10.7	11.9	10.2
Hierarchieebene							
Geschäftsführung	33	4	12.1	44.5	9.5	18.9	8.7
GF -1	101	35	34.7	39.1	10.1	14.8	10.0
GF -2	168	55	32.7	38.4	9.1	13.8	9.8
GF ≤ -3	484	188	38.8	35.1	9.5	11.3	9.4
Selbstständige	58	23	39.7	39.9	12.6	14.1	11.2

Anmerkung. Von 23 Personen fehlt die Angabe der Hierarchieebene. Diese wurden von der Analyse ausgeschlossen.

überprüft haben. Dementsprechend lautet die Kernhypothese für den Zusammenhang zwischen Wertesystemen und Alter:

Hypothese H1: Wertesysteme und Alter korrelieren maximal niedrig ($\leq .30$) miteinander.

In einer Meta-Analyse untersuchten Schwartz und Rubel (2005) mehr als 100 Stichproben aus 70 Ländern auf Geschlechterunterschiede. Sie fanden zwar signifikante Geschlechterunterschiede, die jedoch tendenziell relativ geringe Effektstärken aufwiesen (Median $d = .15$,

maximales $d < .32$). Männer bevorzugten durchweg Werte wie Macht, Leistung und Hedonismus, Frauen hingegen Werte wie Gutmütigkeit und universalistische Werte. Die Hypothesen für Geschlechterunterschiede beim MVSQ orientieren sich an diesen Ergebnissen unter Berücksichtigung der Ergebnisse aus der Untersuchung zur konvergenten Validität (Kapitel 9.2) und lauten wie folgt:

H2a: Frauen bevorzugen das Wertesystem ***Gleichheit*** stärker als Männer.

H2b: Männer scoren höher auf den Wertesystemen ***Macht*** und ***Erfolg***.

Die Hypothesen gelten für die Annäherungswertesysteme und in entgegengesetzter Richtung für die Vermeidungswertesysteme.

Studiengänge und Betriebswirtschaft-Schwerpunkte

Auf Basis der Experteneinschätzung wurden die Hypothesen für die Studiengänge und BW-Schwerpunkte mit mehr als 20 Studierenden abgeleitet. Für die meisten Studiengänge liegen dabei Einschätzungen zur überdurchschnittlichen Ausprägung, bei manchen auch zur unterdurchschnittlichen Ausprägung vor. Für die Formulierung der Hypothesen gilt, dass die genannten Wertesysteme überdurchschnittlich ausgeprägt erwartet werden. Falls eine unterdurchschnittliche Ausprägung erwartet wird, wird dies zusätzlich so angemerkt. Ferner gelten die Hypothesen so wie formuliert für die Annäherungswertesysteme und in umgekehrter Richtung für die Vermeidungswertesysteme.

H3a: Betriebswirtschaft: ***Erfolg***.

H3b: Informatik: ***Verstehen*** und ***Gewissheit***.

H3c: International Relations and Management: ***Nachhaltigkeit***.

Bei den Schwerpunkten im Studiengang Betriebswirtschaft führen die Experteneinschätzungen zu folgenden Hypothesen:

H4a: Personalmanagement: ***Gewissheit*** und ***Gleichheit***.

H4b: Projektmanagement: ***Erfolg***.

H4c: Finanzen: ***Erfolg***.

Aufgabenbereiche und Hierarchieebene

Auch bei den Hypothesen zu den Wertesystemausprägungen je nach Aufgabenbereich gilt, dass sie so verstanden werden sollen, dass die Mittelwerte der genannten Wertesysteme in der Annäherungsdimension höher und in der Vermeidungsdimension niedriger vermutet werden. Wenn niedrigere Ausprägungen von Wertesystem-Annäherung (und dementsprechend höhere Ausprägungen in der Vermeidungsdimension) hypothetisiert werden, wird dies zusätzlich formuliert.

H5a: Vertrieb: Höhere Ausprägungen von ***Erfolg***, sowie niedrigere Ausprägungen von ***Gewissheit*** und ***Geborgenheit***.

H5b: Forschung und Entwicklung (FuE): ***Verstehen***.

H5c: Personal: Höhere Ausprägungen von ***Gewissheit*** und ***Gleichheit*** sowie niedrigere Ausprägungen von ***Macht***.

H5d: IT: ***Verstehen*** und ***Gewissheit***.

H5e: Produktion: ***Gewissheit***.

H5f: Unternehmensführung: Höhere Ausprägungen von ***Erfolg*** und ***Macht*** sowie niedrigere Ausprägungen von ***Gewissheit*** und ***Geborgenheit***.

H5g: Logistik: ***Erfolg***.

H5h: Organisation und Verwaltung (OuV): Höhere Ausprägungen von ***Gewissheit*** und ***Gleichheit*** sowie niedrigere Ausprägung von ***Macht***.

H5i: Kundendienst: Höhere Ausprägung von ***Gleichheit***.

H5j: Marketing: ***Erfolg*** und ***Verstehen***.

H5k: Abteilungsleitung: ***Macht***.

H5l: Finanzen: Höhere Ausprägungen von ***Erfolg*** und niedrigere Ausprägung von ***Gleichheit***.

Zwar wurden im MVSQ fünf verschiedene Optionen zur Angehörigkeit der Hierarchieebene angeboten, allerdings vermuten die Experten nur Zusammenhänge zwischen der obersten und untersten Ebene sowie der Gruppe der Selbstständigen.

H6a: Mitarbeiter, die der höchsten Hierarchieebene angehören, haben höhere Ausprägungen von ***Macht***.

H6b: Mitarbeiter eine Ebene unter der Geschäftsführung (GF) haben keine charakteristischen Wertesystempräferenzen.

H6c: Mitarbeiter zwei Ebenen unter der GF haben ebenso keine charakteristischen Wertesystempräferenzen.

H6d: Mitarbeiter, die der untersten Hierarchieebene angehören, haben niedrigere Ausprägungen von ***Macht*** und höhere Ausprägungen von ***Gewissheit***.

H6e: Selbstständige haben höhere Ausprägungen von ***Verstehen*** und ***Erfolg***.

10.1.2 Ergebnisse

Im Folgenden werden die Ergebnisse der Untersuchungen zur konkurrenten Validität beschrieben. Der Ergebnisteil gliedert sich in drei Teile. Im ersten werden die Zusammenhänge zwischen Wertesystemen und den allgemeinen soziodemografischen Variablen Alter und Geschlecht anhand von Produkt-Moment-Korrelationen bzw. Welch's t-Test für unabhängige Stichproben berichtet Im zweiten Teil werden Wertesysteme der Studierendenstichprobe gegliedert nach Studiengang bzw. Studiengangsschwerpunkt der Betriebswirtschaftsstudierenden mit einfaktoriellen ANOVAs und Post-hoc Tukey's HSD-Tests verglichen. Dem dritten Teil liegt die kumulierte Stichprobe der Berufstätigen zugrunde. Darin werden die Merkmalsausprägungen nach Aufgabenbereichen und Hierarchieebenen ebenfalls mit einfaktoriellen ANOVAs und Tukey's HSD-Tests analysiert.

10.1.2.1 Alter und Geschlecht

Die Korrelationen zwischen Wertesystemen und Alter (Tabelle 50) waren insgesamt niedrig (alle $r \leq |0.25|$). Nur mit dem Wertesystem ***Macht***^{V} liegt eine Korrelation vor, die mit $r = -.25$ im Betrag über $.20$ lag. Zwar wurden die meisten der Korrelationen signifikant, dies ist aber angesichts der großen Stichprobe ($N = 1316$) nicht verwunderlich. Hypothese 1, dass nur geringe Korrelationen zwischen Wertesystemen und Alter bestehen, kann somit bestätigt werden. Die Befunde bestätigen damit die Ergebnisse bestehender Forschung.

Tabelle 50. Korrelationen der Wertesysteme mit Alter.

	GB	MA	GW	ER	GL	VE	NA
MVSQ^A	-.17	.19	-.16	-.14	-.14	.04	.00
MVSQ^V	.08	-.25	.13	-.14	.07	-.06	-.07

Anmerkung. Korrelationen > .09 waren signifikant zum Niveau p < .001, Korrelationen > .08 signifikant zum Niveau p < .01 und Korrelationen > .06 signifikant zum Niveau p < .05.

Bei der Analyse der Wertesystemausprägungen nach Geschlecht (Tabelle 51) wurden in der Annäherungsskala Unterschiede zwischen den Geschlechtern bei allen Wertesystemen außer ***Erfolg***^{A} signifikant. Frauen haben signifikant höhere Ausprägungen auf den Wertesystemen ***Geborgenheit***^{A}, ***Gewissheit***^{A}, ***Gleichheit***^{A}, und ***Nachhaltigkeit***^{A} und Männer bevorzugen signifikant die Wertesysteme ***Macht***^{A} und ***Verstehen***^{A}. Die Hypothese 2a kann demnach voll und 2b halb (für ***Macht***^{A}) bestätigt werden, sie decken jedoch nicht die ganze Bandbreite der Befunde ab. Auffällig ist, dass Frauen auf allen Wertesystemen der Gruppe der *sacrifice-self*-Wertesysteme und Männer auf zwei von drei Wertesystemen der *express-self*-Gruppe höher scoren. Bezüglich der Vermeidungsskala gibt es signifikante Unterschiede auf allen Wertesystemen außer ***Nachhaltigkeit***^{V}. Frauen scoren höher auf ***Macht***^{V}, ***Erfolg***^{V} und ***Verstehen***^{V},

Männer höher auf ***Geborgenheit***^{V}, ***Gewissheit***^{V}, ***Gleichheit***^{V}, und ***Nachhaltigkeit***^{V}. Die Ergebnisse sind somit vom Muster her ähnlich, nur mit jeweils entgegengesetztem Vorzeichen. Für die Vermeidungswertesysteme können die Hypothesen 2a und 2b somit vollständig bestätigt werden.

Die Effektstärken bewegen sich zwischen $d = 0.04$ und $d = 0.6$, mit einer mittleren Effektstärke von 0.3, liegen damit also etwas höher als in der Meta-Analyse von Schwartz und Rubel (2005).

Tabelle 51. Geschlechterunterschiede der Wertesysteme: t-Tests und Effektstärken.

	Frauen		Männer		t-Tests				
	M	SD	M	SD	t	df	p	95 % KI	d
GB^{A}	0.07	0.45	-0.08	0.43	0.41	1272	<.001	-0.199 - -0.105	0.35
MA^{A}	-0.05	0.43	0.05	0.43	1.24	1279	<.001	0.053 - 0.147	0.23
GW^{A}	0.10	0.49	-0.09	0.48	0.48	1277	<.001	-0.249 - -0.144	0.41
ER^{A}	-0.01	0.41	0.01	0.46	4.39	1301	.51	-0.031 - 0.063	0.04
GL^{A}	0.14	0.41	-0.11	0.41	2.54	1247	<.001	-0.293 - -0.203	0.60
VE^{A}	-0.06	0.42	0.07	0.45	0.07	1297	<.001	0.076 - 0.17	0.28
NA^{A}	0.03	0.50	-0.04	0.48	2.27	1314	.02	-0.115 - -0.009	0.13
GB^{V}	-0.07	0.44	0.07	0.42	0.41	1272	<.001	0.096 - 0.189	0.33
MA^{V}	0.08	0.44	-0.08	0.46	1.24	1279	<.001	-0.202 - -0.104	0.34
GW^{V}	-0.08	0.49	0.08	0.46	0.48	1277	<.001	0.108 - 0.212	0.34
ER^{V}	0.12	0.44	-0.10	0.43	4.39	1301	<.001	-0.262 - -0.168	0.50
GL^{V}	-0.05	0.45	0.04	0.42	2.54	1247	<.001	0.042 - 0.137	0.20
VE^{V}	0.07	0.43	-0.07	0.45	0.07	1297	<.001	-0.188 - -0.092	0.32
NA^{V}	0.02	0.42	-0.02	0.41	2.27	1314	.09	-0.084 - 0.006	0.09

Anmerkung. d = Effektstärke Cohen's d mit gepoolter Varianz.

10.1.2.2 Studiengänge und Studienganasschwerpunkte

Im Folgenden werden die Ergebnisse zur Fragestellung untersucht, ob sich Studiengänge durch charakteristische Wertesysteme unterscheiden lassen und ob diese Unterschiede hypothesenkonsistent sind. Dazu wurden einfaktorielle ANOVAs mit je Studiengang bzw. Schwerpunkt als Faktor gerechnet. Zwar ergaben Tests auf Normalität nach Shapiro-Wilk, dass bei den BW-Studierenden ***Nachhaltigkeit*** auf beiden Dimensionen (beide $p < .01$), ***Macht***^{A} im Studiengang IRM ($p < .05$) und ***Gewissheit*** (beide Dimensionen) bei den sonstigen Studiengängen (beide $p < .05$) nicht normalverteilt waren, da jedoch parametrische Varianzanalysen als robust gegen Verletzung der Normalverteilungsannahme gelten (Box, 1953) und da die wichtigere Voraussetzung der Varianzhomogenität der Wertesysteme zwischen den Gruppen mit Levene-Tests auf Varianzhomogenität (Eid et al., 2015) bestätigt werden konnte, wurden für alle Wertesysteme normale, parametrische ANOVAs durchgeführt. Außerdem wurden die Histogramme der Verteilungen begutachtet, die nach Augenmaß lediglich leicht schief, jedoch nie auffallend stark schief oder zweigipflig waren. Auch die unterschiedlichen Stichprobengrößen stellen kein Hindernis bei der Berechnung dar, wenn Varianzhomogenität gegeben ist (Tabachnick & Fidell, 2007).

Die einfaktoriellen Varianzanalysen (Tabelle 52) zeigen signifikante Haupteffekte für eine Mehrzahl der Wertesysteme, wenngleich die Effektstärken alle relativ niedrig sind ($\eta^2 \leq .12$). Auf der Annäherungsskala unterscheiden sich die Mittelwerte von ***Gewissheit***^{A}, ***Erfolg***^{A}, ***Gleichheit***^{A}, ***Verstehen***^{A} und ***Nachhaltigkeit***^{A}, auf der Vermeidungsskala von ***Macht***^{V}, ***Erfolg***^{V}, ***Verstehen***^{V} und ***Nachhaltigkeit***^{V}. Post-hoc haben Tukey's HSD-Tests folgende Unterschiede zwischen den einzelnen Studiengängen gezeigt:

Beim ***Macht***-Wertesystem gibt es nur einen signifikanten Unterschied, und zwar auf der MVSQ^V-Skala. BW-Studierende haben signifikant ($p < .001$) niedrigere Ausprägungen als Studierende der sonstigen Studiengänge.

Auf dem ***Gewissheit***-Wertesystem unterscheiden sich die Studierenden des Studiengangs IRM durch signifikant niedrigere Ausprägungen auf dem ***Gewissheit***^{A}-Wertesystem als Studierende der Informatik ($p < .05$).

Bzgl. des Wertesystems ***Erfolg***^{A} unterscheiden sich die BW-Studierenden hochsignifikant von den Studierenden der sonstigen Studiengänge ($p < .001$) durch deutlich höhere Scores. Bei ***Erfolg***^{V} verhält es sich umgekehrt, hier haben die BW-Studierenden niedrigere Ausprägungen zum Niveau von $p < .001$.

Beim ***Gleichheit***-Wertesystem ist nur der Vergleich zwischen den Ausprägungen der Annäherungswertesysteme der Studierenden der Betriebswirtschaft und der sonstigen Studiengänge signifikant ($p < .05$). Die Gruppe der BW-Studierenden weist hier einen niedrigeren Mittelwert auf.

Auf der Annäherungsskala (***Verstehen***^{A}) weist die Gruppe der Informatik-Studierenden einen signifikant höheren Mittelwerte als die Gruppen der BW-Studierenden ($p < .001$),

Tabelle 52. Mittelwerte und Standardabweichungen der Wertesysteme in Abhängigkeit des Studiengangs, sowie Kennwerte der einfaktoriellen ANOVAs.

	BW		Informatik		IRM		Sonstige		ANOVA	
	M	SD	M	SD	M	SD	M	SD	$F(3, 445)$	η^2
GB^A	0.05	0.45	0.04	0.42	-0.16	0.59	0.08	0.43	1.78	.01
MA^A	-0.04	0.45	-0.25	0.40	-0.16	0.47	-0.13	0.46	2.45	.02
GW^A	0.10	0.53	0.23	0.46	-0.18	0.51	0.06	0.51	2.67*	.02
ER^A	0.13	0.41	-0.05	0.50	-0.05	0.46	-0.10	0.44	10.03***	.06
GL^A	0.04	0.43	0.22	0.56	0.08	0.41	0.17	0.44	3.25*	.02
VE^A	-0.09	0.46	0.39	0.39	-0.03	0.37	0.11	0.45	12.83***	.08
NA^A	-0.12	0.51	0.03	0.48	0.31	0.61	0.26	0.51	19.61***	.12
GB^V	-0.02	0.48	0.04	0.48	0.14	0.53	-0.06	0.44	1.2	.01
MA^V	0.06	0.46	0.30	0.39	0.11	0.52	0.24	0.47	5.6***	.04
GW^V	-0.07	0.52	-0.07	0.49	0.20	0.59	-0.05	0.51	1.78	.01
ER^V	-0.01	0.45	0.08	0.54	0.05	0.36	0.25	0.45	10.2***	.06
GL^V	-0.04	0.49	-0.18	0.57	0.02	0.43	-0.06	0.50	0.78	.01
VE^V	0.13	0.47	-0.29	0.48	0.05	0.42	-0.09	0.46	11.05***	.07
NA^V	0.10	0.43	-0.09	0.46	-0.26	0.43	-0.13	0.40	12.71***	.08

Anmerkung. BW = Betriebswirtschaft; IRM = International Relations and Management; Sonstige = Sonstige Studiengänge; *** p < .001, ** p < .01, * p < .05.

IRM-Studierenden ($p < .05$) und die der sonstigen Studierenden ($p < .05$) auf. Auf der Vermeidungsskala ist nur der Vergleich mit den BW-Studierenden signifikant ($p < .001$). Des Weiteren haben die BW-Studierenden auf beiden Skalen hochsignifikant unterschiedliche Ausprägungen als die Gruppe der Sonstigen ($p < .001$). Bei Annäherungs-Scores ist der Mittelwert niedriger, bei den Vermeidungs-Scores höher.

Bzgl. des Wertesystems ***Nachhaltigkeit*** weisen die IRM-Studierenden hochsignifikant höhere Mittelwerte als die BW-Studierenden ($p < .001$) auf der Annäherungsskala sowie einen hochsignifikant niedrigeren Mittelwert auf der Vermeidungsskala ($p < .001$) auf. Abgesehen davon charakterisieren sich BW-Studierende durch signifikant niedrigere Ausprägungen auf ***Nachhaltigkeit***^{V} und höhere Ausprägungen auf ***Nachhaltigkeit***^{V} (beide $p < .001$).

Insgesamt kann auf Basis dieser festgestellten Unterschiede gesagt werden, dass hohe Ausprägungen auf den Wertesystemen ***Erfolg*** und ***Macht*** (hohe Annäherung und niedrige Vermeidung) sowie niedrige Ausprägungen auf den Wertesystemen ***Verstehen*** und ***Nachhal-***

tigkeit charakteristisch für Studierende der Betriebswirtschaft waren, Informatik-Studierende vor allem durch das Wertesystem ***Verstehen*** und IRM-Studierende tendenziell am ehesten durch hohe Ausprägungen des Wertesystems ***Nachhaltigkeit*** sowie durch niedrige Ausprägungen von ***Gewissheit*** beschrieben werden können.

Die Hypothesen zu diesen Gruppen können somit größtenteils bestätigt werden. Betriebswirtschaftsstudierende lassen sich eindeutig durch das Wertesystem ***Erfolg*** charakterisieren (Hypothese 3a), Informatik-Studierende durch das Wertesystem ***Verstehen*** (Hypothese 3b) und IRM-Studierende durch das Wertesystem ***Nachhaltigkeit***. Nicht bestätigt wurde der Zusammenhang von ***Gewissheit*** und Informatikstudierenden. Zudem wurde festgestellt, dass Betriebswirtschaftsstudierende die Wertesysteme ***Verstehen***^{A} und ***Nachhaltigkeit***^{A} tendenziell niedriger und ***Verstehen***^{V} und ***Nachhaltigkeit***^{V} höher ausgeprägt haben.

Des Weiteren wurden einfaktorielle ANOVAs für jedes Wertesystem über die Stichprobe der Betriebswirtschaftsstudierenden mit dem Faktor Studienschwerpunkt gerechnet (Tabelle 53). Die Normalverteilungsannahme gilt für alle Wertesysteme außer ***Nachhaltigkeit***^{A} in der Stichprobe der sonstigen Schwerpunkte ($p < .05$) und ***Nachhaltigkeit***^{V} in Stichprobe Personalmanagement ($p < .01$). Die p-Werte beziehen sich auf Shapiro-Wilk-Tests, die für alle Gruppen und Wertesysteme durchgeführt wurden. Die Varianzhomogenität ist bei allen Wertesystemen und Gruppen gegeben, da keiner der Levene-Tests signifikant wurde. Somit konnten erneut normale, parametrische ANOVAs durchgeführt werden.

Beim Vergleich der Wertesystemmittelwerte innerhalb der Betriebswirtschaftsstudierenden kategorisiert nach Schwerpunkten fällt auf, dass deutlich weniger Mittelwertsunterschiede signifikant wurden und auch die Effektstärken niedriger waren ($\eta^2 \leq .08$). Dies war insofern plausibel, da es sich um Studierende desselben Studiengangs handelt und somit bereits eine Vorselektion stattfand. Dennoch unterscheiden sich die Schwerpunkt-Gruppen in den Wertesystemen ***Geborgenheit***, ***Gleichheit*** und ***Verstehen*** auf beiden Skalen gleichermaßen. Für diese drei Wertesysteme wurden erneut Post-hoc Tukey's HSD-Tests durchgeführt.

Bei ***Geborgenheit*** haben BW-Studierende mit Schwerpunkt Personalmanagement signifikant höhere Ausprägungen auf der Annäherungsskala als die Studierenden mit Schwerpunkt Finanzen oder sonstigen Schwerpunkten (beide $p < .05$). Auf der Vermeidungsskala haben die Personal-Studierenden nur signifikant niedrigere Ausprägungen als diejenigen mit Finanz-Schwerpunkt ($p < .05$). Letztere haben zudem signifikant niedrigere Ausprägungen als diejenigen mit Projektmanagement-Schwerpunkt auf dem ***Geborgenheit***^{V}-Wertesystem.

Bzgl. ***Gleichheit*** unterscheidet sich Personal- von Finanz- ($p < .01$) und sonstigen Schwerpunkten ($p < .001$) signifikant durch höhere ***Gleichheit***^{A}-Scores und niedrigere ***Gleichheit***^{V}-Scores ($p < .01$ bzw. $< .001$).

Bzgl. des Wertesystems ***Verstehen***^{A} weisen diejenigen mit Finanz-Schwerpunkt signifikant höhere Ausprägungen als alle anderen drei Gruppen auf (alle drei $p < .05$). Auf der Ver-

Tabelle 53. Mittelwerte und Standardabweichungen der Wertesysteme von Betriebswirtschaftsstudierenden in Abhängigkeit des Schwerpunkts, sowie Kennwerte der einfaktoriellen ANOVAs.

	Finanzen		Personal		PJM		Sonstige		ANOVA	
	M	SD	M	SD	M	SD	M	SD	$F(3, 245)$	η^2
GB^A	-0.09	0.31	0.23	0.45	0.03	0.45	0.01	0.46	4.23^{**}	.05
MA^A	0.05	0.51	-0.16	0.39	0.02	0.45	-0.03	0.46	2.01	.02
GW^A	0.08	0.50	0.19	0.52	0.01	0.47	0.08	0.55	0.97	.01
ER^A	0.19	0.51	0.08	0.42	0.15	0.38	0.14	0.39	0.51	.01
GL^A	-0.14	0.34	0.23	0.40	0.12	0.37	-0.03	0.45	7.05^{***}	.08
VE^A	0.20	0.39	-0.13	0.47	-0.16	0.45	-0.11	0.46	3.42^{*}	.04
NA^A	-0.19	0.62	-0.21	0.49	-0.12	0.46	-0.07	0.51	1.17	.01
GB^V	0.22	0.46	-0.09	0.56	-0.15	0.40	0.01	0.47	3.44^{*}	.04
MA^V	0.01	0.47	0.21	0.43	0.03	0.41	0.02	0.48	2.38	.03
GW^V	-0.01	0.47	-0.16	0.53	-0.10	0.43	-0.03	0.55	0.91	.01
ER^V	-0.15	0.42	0.10	0.46	0.00	0.54	-0.03	0.41	2.11	.03
GL^V	0.13	0.38	-0.25	0.52	-0.13	0.46	0.06	0.47	7.04^{***}	.08
VE^V	-0.13	0.44	0.21	0.48	0.17	0.50	0.12	0.46	2.77^{*}	.03
NA^V	0.09	0.50	0.17	0.47	0.01	0.38	0.10	0.41	1.03	.01

Anmerkung. PJM = Projektmanagement; Sonstige = Sonstige Schwerpunkte; *** p < .001, ** p < .01, * p < .05.

meidungsskala haben sie jedoch nur niedrigere Ausprägungen als die Personal-Studierenden ($p < .05$).

In Summe kann festgestellt werden, dass sich BW-Studierende, die sich als Gruppe von anderen Studierendengruppen bereits unterscheiden, auch noch untereinander in einigen Wertesystemen anhand ihrer Schwerpunktwahl unterscheiden und dadurch charakterisieren lassen. Hypothese 4a kann damit teilweise bestätigt werden. Das Wertesystem ***Gleichheit*** kann als charakteristisch für Personal-Studierende verwendet werden. Zudem unterscheiden sich Personal-Studierende durch überdurchschnittliche Ausprägung von ***Geborgenheit***. Nicht bezeichnend ist hingegen das Wertesystem ***Gewissheit***. Hypothese 4b kann somit nicht bestätigt werden, zumindest nicht für den Vergleich innerhalb der BW-Studierenden. Studierende mit Finanz-Schwerpunkt (Hypothese 4c) zeichnen sich nicht wie hypothetisiert durch das Wertesystem ***Erfolg*** aus, sondern durch hohe Ausprägungen auf ***Verstehen***^{A} und niedrigen auf ***Geborgenheit***^{A}.

10.1.2.3 Aufgabenbereiche und Hierarchieebenen

Die Tabellen 54 und 55 zeigen die Mittelwerte aller Wertesysteme aufgeteilt nach Aufgabenbereichen mit mehr $n \geq 20$ Mitarbeitern und die Ergebnisse der entsprechenden ANOVAs. Shapiro-Wilk-Tests ergaben, dass zum .05-Niveau unter den Abteilungsleitern ***Gewissheit***^{A}, im Marketing ***Gleichheit***^{A} und ***Gleichheit***^{V}, in OuV ***Geborgenheit***^{A}, im Personal ***Erfolg***^{A} und ***Gleichheit***^{V}, in den sonstigen Bereichen ***Nachhaltigkeit***^{A} und ***Nachhaltigkeit***^{V} (als einziges $p < .01$) die Hypothese der Normalverteilung abgelehnt werden musste. Dennoch wurden auch hier aufgrund der Robustheit der parametrischen ANOVA gegenüber Abweichungen der Normalität ebendiese berechnet. Bei den Wertesystemen ***Nachhaltigkeit***^{A} und ***Gewissheit***^{V} haben Levene-Tests ergeben, dass die Voraussetzung homogener Varianzen verletzt war (beide $p < .05$). Dementsprechend wurden bei diesen ANOVAs robuste F-Tests für inhomogene Varianzen nach Welch (1951) berechnet. Auch bei diesen ergab sich jeweils ein signifikanter Haupteffekt. Insgesamt zeigten somit zwölf der 14 Varianzanalysen einen signifikanten Haupteffekt bei allerdings insgesamt niedrigen Effektstärken (η^2 zwischen .02 und .10).

Tukey's HSD-Tests ergaben, dass Personen mit Jobs in der Unternehmensführung (UF) signifikant niedrigere Ausprägungen auf ***Geborgenheit***^{A} aufweisen als Mitarbeiter in Forschung- und Entwicklung (FuE), Personal, Vertrieb (alle $p < .05$), Organisation und Verwaltung (OuV, $p < .001$) und sonstigen Aufgabenbereichen ($p < .01$) auf. Auf ***Geborgenheit***^{V} sind die Unterschiede entgegengesetzt, d.h. die Ausprägungen waren höher als in den Bereichen FuE, Marketing, Personal, Vertrieb (alle $p < .05$), OuV und sonstige Bereiche (beide $p < .01$).

Beim ***Macht***-Wertesystem unterscheiden sich zum einen FuE-Mitarbeiter auf der Annäherungsskala durch niedrigere Ausprägungen im Vergleich zu IT, UF (beide $p < .05$) und Vertrieb ($p < .001$). Auf der Vermeidungsskala weisen sie gegenüber IT, UF (beide $p < .01$), Logistik und Vertrieb (beide $p < .001$) signifikant höhere Ausprägungen vor. Logistiker haben des Weiteren auch niedrigere Ausprägungen als Personaler ($p < .01$) und Mitarbeiter sonstiger Aufgabenbereiche ($p < .05$) auf dem ***Macht***^{V}-Wertesystem und Vertriebler haben höhere Ausprägungen auf ***Macht***^{A} als Mitarbeiter im Personal- ($p < .05$) und sonstigen Bereichen ($p < .01$).

Am stärksten fallen beim ***Gewissheit***-Wertesystem wie schon bei ***Geborgenheit*** Unterschiede von Mitarbeitern der Unternehmensführung auf. Diese haben auf der Annäherungsskala im Vergleich mit fast allen Vergleichsgruppen niedrigere Ausprägungen und auf der Vermeidungsskala höhere Ausprägungen von ***Gewissheit***. Die entsprechenden p-Werte liegen bei der MVSQ^A-Skala wie folgt: Abteilungsleiter $p < .05$, Finanzen, FuE, Kundendienst, OuV, Personal, Produktion, Vertrieb und sonstige alle $p < .001$, sowie verglichen mit Marketing und Logistik $p < .01$. Auf der MVSQ^V-Skala sind die Ergebnisse ähnlich in umgekehrter Richtung, d.h. signifikant niedrigere Ausprägungen im Vergleich zu Mitarbeitern aus dem Finanzbereich, FuE, Kundendienst, OuV, Personal, Produktion, Vertrieb und sonstige mit p-Werten $< .001$. Im Vergleich zu Abteilungsleitern, Marketing und Logistik liegen die entsprechenden p-Werte

$< .01$. Abgesehen davon weisen Berufstätige im IT-Bereich niedrigere Ausprägungen von ***Gewissheit***^{A} als Mitarbeiter aus OuV, Personal, Produktion und sonstigen Bereichen auf (alle $p < .05$). Keine Unterschiede gab es für IT-Mitarbeiter auf ***Gewissheit***^{V}.

Beim Wertesystem ***Erfolg*** treten nur signifikante Effekte auf der Vermeidungsdimension auf. Mitarbeiter im Vertrieb haben niedrigere Ausprägungen als Mitarbeiter aus FuE, OuV, Personal und sonstigen Bereichen (alle $p < .001$) sowie Produktion ($p < .05$). Darüber hinaus gab es auch Unterschiede zwischen Mitarbeitern aus der UF, die ebenso niedrigere Ausprägungen als Berufstätige in Personal ($p < .01$), OuV und sonstigen Bereichen (beide $p < .05$) aufweisen.

Bei ***Gleichheit*** haben Personaler höhere Ausprägungen auf ***Gleichheit***^{A} als Beschäftigte in IT, Vertrieb (beide $p < .05$) und Produktion ($p < .01$). Mitarbeiter aus OuV haben ebenso höhere Ausprägungen als ITler, Vertriebler (beide $p < .05$) und als Mitarbeiter aus dem Bereich Produktion ($p < .01$). Bei ***Gleichheit***^{V} traten keine signifikanten Effekte auf.

Bei ***Verstehen*** gibt es einige Unterschiede mit FuE-Mitarbeitern (überdurchschnittlich) sowie mit Vertriebsmitarbeitern (unterdurchschnittlich). FuE-Mitarbeiter haben höhere Ausprägungen auf ***Verstehen***^{A} als Abteilungsleiter ($p < .05$), Mitarbeiter aus Kundendienst, Logistik, Personal, Produktion, Vertrieb und sonstigen Bereichen (alle $p < .001$). Auf der Vermeidungs-Skala haben FuE-Mitarbeiter niedrigere Ausprägungen als Abteilungsleitern, Mitarbeitern aus OuV und UF (alle $p < .05$) sowie der Logistik, Produktion, sonstigen Bereichen ($p < .01$) und Personal ($p < .001$). Vertriebsmitarbeiter haben außerdem niedrigere Ausprägungen auf ***Verstehen***^{A} als Mitarbeiter aus der UF ($p < .05$), der IT ($p < .01$) und sonstige Bereiche ($p < .001$). Auf der Vermeidungsskala wurde bei den Vertrieblern nur der Unterschied (höhere Ausprägung) mit den sonstigen Bereichen signifikant ($p < .01$).

Bezogen auf das ***Nachhaltigkeit***-Wertesystem unterscheidet sich die Gruppe der Vertriebler von Mitarbeitern aus FuE und sonstigen Bereichen hoch signifikant durch niedrigere Ausprägungen auf ***Nachhaltigkeit***^{A} und höheren auf ***Nachhaltigkeit***^{V} ($p < .001$). Zudem haben sie niedrigere Ausprägungen von ***Nachhaltigkeit***^{A} als Mitarbeiter aus der OuV ($p < .05$) und höhere Ausprägungen auf ***Nachhaltigkeit***^{V} als Mitarbeiter aus der IT ($p < .01$).

Zusammenfassend kann aus diesen Ergebnissen abgeleitet werden, dass Wertesysteme durchaus bezeichnenden Charakter für einige Aufgabenbereiche haben können und dass dabei die Annäherungs- und Vermeidungsdimensionen zwar tendenziell, aber nicht immer, entgegengesetzt sind. Im nächsten Abschnitt werden die Ergebnisse auf ihre Hypothesenkonformität geprüft.

Hypothese 5a kann nur teilweise bestätigt werden, da Vertriebsmitarbeiter keine hohen Ausprägungen auf ***Erfolg***^{A}, allerdings besonders niedrige Ausprägungen im Vergleich mit einigen anderen Bereichen auf ***Erfolg***^{V} haben. Es muss abgelehnt werden, dass ***Gewissheit*** und ***Geborgenheit*** charakteristisch für Vertriebsaufgaben sind. Die Ergebnisse hier zeigen keinen Zusammenhang. Bezeichnender sind dagegen hohe Ausprägungen von ***Macht***^{A}, ***Verstehen***^{V} und ***Nachhaltigkeit***^{V}. Hypothese 5b kann vollständig bestätigt werden. FuE-Mitarbeiter wei-

sen höhere Scores von ***Verstehen***^{A} und niedrigere von ***Verstehen***^{V} verglichen mit fast allen anderen Gruppen auf. Bei Hypothese 5c kann wieder ein Teil bestätigt werden, nämlich, dass ***Gleichheit*** wie hypothetisiert charakteristisch für Mitarbeiter aus dem Personalbereich ist. Nicht charakteristisch sind dagegen ***Gewissheit*** und ***Macht***. Hypothese 5d trifft nicht zu und muss verworfen werden. Weder haben ITler auffällige Ausprägungen auf dem Wertesystem ***Verstehen***, noch haben sie hohe Ausprägungen auf ***Gewissheit***^{A}. Im Gegenteil, die ***Gewissheit***^{A}-Scores sind im Vergleich zu vier untersuchten Gruppen niedriger ausgeprägt. Hypothese 5e kann erneut nur eingeschränkt bestätigt werden. Nur im Vergleich zu den Bereichen IT und UF haben Produktions-Mitarbeiter höhere Ausprägungen von ***Gewissheit***^{A}. Zwar war die durchschnittliche Ausprägung von ***Gewissheit***^{A} positiv und höher als viele Werte anderer Bereiche, allerdings kamen diese Signifikanzen vermutlich eher dadurch zustande, dass IT- und UF-Mitarbeiter besonders niedrige Ausprägungen auf ***Gewissheit***^{A} aufweisen. Für ***Gewissheit*** und ***Geborgenheit*** kann Hypothese 5f auf beiden Dimensionen bestätigt werden. UF-Mitarbeiter haben hier extremere Ausprägungen im Vergleich zu vielen Gruppen. Für ***Erfolg*** und ***Macht*** kann dies nicht so eindeutig festgestellt werden. Hier wurden nur drei bzw. ein Mittelwertsunterschied signifikant. Hypothese 5g muss abgelehnt werden, da ***Erfolg*** nicht bezeichnend für Logistiker war. Charakteristisch ist eher ***Macht***^{V} (niedrige Ausprägungen). Zwar wurden einige Mittelwertsvergleiche signifikant, in denen OuV-Mitarbeiter einflossen, als charakteristisch für diesen Aufgabenbereich konnte jedoch kein Wertesystem identifiziert werden. Hypothese 5h muss deshalb abgelehnt werden. Auch Hypothese 5i trifft nicht zu. Für Kundendienst-Mitarbeiter wurde kein typisches Wertesystem gefunden. Ebenso wie bei Hypothese 5h zu den OuV-Mitarbeitern waren zwar Marketing-Mitarbeiter in signifikante Mittelwertsvergleiche involviert, es war jedoch kein Schema zu erkennen. Hypothese 5j wird deshalb abgelehnt. Dieselbe Feststellung kann für Abteilungsleiter und Mitarbeiter im Finanz-Bereich getroffen werden und führt dazu, dass die Hypothesen 5k und 5l verworfen werden.

Beim Vergleich der Wertesysteme in Abhängigkeit der Hierarchieebenen (Tabelle 56) wurden ebenfalls einige Abweichungen der Normalverteilung festgestellt: ***Macht***^{A} bei GF ($p < .05$), ***Verstehen***^{A} bei GF -1 ($p < .05$), ***Erfolg***^{A} bei GF -2 ($p < .05$) sowie ***Geborgenheit***^{A}, ***Macht***^{V} (beide $p < .05$), ***Nachhaltigkeit***^{A} ($p < .001$) und ***Erfolg***^{V} ($p < .01$) bei GF ≤ 3. Wie schon zuvor wurden für diese dennoch parametrische ANOVAs berechnet. Bei den Analysen von ***Macht***^{V} und ***Gewissheit***^{V} wurde aufgrund eines signifikanten Levene-Tests ($p < .05$) der robuste F-Test durchgeführt. Insgesamt ergaben sich relativ wenige Unterschiede und auch die Effektstärken liegen im Vergleich zu den vorherigen Untersuchungen merklich niedriger. Dennoch gibt es auf beiden Skalen Unterschiede bzgl. der Wertesysteme ***Geborgenheit***, ***Macht***, ***Gewissheit*** und ***Nachhaltigkeit*** auf beiden Dimensionen. Auf der Annäherungsskala wurde zudem ein Mittelwertsvergleich des Wertesystems ***Gleichheit***^{A} signifikant. Im Folgenden werden die Ergebnisse von Tukey's HSD-Tests nach Wertesystemen sortiert berichtet.

Bzgl. des Wertesystems ***Geborgenheit*** wurden ähnlich wie zuvor beim Aufgabenbereich UF drei HSD-Tests von Mitarbeitern auf der obersten Hierarchie-Ebene signifikant. Mitarbeiter, die der Geschäftsführung (GF) angehören, haben signifikant niedrigere Ausprägungen auf ***Geborgenheit***^{A} als Mitarbeiter zwei ($p < .05$) und mindestens drei Ebenen ($p < .01$) darunter. Auch haben Mitarbeiter der GF niedrigere Ausprägungen als Selbstständige ($p < .01$). Entgegengesetzt verhält es sich bei ***Geborgenheit***^{V}: Die Scores waren zum Niveau von .01 signifikant höher im Vergleich mit Mitarbeiter mindestens drei Ebenen unter GF und zum Niveau von .05 gegenüber Mitarbeitern zwei Ebenen unter GF und Selbstständigen.

Personen, die auf der Ebene GF -1 angesiedelt waren, wiesen höhere Ausprägungen von ***Macht***^{A} auf, sowohl gegenüber Mitarbeitern zwei ($p < .05$) als auch drei Ebenen darunter ($p < .01$). ***Macht***^{V} war im Vergleich zu GF ≤ 3 signifikant niedriger ausgeprägt ($p < .01$).

Gewissheit differierte im Verhältnis am stärksten zwischen den Hierarchieebenen. Mitarbeiter der GF haben signifikant niedrigere Ausprägungen auf ***Gewissheit***^{A} als Mitarbeiter aus allen Hierarchieebenen darunter ($p < .001$ für GF - 2 und ≤ 3, sowie $p < .01$ für GF - 1) und ebenso als Selbstständige ($p < .05$). Bei ***Gewissheit***^{V} wurden die Unterschiede (höhere Ausprägungen) zu GF -1 ($p < .05$), -2 und ≤ 3 (beide $p < .001$), nicht aber zu Selbstständigen signifikant. Des Weiteren scoren Mitarbeiter GF ≤ 3 höher auf ***Gewissheit***^{A} als Mitarbeiter GF - 1 ($p < .05$). Ansonsten unterscheiden sich noch die Selbstständigen insofern von den anderen Gruppen, dass sie niedrigere Ausprägungen auf ***Gewissheit***^{A} haben als die Gruppe der GF ≤ 3 ($p < .05$) und höhere Ausprägungen auf ***Gewissheit***^{V} als eben diese Gruppe ($p < .01$) und auch als Mitarbeiter zwei Ebenen unter GF ($p < .05$)

Bei ***Gleichheit*** scoren lediglich Selbstständige höher auf der Annäherungsdimension als Mitarbeiter der GF ($p < .05$) und bei ***Nachhaltigkeit***^{A} hatten Selbstständige und Mitarbeiter der Gruppe der zweiten Ebene unter GF höhere Ausprägungen als Mitarbeiter der Gruppe GF ≤ 3 (beide $p < .01$). Bei ***Erfolg*** und ***Verstehen*** wurden keine Mittelwertsunterschiede signifikant.

Zu den Hypothesen lässt sich Folgendes sagen: Hypothese 6a, die die Mitarbeiter der GF-Ebene betrifft, muss abgelehnt werden. Anstatt durch ***Macht*** können diese Mitarbeiter eher durch niedrige Ausprägungen auf ***Gewissheit***^{A} und ***Geborgenheit***^{A} sowie hohen Ausprägungen auf ***Gewissheit***^{V} und ***Geborgenheit***^{V} beschrieben werden. Die Hypothesen 6b und 6c müssen tendenziell verworfen werden, da sich Mitarbeiter auf einer und zwei Ebenen unter GF sehr wohl durch Wertesysteme charakterisieren lassen, zumindest im Vergleich mit der GF-Ebene. Sie haben z.B. höhere Werte auf ***Gewissheit***^{A} und niedrigere auf ***Gewissheit***^{V}. Hypothese 6d gilt teilweise. Mitarbeiter der unteren Hierarchieebenen scoren nur im Vergleich zu Mitarbeitern der ersten Ebene unter GF niedriger auf ***Macht***^{A} ($p < .01$) und höher auf ***Macht***^{V} ($p < .01$). Bzgl. des Wertesystems ***Gewissheit*** trifft die Hypothese 6d weitestgehend zu, denn Mitarbeiter der unteren Ebenen hatten signifikant niedrigere Ausprägungen als Mitarbeiter der GF, GF -1 und Selbstständige.

Tabelle 54. Mittelwerte und Standardabweichungen der Wertesysteme in Abhängigkeit des Jobs, sowie Kennwerte der einfaktoriellen ANOVAs.

	AL		Finanzen		FuE		IT		Kundendienst		Logistik		Marketing		OuV	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
GB^A	-0.13	0.47	-0.06	0.47	-0.03	0.42	-0.13	0.42	-0.03	0.43	-0.14	0.45	0.01	0.37	0.15	0.38
MA^A	0.15	0.45	0.11	0.37	-0.06	0.40	0.16	0.39	0.03	0.34	0.16	0.32	-0.02	0.40	-0.02	0.49
GW^A	-0.13	0.57	0.05	0.54	-0.04	0.45	-0.25	0.44	0.07	0.53	-0.10	0.37	-0.05	0.51	0.14	0.53
ER^A	-0.03	0.47	0.07	0.42	-0.03	0.45	0.07	0.38	-0.07	0.43	-0.13	0.34	-0.04	0.50	-0.06	0.40
GL^A	-0.13	0.39	-0.06	0.41	-0.08	0.41	-0.13	0.41	-0.13	0.33	-0.09	0.40	0.05	0.41	0.19	0.41
VE^A	-0.06	0.42	0.02	0.36	0.24	0.41	0.11	0.40	-0.13	0.38	-0.12	0.43	0.09	0.49	-0.03	0.37
NA^A	-0.05	0.49	-0.17	0.45	0.03	0.41	-0.06	0.44	-0.12	0.41	-0.00	0.42	-0.09	0.42	0.09	0.47
GB^V	0.13	0.43	-0.05	0.38	0.01	0.42	0.05	0.36	0.03	0.42	0.04	0.41	-0.07	0.41	-0.09	0.40
MA^V	-0.19	0.45	-0.18	0.40	0.09	0.44	-0.19	0.45	-0.16	0.48	-0.33	0.32	-0.04	0.37	-0.04	0.38
GW^V	-0.02	0.52	-0.11	0.39	0.05	0.43	0.14	0.47	-0.06	0.43	0.00	0.31	-0.01	0.45	-0.14	0.50
ER^V	-0.17	0.37	-0.19	0.43	0.00	0.42	-0.12	0.38	0.03	0.45	-0.02	0.40	-0.01	0.51	0.15	0.45
GL^V	-0.01	0.42	0.06	0.40	0.01	0.40	0.09	0.38	0.10	0.34	-0.03	0.38	0.07	0.47	-0.05	0.38
VE^V	0.07	0.44	-0.04	0.34	-0.25	0.39	-0.06	0.40	-0.02	0.44	0.09	0.48	-0.05	0.53	0.05	0.37
NA^V	0.01	0.44	-0.06	0.37	-0.07	0.36	-0.08	0.38	0.15	0.35	0.06	0.35	-0.04	0.39	-0.03	0.37

Anmerkung. AL = Abteilungsleitung; FuE = Forschung & Entwicklung; OuV = Organisation und Verwaltung; *** p < .001, ** p < .01, * p < .05.

Tabelle 55. Mittelwerte und Standardabweichungen der Wertesysteme in Abhängigkeit des Jobs, sowie Kennwerte der einfaktoriellen ANOVAs (Fortsetzung).

	Personal		Produktion		Sonstige		UF		Vertrieb		ANOVA	
	M	SD	M	SD	M	SD	M	SD	M	SD	$F(12, 854)$	η^2
GB^A	-0.00	0.48	-0.10	0.36	0.00	0.45	-0.31	0.41	-0.03	0.40	2.57^{**}	.03
MA^A	-0.06	0.40	0.08	0.41	-0.01	0.43	0.22	0.38	0.16	0.41	3.78^{***}	.05
GW^A	0.04	0.50	0.07	0.42	-0.01	0.46	-0.53	0.36	-0.04	0.42	5.64^{***}	.07
ER^A	-0.04	0.46	-0.10	0.42	-0.06	0.43	0.09	0.47	0.08	0.42	1.69	.02
GL^A	0.12	0.42	-0.17	0.38	0.00	0.44	-0.11	0.43	-0.08	0.39	3.3^{***}	.04
VE^A	-0.09	0.36	-0.12	0.45	0.06	0.41	0.11	0.38	-0.16	0.39	7.52^{***}	.10
NA^A	-0.04	0.47	0.04	0.40	0.06	0.51	-0.04	0.43	-0.20	0.38	3.6^{***a}	.04
GB^V	-0.00	0.43	0.10	0.40	-0.02	0.44	0.29	0.37	0.02	0.39	2.25^{**}	.03
MA^V	0.02	0.39	-0.10	0.41	-0.03	0.44	-0.25	0.45	-0.15	0.40	4.68^{***}	.06
GW^V	0.01	0.56	-0.11	0.43	0.03	0.47	0.43	0.39	0.04	0.41	4.47^{***b}	.05
ER^V	0.11	0.42	0.02	0.38	0.03	0.44	-0.23	0.37	-0.23	0.39	5.73^{***}	.07
GL^V	-0.07	0.44	0.06	0.34	-0.01	0.43	0.03	0.43	0.12	0.37	1.56	.02
VE^V	0.06	0.43	0.04	0.42	-0.06	0.42	0.02	0.39	0.12	0.38	5.44^{***}	.07
NA^V	0.05	0.44	0.04	0.30	-0.07	0.43	-0.03	0.45	0.17	0.42	3.65^{***}	.05

Anmerkung. PJM = Projektmanagement; Sonstige = Sonstige Aufgabenbereiche; UF = Unternehmensführung; *** p < .001, ** p < .01, * p < .05; ^{a} robuster F-Test nach Welch (1951) mit df = (12, 196); ^{b} robuster F-Test nach Welch (1951) mit df = (12, 197)

Tabelle 56. Mittelwerte und Standardabweichungen der Wertesysteme in Abhängigkeit der Hierarchieebene, sowie Kennwerte der einfaktoriellen ANOVAs.

	GF		-1		-2		≤ -3		Selbstst.		ANOVA	
	M	SD	M	SD	M	SD	M	SD	M	SD	$F(4, 839)$	η^2
GB^A	-0.04	0.43	-0.02	0.43	-0.11	0.39	-0.29	0.49	0.02	0.48	3.94^{**}	.02
MA^A	0.03	0.43	0.02	0.40	0.18	0.45	0.18	0.37	0.05	0.41	4.45^{**}	.02
GW^A	-0.05	0.45	0.02	0.46	-0.13	0.46	-0.46	0.47	-0.17	0.50	11^{***}	.05
ER^A	-0.09	0.42	-0.01	0.43	0.02	0.44	0.08	0.54	0.08	0.43	2.68^{*}	.01
GL^A	-0.06	0.45	-0.04	0.41	-0.07	0.42	-0.21	0.37	0.05	0.39	2.2	.01
VE^A	0.01	0.43	0.01	0.42	0.04	0.43	-0.01	0.42	-0.02	0.44	0.27	.00
NA^A	0.06	0.46	-0.09	0.43	0.00	0.51	0.02	0.43	0.12	0.48	5.67^{***}	.03
GB^V	0.03	0.39	-0.02	0.41	0.08	0.37	0.25	0.45	-0.01	0.51	4.32^{**}	.02
MA^V	-0.10	0.40	-0.04	0.42	-0.19	0.52	-0.25	0.37	-0.06	0.42	4.02^{**a}	.02
GW^V	0.00	0.45	-0.02	0.45	0.09	0.46	0.36	0.48	0.19	0.46	7.54^{***b}	.04
ER^V	-0.01	0.45	-0.04	0.42	-0.11	0.44	-0.18	0.45	0.05	0.44	2.36	.01
GL^V	0.00	0.42	0.03	0.39	0.04	0.42	0.13	0.35	-0.01	0.47	0.84	.00
VE^V	-0.05	0.41	-0.02	0.43	0.03	0.46	0.01	0.44	-0.02	0.45	0.59	.00
NA^V	-0.05	0.41	0.04	0.39	-0.07	0.44	-0.02	0.43	-0.00	0.43	2.84^{*}	.01

Anmerkung. GF = Geschäftsführung; -1 = eine Ebene unter GF; -2 = zwei Ebenen unter GF; ≤ -3 = mindestens drei Ebenen unter GF; Selbstst. = Selbstständige; *** p < .001, ** p < .01, * p < .05; ^{a} robuster F-Test nach Welch (1951) mit df = (4, 143); ^{b} robuster F-Test nach Welch (1951) mit df = (4, 142)

10.1.3 Diskussion

In dieser Untersuchung wurde die konkurrente Validität anhand mehrerer Kriterien begutachtet. Es wurden Zusammenhänge zwischen Wertesystemen mit Alter und Geschlecht, Studiengang und Studienschwerpunkt sowie Aufgabenbereich und Zugehörigkeit zur Hierarchieebene analysiert. Dabei haben sich einige signifikante Ergebnisse ergeben.

Bei Alter und Geschlecht waren die Ergebnisse weitestgehend hypothesenkonsistent, wobei die Unterschiede zwischen den Geschlechtern im Vergleich zur Meta-Analyse von Schwartz und Rubel (2005) deutlicher ausfallen. Dies kann als Hinweis gewertet werden, die Formulierungen der Items auf ihre Tonalität bzgl. Geschlechter einer Überprüfung zu unterziehen. Als weitergehende Untersuchung könnten z.B. die Maskulinitäts- und Feminitätsdimension nach Hofstede (1984) mit Wertesystemen in Bezug gesetzt werden (am besten in länderübergreifenden Stichproben), um dadurch zu überprüfen, ob diese Unterschiede der Geschlechter tatsächlich häufiger auftreten, oder ein Artefakt der deutschen Version des Fragebogens sind.

Bei den Untersuchungen zu charakteristischen Wertesystemen in Studiengängen und BW-Schwerpunkten wurden einige (auch über die Hypothesen hinausgehende) signifikante Effekte gefunden. Von den drei Hypothesen zu den Studiengängen konnten zwei vollständig und die Dritte zur Hälfte bestätigt werden. Weniger treffend waren die Hypothesen bei den BW-Schwerpunkten, obgleich sich auch hier unerwartete Effekte zeigten. Es wurden zwei der drei Hypothesen vollständig und eine dritte teilweise verworfen. Somit ist es scheinbar schwieriger, Schwerpunktgruppen anhand ihrer Wertesystempräferenzen zu beschreiben als ganze Studiengruppen. Dieses Ergebnis ist logisch, da die Stichproben der Schwerpunkte allesamt Studierende der Betriebswirtschaft waren.

Ein ähnliches Muster hat sich bei den Hypothesen zu den Aufgabenbereichen gezeigt. Von den zwölf Hypothesen wurden eine vollständig und vier teilweise angenommen sowie sieben Hypothesen abgelehnt. Darüber hinaus zeigten sich auch einige nicht erwartete Effekte. Dass weniger als die Hälfte der charakteristischen Wertesysteme von den Experten richtig eingeschätzt wurden, könnte einerseits ein Indiz dafür sein, dass die subjektive Einschätzung von typischen Wertesystemen schwierig ist und deshalb der Einsatz von Fragebögen wie dem MVSQ empfehlenswert ist, andererseits sind auch einige der untersuchten Stichproben relativ klein ($n < 30$), weswegen die vorliegenden Ergebnisse keinesfalls als repräsentativ für die Grundgesamtheit gesehen werden dürfen.

Bei den Hierarchieebenen trafen die Hypothesen weniger zu. Zwei der fünf Hypothesen wurden teilweise bestätigt und drei wurden abgelehnt. Es kann gesagt werden, dass es signifikante Unterschiede bzgl. einiger Wertesysteme zwischen den Hierarchieebenen gibt, diese jedoch tiefer gehender Untersuchungen bedürfen. Auch stellt sich die Frage, inwiefern die Charakteristika der Hierarchieebenen in unterschiedlich großen Unternehmen vergleichbar sind. Während zwei Ebenen unter der GF in einem Großkonzern bereits weit oben in der Hierarchie ist, gilt das nicht für kleine Unternehmen, die vielleicht insgesamt nur zwei oder drei

Hierarchieebenen haben. Eine differenzierte Abfrage im MVSQ wäre für diesen Zusammenhang wünschenswert.

Insgesamt kann also abgeleitet werden, dass es einige überzufällige Zusammenhänge zwischen charakteristischen Wertesystemen von Studiengängen, Studienschwerpunkten, Aufgabenbereichen und Hierarchieebenen gibt. Allein die Fülle der signifikanten Ergebnisse kann als Hinweis für die konkurrente Validität der Wertesysteme gewertet werden. Zwar bedürfen die Zusammenhänge einer gründlicheren Erforschung (mehr Studiengänge, größere und unabhängigere Stichproben), dennoch stellt dieser Ansatz eine vielversprechende Möglichkeit dar, um ein aufschlussreiches Hilfsmittel für Fragen der Studienwahl und Berufsorientierung zu sein. Auch in der Personalauswahl können Erkenntnisse zu Passung von Wertesystemen und Aufgabenbereichen eine nützliche Hilfestellung geben, um Fehlbesetzungen sowohl zugunsten des Bewerbers als auch der Firma zu reduzieren. Des Weiteren lässt sich die Fragestellung ableiten, welche Auswirkungen die Kongruenz von Wertesystemen in Aufgabenbereich und individuellen Wertesystempräferenzen mit sich bringt. Sind Personen mit passenden Profilen motivierter, zufriedener und bringen sie mehr Leistung? Das sind Fragen, die in weiterführenden Studien untersucht werden sollten.

Darüber hinaus kann zur Bipolaritätsfrage gesagt werden, dass es einige Unterschiede zwischen der Annäherungs- und Vermeidungsdimension gab, d.h. diese nicht durchgehend in entgegengesetzter Richtung auftraten. Zum Beispiel unterscheiden sich Vertriebler durch höhere Ausprägungen auf $\boldsymbol{Macht}^{A}$ und Logistiker durch niedrigere Ausprägungen auf $\boldsymbol{Macht}^{V}$. $\boldsymbol{Erfolg}^{V}$ zeigt signifikante Zusammenhänge mit Vertrieb und UF gleichermaßen, nicht aber $\boldsymbol{Erfolg}^{A}$. Letzteres Ergebnis könnte so gedeutet werden, dass Menschen eher dann in Vertriebs- oder UF-Jobs arbeiten, wenn sie ***Erfolg*** weniger ablehnen als der Rest, also die Vermeidungsmotivation mehr Einfluss darauf hat, welchen Job jemand eher bzw. zeitlich langfristiger hat. Auch hier besteht noch viel Forschungsbedarf.

Grenzen der Bedeutung dieser Studienergebnisse können darin gesehen werden, dass die ermittelten Effektstärken gemessen an der gängigen Faustregel (Cohen, 1988) als gering eingestuft werden müssen. Andererseits kann an dieser Stelle auf die Empfehlung von Thompson (2007) verwiesen werden, wonach der Vergleich mit ähnlichen Studien den Faustregeln vorzuziehen sei. Hierfür kann z.B. die Meta-Analyse von Verquer et al. (2003) herangezogen werden, in die größtenteils Studien zum Zusammenhang von Person-Organisation-Fit und Werten einfließen. Dort werden mittlere Effektstärken zwischen -.18 und .28 berichtet. Demnach liegen die hier ermittelten Effektstärken in durchaus erwartbaren Größenordnungen.

10.2 Studie zur prädiktiven Validität

Wie in der vorherigen Untersuchung zur konkurrenten Validität geschlussfolgert wurde, stellt sich die Frage, ob die Kongruenz zwischen Wertesystemen und Aufgabenbereich zu einer erhöhten Motivation führt. Diese Fragestellung war Ausgangspunkt dieser Studie. Dazu wurde ein *echtes* Experiment durchgeführt, in dem den Versuchspersonen drei verschiedene Aufgaben präsentiert wurden, die so entworfen wurden, dass sie kongruent mit unterschiedlichen Wertesystemen sein sollten. Je nach Passung zwischen Wertesystem und Aufgabe wurden dann erhöhte Ausmaße von Intensität der Motivation hypothetisiert. Da die Wertesysteme zeitlich vor der Motivationsintensität gemessen wurde, kann dies als Untersuchung der prädiktiven Validität der Wertesysteme beschrieben werden, denn ein Merkmal (hier Wertesystem) besitzt dann prädiktive Validität, wenn es ein Außenkriterium vorhersagen kann, das zeitlich *nach* dem Merkmal gemessen wurde (Groth-Marnat, 2003; Hartig et al., 2012).

An dieser Stelle ist angebracht, an die drei Dimensionen von Motivation (Richtung, Intensität und Persistenz, Kapitel 2.6) zu erinnern und den Bezug zu dieser Untersuchung herzustellen. In dieser Studie soll die Intensität der Motivation in Abhängigkeit der Richtung untersucht werden, wobei die Kernhypothese lautet: Je mehr die Richtung stimmt, d.h. die Aufgabe kongruenter zum Wertesystem ist, desto höher ist die Intensität der Motivation.

Zudem wurden die VPn in zwei Gruppen aufgeteilt, in denen als weiteren Faktor der Einfluss eines monetären Anreizes untersucht werden sollte. Die zugehörige Hypothese lautete, dass das Motivationsempfinden positiv mit dem Wertesystem $\boldsymbol{Erfolg}^{A}$ und negativ mit $\boldsymbol{Erfolg}^{V}$ zusammenhängt.

10.2.1 Hintergrund

Im theoretischen Teil werden die Grundlagen für die Auswahl und Konstruktion der Experimentalunterlagen gelegt. Die Ergebnisse aus der Untersuchung zur konkurrenten Validität (Kapitel 10.1) haben gezeigt, dass es für das Wertesystem ***Gewissheit*** besonders bezeichnend ist, nicht in leitenden Tätigkeiten vorzukommen (weder Unternehmensführung noch Abteilungsleitung). Nachdem es sich dabei wiederum um Positionen handelt, für die vor allem das Fällen von Entscheidungen (Rajagopalan & Datfa, 1996) charakteristisch ist, kann im Umkehrschluss gefolgert werden, dass Aufgabenbereiche, die eine gute Passung zum ***Gewissheit***-Wertesystem haben, wenig Raum für Entscheidungen brauchen. Diese Beschreibung ist auch mit den dieses Wertesystem kennzeichnenden Werten wie Regeltreue und Disziplin stimmig, denn bei der Einhaltung von Regeln und der disziplinierten Abarbeitung von Aufgaben nach vorgegebenen Schemata ist es nicht erforderlich, Entscheidungen zu treffen. Aufgaben, die mit ***Gewissheit*** kongruent sind, sollten demnach wenig Spielraum für eigene Entscheidungen lassen. Aufgaben mit wenig Entscheidungsspielraum erfordern klare Vorgaben der Bearbeitungsweise und haben aufgrund dessen vermutlich eher repetitiven Charakter.

Außerdem hat diese Untersuchung ergeben, dass ***Erfolg*** typisch für Studierende der Betriebswirtschaft war und die Vermeidung davon niedrig bei Personen mit Aufgaben wie Verkauf oder Unternehmensführung ausgeprägt war. Das Fällen von Entscheidungen entspricht Menschen mit niedrigen Ausprägungen auf $\boldsymbol{Erfolg}^{V}$ dementsprechend mehr. Für Erfolg in Vertriebsaufgaben sind außerdem ein hohes Maß an Flexibilität und Zielorientierung erforderlich (Dubinsky et al., 1997). Diese Beschreibung kann als kongruent zur Definition des Wertesystems ***Erfolg*** mit den zentralen Werten Ergebnisorientierung, Pragmatismus und Wettbewerb gesehen werden. Aufgaben, die wirtschaftliches Denken und Handeln abbilden, spiegeln dieses Wertesystem somit wieder.

Das ***Gleichheit***-Wertesystem trat verstärkt im Personalbereich auf, sowohl unter den Studierenden, als auch in der Berufstätigenstichprobe. Das Merkmal dieses Wertesystems, das vermutlich am besten zu diesen Aufgabenbereichen passt, ist die Orientierung hin zum Zwischenmenschlichen. Aufgaben, die durch das ***Gleichheit***-Wertesystem charakterisiert werden können, haben demnach damit zu tun, Konsens und Harmonie in Gruppen zu erreichen. Dafür erscheinen Gruppenarbeiten wesentlich passender als Aufgaben, die alleine bearbeitet werden.

Für das Wertesystem ***Verstehen*** waren eindeutig forschende Tätigkeiten charakteristisch und für diese stellt Innovativität eine der wichtigsten Eigenschaften dar (Elkins & Keller, 2003; Judge et al., 1997). Innovativität wiederum ist einer der Hauptwerte des ***Verstehen***-Wertesystems. Brainstorming oder konzeptuelle Denkarbeit sind Tätigkeiten, die in besonderem Maße zu diesem Wertesystem passen, da sie einerseits Freiraum für Ideen, also Innovativität lassen und andererseits erfordern, sich tiefgehend mit einer Materie zu befassen, sie also zu erforschen.

Beim Wertesystem ***Geborgenheit*** gingen tendenziell ähnlich wie bei ***Gewissheit*** Aufgaben der Unternehmensführung mit niedrigen Ausprägungen auf der Annäherungs- und hohen auf der Vermeidungsdimension im Vergleich zu einigen anderen Aufgabenbereichen einher. Auch bei ***Macht*** und ***Nachhaltigkeit*** konnten nur Unterschiede zwischen wenigen der untersuchten Gruppen festgestellt werden.

Da Anreizsysteme häufig auf monetärer Basis operieren und Geld eine bedeutende Rolle als Incentive darstellt (Latham, 2012), liegt es nahe einen monetären Anreiz als manipulierbare Variable in das Experiment aufzunehmen. Abgesehen davon herrschen in der Forschungsliteratur gegensätzliche Ansichten. Während manche Autoren zu dem Schluss kommen, dass Geld als Anreiz eher förderlich auf Motivation wirkt (Lea & Webley, 2006), konstatieren andere eine hemmende Wirkung (Furnham, 2014). Dieser Gegensatz könnte mithilfe der Wertesysteme so erklärt werden, dass ein monetärer Anreiz je nach Wertesystempräferenz entweder positiv oder negativ mit Motivation interagiert. Auch diese Kontroverse stellt in der vorliegenden eine Untersuchung eine Motivation dar, eine monetäre Anreiz-Variable aufzunehmen.

Um Motivation zu messen, bieten sich unterschiedliche Verfahren an, je nachdem welcher Aspekt der Motivation gemessen werden soll (Rheinberg & Vollmeyer, 2012). Wie in Kapitel 2.6 dargestellt, geht es bei der Intensität der Motivation um das gefühlte Erleben bei der Ausführung einer Tätigkeit. Dementsprechend sollte sich auch das Messinstrument direkt auf die Qualität des Empfindens bei der Tätigkeitsausführung beziehen. Flow, positiver und negativer Affekt und intrinsische Motivation (im Sinne von Deci & Ryan, 1985) sind Konzepte, die diese Bedingung erfüllen (Rheinberg, 2004).

10.2.2 Methode

An dieser Studie nahmen 166 Studierende teil, die in Kapitel 3.7 als Stichprobe IIIb beschrieben wurden. Fünf weitere Personen wurden von der Untersuchung ausgeschlossen, da sie die Materialien nicht vollständig bearbeitet hatten. Die Teilnehmer wurden in verschiedenen Lehrveranstaltungen der Fakultäten Betriebswirtschaft, Allgemeinwissenschaften und Mikrosystemtechnik, Bauingenieurwesen, Elektro-und Informationstechnik, Informatik und Mathematik und Angewandte Sozial- und Gesundheitswissenschaften der OTH Regensburg sowie Lehrveranstaltungen der Institute für Psychologie, Information und Medien, Sprache und Kultur sowie Kunsterziehung der Universität Regensburg angeworben. Alle Teilnehmer erhielten die Möglichkeit zur individuellen Testergebnisrückmeldung in Form eines mündlichen Vortrags und eines schriftlichen Ergebnisberichts. Studierende der Psychologie hatten zudem die Möglichkeit, zwei Versuchspersonen-Stunden zu erhalten. Der Anteil Psychologie-Studierender war allerdings relativ gering, da vornehmlich in anderen Fakultäten geworben wurde.

10.2.2.1 Material

Das Material bestand aus drei Aufgaben und je mehreren Fragen, die die Intensität der Motivation messen sollten. Die Konstruktion der Aufgaben wurde auf allgemeiner Ebene von folgenden Kriterien geleitet:

1. Die Beabreitungszeit sollte bei allen Aufgaben ungefähr gleich lang sein und insgesamt für alle drei Aufgaben eine Stunde nicht überschreiten.
2. Die Aufgaben sollten individuell, d.h. alleine bearbeitet werden können.
3. Die Aufgaben sollten vergleichbar in ihrer Schwierigkeit sein.
4. Die Aufgaben sollten je kongruent zu einem Wertesystem sein.

Um auszuschließen, dass Ermüdungseffekte Einfluss auf die Untersuchungsergebnisse haben, wurde die Zahl der Aufgaben auf drei festgelegt, für die jeweils 15 Minuten zur Bearbeitung zur Verfügung standen. Inklusive der Bearbeitung der Fragen zur Intensität der

Motivation (maximal 4 Minuten pro Aufgabe) konnte somit das zuvor formulierte Kriterium der Gesamtbearbeitungsdauer von einer Stunde eingehalten werden.

Die drei Aufgaben wurden auf Basis der in Abschnitt 10.2.1 aufgeführten empirischen Ergebnisse und theoretischen Überlegungen hin konstruiert. Dabei wurde je eine Aufgabe für eines der Wertesysteme ***Gewissheit***, ***Erfolg*** und ***Verstehen*** erstellt. ***Gleichheit*** hat sich als wenig geeignet für das Vorhaben dieser Untersuchung gezeigt, denn es passt weniger gut zur Einzelbearbeitung. Sondern eine Gruppenarbeit wäre besser geeignet, um eine zu diesem Wertesystem kongruente Aufgabe darzustellen. Für die Wertesysteme ***Geborgenheit***, ***Macht*** und ***Nachhaltigkeit*** war die empirische Basis weniger eindeutig, weshalb die Wahl der untersuchten Wertesysteme auf die genannten fiel.

Als übergreifendes Setting wurde ein Studentenjob im (erfundenen) „Café am Campus“ gewählt. Die Studierenden sollten sich vorstellen, dass sie in besagtem Café arbeiten und Aufgaben aus verschiedenen Bereichen zu bearbeiten haben. Insgesamt firmierte die Studie unter dem Namen „Regensburger Aufgabenstudie“. Im Folgenden werden die Aufgaben kurz beschrieben, Anhang C enthält die vollständigen Unterlagen. Die Aufgaben wurden zudem in zwei Pretests auf ihre Brauchbarkeit (Verständlichkeit und Bearbeitungsdauer) untersucht (Diekmann, 2010). Im ersten Pretest wurden dazu mit zwei Personen die Methode des *lauten Denkens* gewählt. Im zweiten Pretest haben zwölf Personen die Unterlagen bearbeitet und wurden retrospektiv befragt. Dabei wurden sowohl die Verständlichkeit der Unterlagen bestätigt, als auch das Erkennbarkeit der Kongruenz der Aufgaben mit unterschiedlichen Wertesystemen verbal geäußert.

Sachbearbeitungsaufgabe (kongruent mit *Gewissheit*)

Bei der Sachbearbeitungsaufgabe sollten in 50 vordefinierte Felder auf zwei DIN A4 Blätter jeweils exakt derselbe Text geschrieben werden. Die ausgeschnittenen Felder sollten für eine Werbe-Aktion als Flyer verwendet werden, so die Begründung der Aufgabe. Der Text wurde nicht nur in Wortlaut („Café am Campus Croissants für 50 Cent“), sondern auch Form und Zeilenumbruch vorgegeben (siehe Anhang C). Es bestand überhaupt kein Spielraum, um eigene Entscheidungen zu treffen. Die Aufgabe erforderte somit Disziplin und striktes Ausführen nach Vorgabe und sollte demnach kongruent mit dem ***Gewissheit***-Wertesystem sein.

Gewinnmaximierungsaufgabe (kongruent mit *Erfolg*)

Diese Aufgabe beinhaltete eine Serie von 15 Entscheidungssituationen, in denen in Anlehnung an Kahneman und Tversky (1979) Situationen skizziert wurden, in denen zwischen zwei oder drei unsicheren Antwortmöglichkeiten abgewägt werden musste, welche Option den größten finanziellen Gewinn bringt. Dabei wurde ein pragmatisches Vorgehen gefordert, da keine

Angaben darüber gemacht wurden, wie die Entscheidungen getroffen werden sollten, sondern nur die Zielvorgabe ausgegeben wurden, den Gewinn durch die Entscheidungen zu maximieren.

Konzeptentwicklungsaufgabe (kongruent mit *Verstehen*)

Bei dieser Aufgabe waren die Versuchspersonen aufgefordert, ein Konzept für ein gemeinsames Studierendenhaus von Universität und OTH Regensburg mit integrierter Gastronomie zu erarbeiten. Es wurden viele Anregungen in Form von Fragen und wenig Einschränkungen gegeben, sowie viel Raum für eigene Ideen gelassen. Die Aufgabe kann damit einerseits als theoretisch und andererseits als komplex bezeichnet werden. Sie entspricht damit den Werten des Wertesystems ***Verstehen***.

Monetärer Anreiz

Die Unterlagen wurden in zwei Gruppen mit unterschiedlichen experimentellen Bedingungen aufgeteilt (mit und ohne monetärem Anreiz). Monetärer Anreiz wurde dabei durch die Ausschreibung eines Gewinns operationalisiert. Die Gruppe ohne Anreiz erhielt vor der Bearbeitung der ersten Aufgabe den Hinweis, dass es sich bei den Aufgaben *nicht* um einen Leistungstest handelt und es in der Untersuchung lediglich darum geht, die Aufgaben untereinander zu vergleichen. Die Gruppe mit Anreiz erhielten stattdessen den Hinweis, dass die drei besten Teilnehmer einen Geldgewinn erhalten (50, 25 und 15 Euro) und dass die Gewinner anhand von Leistungspunkten ermittelt werden, die bei der Sachbearbeitungsaufgabe die Anzahl der Fehler, bei der Gewinnmaximierung die Höhe des ermittelten Gewinns und bei der Konzeptentwicklungsaufgabe die Anzahl der Ideen waren. Bei allen drei Aufgaben erfolgte in der Anreiz-Gruppe am Ende der Anleitung zudem der Hinweis auf das jeweilige Leistungsmaß.

Flow-Kurzskala

Bei der Flow-Kurzskala (FKS) handelt es sich um ein deutschsprachiges Instrument, das entwickelt wurde, um das Flow-Erleben zu messen (Rheinberg et al., 2003). Es besteht aus insgesamt 16 Items, von denen zehn Items dafür verwendet werden, die beiden Flow-Faktoren Absorbiertheit und glatter Verlauf mit vier bzw. sechs Items zu erfassen. Zusammengenommen ergeben diese zehn Items einen Wert für Flow. Darüber hinaus enthält die Skala drei Items zur Messung einer Besorgniskomponente. Die Items dieser drei Faktoren wurden jeweils in einer fünfstufigen Rating-Skala präsentiert, wobei die Stufen mit „trifft gar nicht zu“, „trifft wenig zu“, „trifft etwas zu“, „trifft ziemlich zu“ und „trifft voll und ganz zu“ markiert wurden. Zum Schluss enthält die Skala noch drei Items, mit denen mit je einem Item die Einschätzung der eigenen Fähigkeit, Anforderung sowie Passung zwischen Fähigkeit und Anforderung abgefragt wird. Diese wurden auf sieben Stufen erfasst. Hintergrund dazu ist, dass festgestellt wurde, dass

Flow dann maximal wird, wenn die Anforderungen der Situation zu den eigenen Fähigkeiten passen (oder diese leicht überschreiten; Csikszentmihalyi, 1975; Rheinberg, 2010).

Nach Rheinberg et al. (2003) wurden in mehreren Untersuchungen Cronbach's αs von um die .90 festgestellt, die dem Instrument somit eine hohe Messgenauigkeit bescheinigen. Auch die Validität kann als gegeben angesehen werden, zum einen der kurvilineare Zusammenhang zwischen Flow und der Passung von Anforderung und Fähigkeiten repliziert werden konnte und andererseits hypothesenkonsistente Zusammenhänge zwischen Flow und Lernleistung sowie Erfolgszuversicht und Misserfolgsfurcht gezeigt werden konnten (Rheinberg et al., 2003).

Kurzskala intrinsischer Motivation

Die Kurzskala intrinsischer Motivation (KIM) ist eine adaptierte, zeitökonomische Version des Intrinsic Motivation Inventory von Deci und Ryan[1] (Wilde et al., 2009). Sie besteht aus vier Faktoren, von denen die zwei Faktoren *Vergnügen* und *wahrgenommene Kompetenz* in dieser Untersuchung verwendet wurden. Die anderen beiden Faktoren *wahrgenommene Wahlfreiheit* und *Druck/Anspannung* wurden ausgelassen, da sie einerseits nicht zutrafen (es gab keine Wahlfreiheit in diesem Experiment) und andererseits die interne Konsistenz des Faktors Druck/Anspannung um .50 relativ gering war und diese Dimension in der geplanten Untersuchung eine untergeordnete Rolle spielt. Die beiden hier verwendeten Konstrukte wurden jeweils mit drei Items auf einer fünfstufigen Likert-Skala mit den Stufen „trifft gar nicht zu", „trifft wenig zu", „trifft etwas zu", „trifft ziemlich zu" und „trifft voll und ganz zu" gemessen. Für diese beiden Faktoren stellten Wilde et al. (2009) gute interne Konsistenzen fest (α in mehreren Untersuchungen zwischen .79 und .89) und konnten Hinweise auf die Validität der Skala durch eine Hauptkomponentenanalyse im Rahmen einer empirischen Untersuchung ermitteln.

Kurzskalen zur Erfassung der Positiven Aktivierung, Negativen Aktivierung und Valenz

Diese Skala (PANAVA-KS) wurde entwickelt, um die drei Erlebensdimensionen positive Aktivierung (PA), negative Aktivierung (NA) und Valenz (VA) zu erfassen (Schallberger, 2005). Sie setzt sich aus insgesamt 10 bipolaren Items zusammen, die auf einer siebenstufigen Rating-Skala die Stimmung der VP abfragen. Dabei erfassen jeweils vier Items PA und NA sowie zwei Items VA. Die internen Konsistenzen der Skala lagen in einer Untersuchung mit N = 269 Personen bei .83 für PA, .76 für NA und .74 für VA und befinden sich somit in einer zufriedenstellenden Größenordnung. Untersuchungen zur Konstruktvalidität konnten bestätigen, dass Messungen der PANAVA-KS mit dem Originalinstrument PANAS (Krohne et al., 1996; Watson et al., 1988) weitestgehend übereinstimmen.

[1] Für eine Studie zur Validierung des Originalinstruments siehe McAuley et al. (1989).

10.2.2.2 Versuchsablauf und -design

Bei der Studie handelte es sich um ein 2 (Anreiz: ja vs. nein) x 1 (Aufgabe)-faktorielles Design mit Messwiederholung auf dem letzten Faktor, wobei bei jeder Messwiederholung eine andere Aufgabe zu bearbeiten war und die Reihenfolge der Aufgaben vollständig randomisiert wurde. Als abhängige Variablen (AV) wurden Flow und Besorgnis (BE) aus der FKS-Skala, Vergnügen (V) und wahrgenommene Kompetenz (WK) aus dem KIM sowie die drei Faktoren positive Aktivierung (PA), negative Aktivierung (NA) und Valenz (VA) aus dem PANAVA-KS erhoben. Die Zusammenhänge zwischen den Wertesystemen und AVs bei einer Aufgabe wurden in Korrelations- oder Regressionsanalysen ermittelt, wobei jede Analyse bei jeder Aufgabe durchgeführt wurde.

Die entsprechenden Hypothesen lauten für alle drei Wertesystem-Aufgaben-Kombinationen für die Gruppe der VPn ohne monetären Anreiz wie folgt: Der Zusammenhang zwischen Annäherungswertesystem und den AVs Flow, V, WK, PA und VA ist in positiver Richtung und mit den AVs BE und NA in negativer Richtung. Bei den Vermeidungswertesystemen werden die Zusammenhänge in entgegengesetzter Richtung erwartet.

In der Gruppe mit monetärem Anreiz gilt zudem, dass die Zusammenhänge mit $\boldsymbol{Erfolg}^{A}$ bzw. $\boldsymbol{Erfolg}^{V}$ bei der GM-Aufgabe jeweils in der hypothetisierten Richtung stärker ausfallen. Bei den anderen Aufgaben sollen die Zusammenhänge zwischen den AVs und Wertesystemen aufgrund fehlender Vergleichsliteratur explorativ betrachtet werden.

10.2.3 Ergebnisse Voruntersuchung

In einer Voruntersuchung wurden die internen Konsistenzen der verwendeten Instrumente zur Messung des Motivationserlebens berechnet. Da Flow, KIM und PANAVA bei jeder Aufgabe einzeln erhoben wurden, konnten für jede Aufgabe gesondert die Cronbachs α berechnet werden. Tabelle 57 zeigt die entsprechenden Werte, die alle im akzeptablen bis guten und sehr guten Bereich liegen (M $= .83$, Spanne von .72 bis .93). Im Vergleich der Aufgaben war die Messgenauigkeit bei der Sachbearbeitungsaufgabe mit $\alpha = .80$ im Schnitt etwas niedriger als bei der Gewinnmaximierungs- ($\alpha = .85$) und Konzeptentwicklungsaufgabe ($\alpha = .84$). Insgesamt waren die Instrumente somit geeignet, um die abhängigen Variablen dieser Untersuchung verlässlich zu erheben.

10.2.4 Ergebnisse Hauptuntersuchung

Um die Aufgaben miteinander zu vergleichen, wurden zunächst die Einschätzungen der Schwierigkeit der Aufgaben (Tabelle 58), die Einschätzungen der eigenen Fähigkeit der VPn (Tabelle 59) und die Einschätzungen der Höhe der Anforderungen der Aufgaben an die VPn (Tabelle 60) berechnet.

Tabelle 57. Interne Konsistenzen (Cronbachs α) der Skalen zur Messung der Intensität der Motivation pro Aufgabe.

	Aufgabe		
Skala	SB	GM	KE
Flow	.83	.86	.91
Flow BE	.72	.82	.77
KIM V	.80	.89	.88
KIM WK	.93	.91	.91
PA	.83	.87	.87
NA	.77	.84	.80
VA	.76	.77	.75

Anmerkung. BE = Besorgnis; V = Vergnügen; WK = Wahrgenommene Kompetenz; PA = Positive Aktivierung; NA = Negative Aktivierung; VA = Valenz; SB = Sachbearbeitung; GM = Gewinnmaximierung; KE = Konzeptentwicklung.

Tabelle 58. Mittelwerte und Standardabweichungen der Einschätzungen der Schwierigkeit der Aufgaben in Abhängigkeit der experimentellen Bedingung.

	Aufgabe					
	SB		GM		KE	
Experimentelle Bedingung	M	SD	M	SD	M	SD
ohne Anreiz (N = 80)	1.80	1.48	4.26	1.49	3.70	1.58
mit Anreiz (N = 86)	1.79	1.44	4.52	1.63	3.74	1.43

Anmerkung. SB = Sachbearbeitung; GM = Gewinnmaximierung; KE = Konzeptentwicklung.

Für jede dieser drei Einschätzungen wurde eine 2 (Anreiz) x 1 (Aufgabe)-faktorielle ANOVA mit Messwiederholung auf dem letzten Faktor gerechnet. Diese ergaben bei keiner der Aufgaben einen signifikanten Haupteffekt bzgl. der experimentellen Bedingung des Anreizes, allerdings signifikante Unterschiede zwischen den Aufgaben bei allen drei Eigenschaften. Bzgl. der Schwierigkeit ergaben ANOVA ($F(2, 330) = 141.97$, $p =< .001$, $\eta^2 = .35$) und paarweise t-Tests mit verbundener Stichprobe und Bonferroni-Korrektur hochsignifikante Unterschiede ($p < .001$) zwischen allen drei Aufgaben. Die Gewinnmaximierungsaufgabe wurde dabei am schwierigsten und die Sachbearbeitungsaufgabe als am leichtesten empfunden. Zur Überprüfung der Varianzhomogenität wurde Maulchy's Test auf Sphärizität (Mauchly, 1940) mit

Tabelle 59. Mittelwerte und Standardabweichungen der Einschätzungen der eigenen Fähigkeiten in Abhängigkeit der experimentellen Bedingung.

	Aufgabe					
	SB		GM		KE	
Experimentelle Bedingung	M	SD	M	SD	M	SD
ohne Anreiz (N = 80)	4.80	1.98	3.85	1.46	4.46	1.41
mit Anreiz (N = 86)	4.53	1.86	3.64	1.56	4.17	1.56

Anmerkung. SB = Sachbearbeitung; GM = Gewinnmaximierung; KE = Konzeptentwicklung.

Tabelle 60. Mittelwerte und Standardabweichungen der Einschätzungen der Anforderungen der Aufgaben in Abhängigkeit der experimentellen Bedingung.

	Aufgabe					
	SB		GM		KE	
Experimentelle Bedingung	M	SD	M	SD	M	SD
ohne Anreiz (N = 80)	1.66	1.20	4.22	1.19	4.10	1.06
mit Anreiz (N = 86)	1.70	1.14	4.52	1.17	4.20	0.70

Anmerkung. SB = Sachbearbeitung; GM = Gewinnmaximierung; KE = Konzeptentwicklung.

dem R-Paket `ez` (Lawrence, 2015) berechnet. Dieser wurde nicht signifikant und zeigte somit Varianzhomogenität an.

Bei der Einschätzung der eigenen Fähigkeiten ($F(2, 330) = 12.89$, $p =< .001$, $\eta^2 = .05$) sind sich die Mittelwerte zwar deutlich ähnlicher, doch auch hier wurden die Unterschiede signifikant. Da Maulchy's Test auf Sphärizität signifikant wurde ($p < .01$) und somit Heterogenität der Varianzen anzeigte, wurde die Geisser-Greenhouse-Korrektur der p-Werte berechnet. Auch unter Berücksichtigung des korrigierten p-Werts wurde der Haupteffekt des Faktors Aufgabe signifikant ($p < .001$). Interessanterweise wurden nicht die eigenen Fähigkeiten bei der besonders leichten Sachbearbeitungsaufgabe signifikant höher eingeschätzt, sondern die Fähigkeiten waren bei der GM-Aufgabe signifikant niedriger als bei der KE- ($p < .01$) und der SB-Aufgabe ($p < .001$).

Bei den Anforderungen der Aufgaben ($F(2, 330) = 365.97$, $p =< .001$, $\eta^2 = .56$) wurde die SB-Aufgabe als besonders niedrig eingestuft. Die entsprechenden p-Werte sind jeweils im Vergleich zu GM- und KE-Aufgabe $< .001$. Bei diesen p-Werten handelt es sich um Geisser-

Greenhouse-korrigierte p-Werte, da wie schon bei der Varianzanalyse der Fähigkeiten, die Varianzen nach Maulchy's Test auf Sphärizität heterogen waren ($p < .05$).

Insgesamt muss damit gesagt werden, dass die Sachbearbeitungsaufgabe besonders niedrige Anforderungen stellt und einfach ist. Demnach ist sie nur bedingt mit den beiden anderen Aufgaben vergleichbar. Diese Befunde können auch dahingehend gewertet werden, dass die SB-Aufgabe nicht die Bedingung für Flow erfüllt, nach der die subjektive Einschätzung der Passung von Fähigkeiten und Anforderungen gegeben sein muss (Csikszentmihalyi, 1975).

Als nächstes wurden Korrelationen zwischen den AVs und Wertesystemen bei den jeweils kongruenten Aufgaben berechnet. Tabelle 61 zeigt die Zusammenhänge der AVs mit ***Gewissheit*** bei der SB-Aufgabe. Dabei muss festgestellt werden, dass auf der Annäherungsdimension kein einziger Zusammenhang signifikant wurde und die Korrelationskoeffizienten dementsprechend niedrig ausfielen. Insbesondere für Flow als AV ist dieses Ergebnis wenig überraschend, da die Bedingung der Passung von Anforderungen und Fähigkeiten nicht erfüllt war. Auf der Vermeidungsdimension waren die Koeffizienten zwar etwas höher, dennoch wurde nur die Korrelation zwischen Valenz und ***Gewissheit***^{V} in der Gruppe mit Anreiz signifikant. Darüber hinaus kann allerdings beobachtet werden, dass mit Ausnahme von Flow und Vergnügen (deren Korrelationen bei bei $r = 0$ liegen) die Korrelationen zwischen den AVs und ***Gewissheit***^{A} sowie ***Gewissheit***^{V} in beiden experimentellen Bedingungen gleichermaßen in die hypothetisierte Richtung deuten. Somit kann gesagt werden, dass die Befunde über den Zusammenhang zwischen Valenz und ***Gewissheit***^{V} hinaus zumindest ansatzweise für die Gültigkeit der Hypothesen sprechen. Abgesehen davon wurden keine weiteren Effekte gefunden.

Tabelle 61. Korrelationen zwischen ***Gewissheit*** und AVs bei der Sachbearbeitungsaufgabe.

	Annäherung		Vermeidung	
	ohne A	mit A	ohne A	mit A
Flow	.00	-.02	-.11	-.12
Flow BE	-.08	.08	.06	-.05
KIM V	.00	.08	-.07	-.13
Kim WK	.10	-.09	-.11	.02
PA	.13	.01	-.19	-.11
NA	-.08	.02	.08	.15
VA	.09	.07	-.14	-.24*

Anmerkung. BE = Besorgnis; V = Vergnügen; WK = Wahrgenommene Kompetenz; PA = Positive Aktivierung; NA = negative Aktivierung; VA = Valenz; A = Anreiz; * p < .05.

Bei den Korrelationen der AVs mit ***Erfolg*** bei der GM-Aufgabe (Tabelle 62) zeigten sich keine signifikanten Korrelationen in der Gruppe ohne monetären Anreiz. In der Gruppe mit monetärem Anreiz wurden jedoch mehrere Korrelationen signifikant. Zum einen korrelierte der Besorgnisfaktor von Flow positiv ($r = .23$) und auch die positive Aktivierung (PA) ($r = .27$) positiv mit ***Erfolg***^{A} (beide zum .05 Niveau). Auf der Vermeidungsdimension korrelierten zum .05-Niveau Besorgnis mit $r = -.26$, Vergnügen mit $r = -.22$ und positive Aktivierung mit $r = -.25$ mit ***Erfolg***^{V}. Der Zusammenhang von Wertesystem und einigen AVs, die Motivationsempfinden indizieren, wurde also nur in der Gruppe mit Anreiz signifikant.

Tabelle 62. Korrelationen zwischen ***Erfolg*** und AVs bei der Gewinnmaximierungsaufgabe.

	Annäherung		Vermeidung	
	ohne A	mit A	ohne A	mit A
Flow	.13	.11	-.08	-.17
Flow BE	.06	.23*	-.12	-.26*
KIM V	.01	.17	-.02	-.22*
Kim WK	-.02	.04	-.08	-.16
PA	.07	.27*	-.07	-.25*
NA	-.13	.01	.04	.02
VA	.16	.21	-.08	-.14

Anmerkung. BE = Besorgnis; V = Vergnügen; WK = Wahrgenommene Kompetenz; PA = Positive Aktivierung; NA = negative Aktivierung; VA = Valenz; A = Anreiz; * p < .05.

Da die Unterschiede der Korrelationen zwischen den Gruppen relativ deutlich waren, kann in diesem Zusammenhang eine Interaktion, genauer gesagt ein Moderatoreffekt von Anreiz vermutet werden (Baron & Kenny, 1986). Um dies zu überprüfen, wurden mehrere Regressionsmodelle mit dem entsprechenden Wertesystem, Anreiz und dem Wertesystem $\times$ Anreiz-Interaktionsterm als Prädiktoren geschätzt (Jaccard & Turrisi, 2003). Zwar wurde in keinem der Modelle ein signifikanter Interaktionseffekt zwischen Wertesystem und Anreiz festgestellt, allerdings zeigten sich mehrere signifikante Haupteffekte unter Kontrolle des Anreizes. Für die Überprüfung der Haupteffekte wurden mehrere multiple Regressionsmodelle je mit Wertesystem und Anreiz als Prädiktoren ohne Interaktionsterm gerechnet (Crawford et al., 2014). Dabei wurden folgende Haupteffekte signifikant: von ***Erfolg***^{A} auf PA (B = 0.51, $t(162) = 2.25$, $p < .05$, $F(2, 162) = 3.04$, $p = .05$, $R^2 = .04$) und auf VA (B = 0.6, $t(160) = 2.34$, $p < .05$, $F(2, 160) = 3.2$, $p < .05$, $R^2 = .04$) sowie von ***Erfolg***^{V} auf Flow Besorgnis, (B = -0.44, $t(163) = 2.43$, $p < .05$, $F(2, 163) = 3.03$, $p = .05$, $R^2 = .04$), auf PA (B = -0.46, $t(162) = 2.1$, $p < .05$, $F(2, 162) = 2.72$, $p = .07$, $R^2 = .03$). Zu den eben berichteten Haupteffekten ist zudem zu sagen,

dass die Modellanpassungen gering waren, mit teilweise $p > .05$. Dies ging darauf zurück, dass in allen Modellen Anreiz als Prädiktor enthalten war, der jedoch in keinem Modell signifikant wurde (siehe auch Tabellen 58 bis 60 jeweils GM-Aufgabe) und dadurch die Gesamtpassung der Modelle verschlechterte, obwohl das Wertesystem je einen signifikanten Effekt zeigte.

Die Voraussetzungen für lineare Regression waren für alle untersuchten Zusammenhänge erfüllt. Weder die Durbin-Watson-Tests auf Autokorrelation, noch die Breusch-Pagan-Tests auf Homoskedastizität wurden signifikant und auch Multikollinearität war nicht vorhanden (alle VIFs nahe 1). In Summe kann somit gesagt werden, dass die Wertesysteme ***Erfolg***^{A} und ***Erfolg***^{V} ungeachtet der Anreiz-Bedingung einige der abhängigen Variablen beeinflussen, wenngleich die Hypothese abgelehnt werden muss, dass Anreiz in diesem Zusammenhang als Moderator wirkt.

Tabelle 63. Korrelationen zwischen ***Verstehen*** und AVs bei der Konzeptentwicklungsaufgabe.

	Annäherung		Vermeidung	
	ohne A	mit A	ohne A	mit A
Flow	.20	-.01	-.29**	.16
Flow BE	.13	-.06	-.16	.07
KIM V	.24*	-.06	-.27*	.22*
Kim WK	.22*	-.06	-.27*	.18
PA	.12	-.01	-.22	.13
NA	-.09	.02	.13	-.16
VA	.14	-.08	-.19	.22*

Anmerkung. BE = Besorgnis; V = Vergnügen; WK = Wahrgenommene Kompetenz; PA = Positive Aktivierung; NA = negative Aktivierung; VA = Valenz; A = Anreiz; ** p < .01, * p < .05.

Tabelle 63 enthält die Korrelationen der AVs mit ***Verstehen*** bei der KE-Aufgabe. Hierbei wurden im Vergleich die meisten Korrelationen signifikant. In der Gruppe ohne Anreiz hängt ***Verstehen***^{A} positiv mit Vergnügen ($r = .24$) und wahrgenommener Kompetenz ($r = .22$) zusammen ($p < .05$). Zudem unterscheiden sich die Korrelationen in der Gruppe mit Anreiz insofern von den Korrelationen in der Gruppe ohne Anreiz, dass sie ausnahmslos ein gegenteiliges Vorzeichen haben, wenngleich keine der Korrelationen signifikant wurde. Auf der Vermeidungsdimension korreliert ***Verstehen***^{V} in der Gruppe ohne Anreiz signifikant mit Flow ($r = -.29$, $p < .01$), Vergnügen und wahrgenommener Kompetenz (beide $r = -.27$, $p < .05$) und in der Gruppe mit Anreiz mit Vergnügen ($r = .22$) und Valenz ($r = .22$, beide $p < .05$). Vor allem im Zusammenwirken mit der Vermeidungsdimension von ***Verstehen*** scheint Anreiz einen großen Einfluss auf das Motivationsempfinden zu haben. Bei allen AVs liegen die Vorzei-

chen in umgekehrter Richtung in den beiden Gruppen vor. Deshalb wurden erneut mehrere Regressionsmodelle berechnet, um auf Moderatoreffekte von Anreiz zu testen. Zudem wurden Modelle ohne Interaktionsterm spezifiziert, um nach Haupteffekten zu suchen.

In allen Regressionsmodellen mit ***Verstehen***^{A} und Anreiz als Prädiktoren und ohne Interaktionsterm zeigte lediglich das Modell für wahrgenommene Kompetenz (KIM) als AV einen Haupteffekt (B = 0.34, $t(162) = 2.04$, $p < .05$), wobei die Modellanpassung ungenügend war ($F(2, 162) = 2.2$, $p = .11$, $R^2 = .03$). Auch hier liegt die Ursache dafür am nicht-signifikanten Prädiktor Anreiz.

Tabelle 64. Lineare Regression mit Interaktionseffekt zwischen ***Verstehen***^{A} und Anreiz bei Konzeptentwicklungsaufgabe auf Vergnügen.

Prädiktor	B	SE	t	p
$(Konstante)$	2.83	0.1	29.53	<.001
Verstehen^A	0.51	0.21	2.49	<.05
Anreiz	0.05	0.13	0.38	.70
Verstehen^A $\times$ Anreiz	-0.62	0.3	2.03	<.05
R^2	.04			
ΔR^2	.02			
F-Test	$F(3, 162) = 2.2$, $p = .09$			

Anmerkung. N = 166; B = Regressionsgewicht; SE = Standardfehler; A = Annäherung; ΔR^2 = Differenz zum Modell ohne Interaktionsterm.

Allerdings wurden die Interaktionsterme für mehrere Modelle signifikant. Diese sind in den Tabellen 64 bis 67 dargestellt und zeigen, dass Anreiz einen moderierenden Effekt im Zusammenhang zwischen ***Verstehen***^{A} und Vergnügen sowie ***Verstehen***^{V} und Flow, Vergnügen und Positive Aktivierung hat. In Abbildung 11 sind die dazugehörigen Diagramme abgebildet. Zwar waren die p-Werte der Modelle erneut relativ hoch, was wieder daran liegt, dass bei der Untersuchung einer Moderation diejenigen Prädiktoren, die die Interaktion konstituieren, separat als Prädiktoren im Modell enthalten sein müssen (Jaccard & Turrisi, 2003). Anreiz wurde dabei kein einziges Mal signifikant und die Interaktion erklärte in allen Fällen den größeren Teil der Varianz (siehe ΔR^2 in den jeweiligen Tabellen). Auch alle Regressionsmodelle bzgl. der KE-Aufgabe wurden auf Erfüllung der Voraussetzungen geprüft. Dabei wurden weder Heteroskedastizität, Autokorrelation noch Multikollinearität festgestellt, weswegen die Voraussetzungen für die Spezifizierung der Modelle erfüllt waren.

Bei der KE-Aufgabe kann somit gesagt werden, dass der monetäre Anreiz einen schädlichen Effekt auf Vergnügen, Flow und Positive Aktivierung in Abhängigkeit des Wertesystems

Tabelle 65. Lineare Regression mit Interaktionseffekt zwischen ***Verstehen***^{V} und Anreiz bei Konzeptentwicklungsaufgabe auf Flow.

Prädiktor	B	SE	t	p
$(Konstante)$	3.71	0.09	43.45	<.001
Verstehen^V	-0.33	0.19	1.72	.09
Anreiz	-0.04	0.12	0.36	.72
Verstehen^V $\times$ Anreiz	0.68	0.27	2.51	<.05
R^2	.04			
ΔR^2	.04			
F-Test	$F(3, 160) = 2.11, p = .10$			

Anmerkung. N = 166; B = Regressionsgewicht; SE = Standardfehler; V = Vermeidung; ΔR^2 = Differenz zum Modell ohne Interaktionsterm.

Tabelle 66. Lineare Regression mit Interaktionseffekt zwischen ***Verstehen***^{V} und Anreiz bei Konzeptentwicklungsaufgabe auf Vergnügen.

Prädiktor	B	SE	t	p
$(Konstante)$	2.88	0.1	28.95	<.001
Verstehen^V	-0.44	0.21	2.11	<.05
Anreiz	0.01	0.14	0.1	.92
Verstehen^V $\times$ Anreiz	0.85	0.31	2.77	<.01
R^2	.05			
ΔR^2	.05			
F-Test	$F(3, 162) = 2.66, p = .05$			

Anmerkung. N = 166; B = Regressionsgewicht; SE = Standardfehler; V = Vermeidung; ΔR^2 = Differenz zum Modell ohne Interaktionsterm.

Verstehen hat. Während in der Gruppe ohne Anreiz das Motivationsempfinden signifikante Zusammenhänge in hypothetisierter Richtung mit ***Verstehen*** zeigt, waren diese Zusammenhänge in der Gruppe mit Anreiz entgegengesetzt.

Zum Abschluss dieser Untersuchung wurde überprüft, ob signifikante Korrelationen der abhängigen Variablen mit Wertesystemen auftreten, die jeweils inkongruent zu den Aufgaben sind. Die entsprechenden Korrelationen sind in den Tabellen 68 bis 70 aufgeführt, wobei der Übersichtlichkeit halber nur die signifikanten Korrelationen abgebildet sind.

Tabelle 67. Lineare Regression mit Interaktionseffekt zwischen ***Verstehen***^{V} und Anreiz bei Konzeptentwicklungsaufgabe auf Positive Aktivierung.

Prädiktor	B	SE	t	p
$(Konstante)$	5.12	0.13	40.41	<.001
Verstehen^V	-0.35	0.27	1.33	.18
Anreiz	-0.12	0.17	0.67	.50
Verstehen^V $\times$ Anreiz	0.97	0.39	2.47	<.05
R^2	.04			
ΔR^2	.04			
F-Test	$F(3, 160) = 2.2, p = .09$			

Anmerkung. N = 166; B = Regressionsgewicht; SE = Standardfehler; V = Vermeidung; ΔR^2 = Differenz zum Modell ohne Interaktionsterm.

Bei der SB-Aufgabe (Tabelle 68) sind dabei insgesamt 21, bei der GM-Aufgabe (Tabelle 69) 22 und bei der KE-Aufgabe 68) 22 von jeweils 168 Korrelationen signifikant. Die Zahl von 168 Korrelationen erklärt sich so, dass sieben AVs mit zwölf Wertesystemen in zwei experimentellen Bedingung miteinander korreliert wurden, wobei bei jeder Aufgabe die Korrelationen von zwei Wertesystemen (Annäherungs- und Vermeidungsdimension des zur Aufgabe kongruenten Wertesystems) nicht mehr berechnet wurden, da diese bereits zuvor berichtet wurden.

Aus diesen Korrelationen kann gesehen werden, ob andere als die erwarteten Wertesysteme mit den entsprechenden Aufgaben in Zusammenhang standen oder vice versa die Aufgaben kongruent zu anderen als den hypothetisierten Wertesystemen waren. Um diese Frage zu beantworten, sollen nur die Korrelationen unter der Bedingung ohne Anreiz betrachtet werden, da hierbei die Kongruenz zwischen Wertesystem und Aufgabe ohne Einfluss der experimentellen Bedingung zum Ausdruck kommt. Bzgl. der SB-Aufgabe (Tabelle 68) kann dabei gesehen werden, dass diese in negativem Zusammenhang mit ***Macht***^{A} (PA und VA), ***Verstehen***^{A} (WK und VA) und ***Nachhaltigkeit***^{A} (WK) steht. Außerdem korrelieren in der Gruppe ohne Anreiz ***Geborgenheit***^{V} negativ mit PA und VA, ***Erfolg***^{V} positiv mit Flow, Vergnügen und VA sowie ***Nachhaltigkeit***^{V} mit WK. Aus diesen Korrelationen kann abgeleitet werden, dass kein Wertesystem kongruenter zur SB-Aufgabe ist als ***Gewissheit***, jedoch mehrere Wertesysteme das Gegenteil von kongruent sind. Des Weiteren ist aufzuführen, dass mehrere Korrelationen mit ***Gleichheit***^{A} und ***Gleichheit***^{V} in der Gruppe mit monetärem Anreiz signifikant wurden.

Bei der GM-Aufgabe (Tabelle 69) zeigen sich ebenfalls einige signifikante Korrelationen, wobei auch hier gilt, dass kein Wertesystem kongruenter zu dieser Aufgabe war als ***Erfolg***, sich jedoch erneut mehrere Wertesysteme anti-kongruent zeigten. Allen voran ***Gleichheit***^{A}

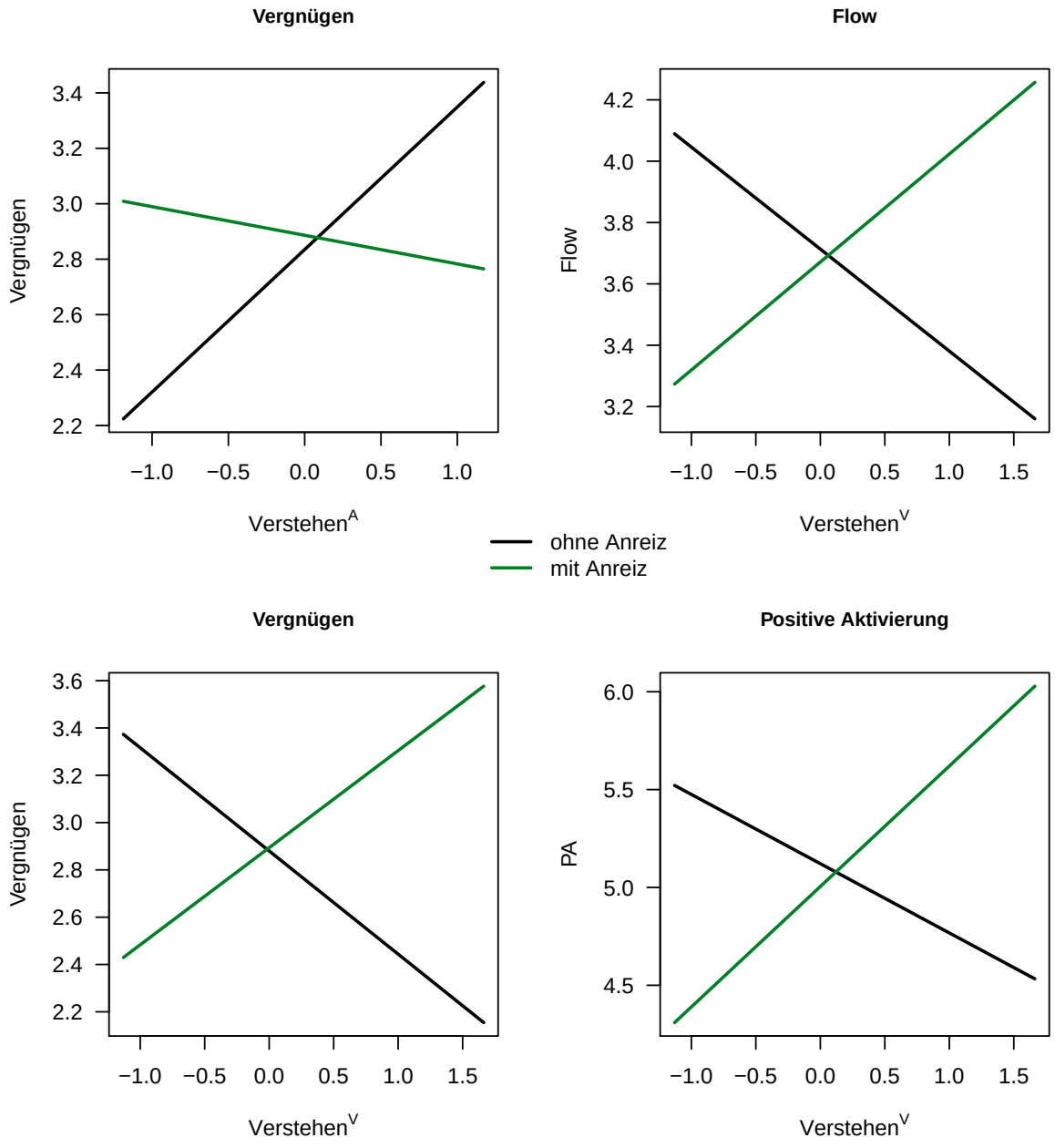

Abbildung 11. Ergebnisse der Moderatoranalyse bei der Konzeptentwicklungsaufgabe: Einfluss von monetärem Anreiz auf Intensität der Motivation in Abhängigkeit des ***Verstehen***-Wertesystems.

mit drei signifikanten Korrelationen (mit Vergnügen, WK und VA) und $\boldsymbol{Nachhaltigkeit}^{A}$, das mit PA signifikant negativ korreliert. Die Korrelationen einiger Wertesysteme mit dem Flow-Besorgnis-Faktor sind an dieser Stelle schwer zu interpretieren.

Bei der KE-Aufgabe zeigte sich $\boldsymbol{Erfolg}^{V}$ als kongruent zur Aufgabe, da Flow, Vergnügen und Valenz signifikant mit diesem Wertesystem korrelieren. Auch $\boldsymbol{Macht}^{V}$ wies eine signifikante

Korrelation auf, und zwar mit Vergnügen. Des Weiteren gab es Zusammenhänge der KE-Aufgabe mit ***Macht***^{A} bei monetärem Anreiz, da Flow, WK und VA signifikant korrelieren.

In Summe zeigen diese Korrelationen an, dass die Aufgaben-Wertesystem-Kongruenzen am größten so wie ursprünglich hypothetisiert waren, wobei sich jedoch einige der Wertesysteme als das Gegenteil von kongruent erwiesen und deshalb als anti-kongruent bezeichnet werden können. An der Gültigkeit der Befunde zum Einfluss der Wertesystem-Kongruenz auf die Intensität von Motivation ändern diese Korrelationen jedoch wenig, da sich kein Wertesystem als kongruenter als die beabsichtigten Wertesysteme erwiesen.

Tabelle 68. Korrelationen der AVs mit inkongruenten Wertesystemen bei der Sachbearbeitungsaufgabe.

	GB	GB	MA	ER	ER	GL	VE	VE	NA	NA
Anreiz	ohne	mit	ohne	ohne	mit	mit	ohne	mit	ohne	mit
Annäherung										
Flow										
BE										
V						.25*				
WK							-.25*		-.22*	
PA			-.24*							
NA										
VA			-.26*			.22*	-.26*			
Vermeidung										
Flow				.23*		-.24*				
BE					-.25*					
V				.37***		-.24*				-.25*
WK								.24*	.32**	
PA	-.32**									
NA										
VA	-.31**	-.22*		.34**		-.32**				-.26*

Anmerkung. Nicht signifikante Korrelationen sind nicht enthalten. BE = Besorgnis; V = Vergnügen; WK = Wahrgenommene Kompetenz; PA = Positive Aktivierung; NA = negative Aktivierung; VA = Valenz; GB = Geborgenheit; MA = Macht; ER = Erfolg; GL = Gleichheit; VE = Verstehen; NA = Nachhaltigkeit; * p < .05, ** p < .01.

Tabelle 69. Korrelationen der AVs mit inkongruenten Wertesystemen bei der Gewinnmaximierungsaufgabe.

Anreiz	GB ohne	MA ohne	MA mit	GW mit	GL ohne	GL mit	VE mit	NA ohne	NA mit
Annäherung									
Flow						-.24*			
BE			.26*					-.35**	
V					-.23*	-.34**			
WK					-.29**				
PA						-.31**		-.36***	
NA				-.28**					
VA				.29**	-.24*	-.28*			
Vermeidung									
Flow									
BE		-.24*	-.24*	.24*		.22*			
V						.27*	-.29**		
WK									
PA	-.23*								
NA				.34**					
VA				-.25*					.24*

Anmerkung. Nicht signifikante Korrelationen sind nicht enthalten. BE = Besorgnis; V = Vergnügen; WK = Wahrgenommene Kompetenz; PA = Positive Aktivierung; NA = negative Aktivierung; VA = Valenz; GB = Geborgenheit; MA = Macht; GW = Gewissheit; GL = Gleichheit; VE = Verstehen; NA = Nachhaltigkeit; * $p < .05$, ** $p < .01$, *** $p < .001$.

10.2.5 Diskussion

In diesem Kapitel wurde eine experimentelle Untersuchung zur prädiktiven Validität der Wertesysteme berichtet. Dazu wurden drei Aufgaben konstruiert, die zu drei Wertesystemen kongruent waren und mit der experimentellen Bedingung eines monetären Anreizes insgesamt 166 VPn präsentiert wurden. Es wurde hypothetisiert, dass die Intensität der Motivation, die durch die Konstrukte Flow, Vergnügen, wahrgenommene Kompetenz sowie positive und negative Aktivierung und Valenz operationalisiert wurde, einen Zusammenhang mit den zur

Tabelle 70. Korrelationen der AVs mit inkongruenten Wertesystemen bei der Konzeptentwicklungsaufgabe.

	GB	MA	MA	ER
Anreiz	ohne	ohne	mit	ohne
Annäherung				
Flow			.22*	
BE	-.24*			
V				
WK			.25*	
PA				
NA				
VA			.22*	
Vermeidung				
Flow				.29**
BE				
V		.25*		.28*
WK				
PA				
NA				
VA				.27*

Anmerkung. Nicht signifikante Korrelationen sind nicht enthalten. BE = Besorgnis; V = Vergnügen; WK = Wahrgenommene Kompetenz; PA = Positive Aktivierung; NA = negative Aktivierung; VA = Valenz; GB = Geborgenheit; MA = Macht; GW = Gewissheit; ER = Erfolg; GL = Gleichheit; NA = Nachhaltigkeit; * p < .05, ** p < .01.

Aufgabe kongruenten Wertesystemen aufweist. Monetärer Anreiz wurde dabei als potenziell interagierende Variable aufgenommen.

Zunächst kann festgestellt werden, dass bei keiner Aufgabe ein signifikanter Haupteffekt der experimentellen Bedingung (monetärer Anreiz) bzgl. der AVs (Maße der Intensität der Motivation) gefunden wurden. Das bedeutet, dass das Vorhandensein oder Fehlen eines monetären Anreizes innerhalb einer Aufgabe keinen Unterschied bzgl. des Motivationsempfindens machte. Zwar gab es signifikante Unterschiede der Intensität der Motivation zwischen den Aufgaben, diese müssen jedoch in der Unterschiedlichkeiten Aufgaben begründet liegen. Auch

die Eigenschaften (Schwierigkeit und Anforderungen) der Aufgaben, die in der Flow-Kurzskala abgefragt wurden, bestätigen diese Schlussfolgerung. Als Folge daraus wurden die AVs nicht weiter bzgl. Unterschiede zwischen den Aufgaben untersucht. Um in zukünftigen Untersuchungen auch Untersuchungen zum Unterschied zwischen Aufgaben plausibel durchführen zu können, sollte weitere Arbeit in die Entwicklung gleich schwieriger Aufgaben investiert werden.

Insgesamt zeigte sich bei zwei der drei Aufgaben mehrere hypothetisierte signifikante Zusammenhänge. Bei der Gewinnmaximierungsaufgabe wurde der Zusammenhang zwischen ***Erfolg***^{A} und PA sowie NA signifikant (Haupteffekt in Regression). Außerdem waren die Korrelationen zwischen ***Erfolg***^{A} und PA in der Gruppe mit monetärem Anreiz signifikant. Da jedoch keine Interaktion zwischen ***Erfolg***^{A} und Anreiz gefunden wurde, kann zwar die Hypothese zum Zusammenhang von ***Erfolg***^{A} und Intensität der Motivation bei Kongruenz der Aufgabe mit ***Erfolg*** bestätigt werden, nicht jedoch der hypothetisierte Zusammenhang zwischen monetärem Anreiz und ***Erfolg***^{A}. Wobei andererseits die signifikanten Korrelationen nur unter der Bedingung von monetärem Anreiz auftraten. Bei ***Erfolg***^{V} zeigte sich auch ein signifikanter Haupteffekt (mit umgekehrtem Vorzeichen) auf PA, zudem wurde die Korrelation mit Vergnügen signifikant (auch umgekehrtes Vorzeichen). Da auch hier keine Interaktion mit Anreiz festgestellt wurde, dennoch nur Korrelationen unter der Anreiz-Bedingung signifikant wurden, gilt dieselbe Aussage bzgl. der Hypothesen wie bei ***Erfolg***^{A}.

Bei der Konzeptentwicklungsaufgabe zeigten sich signifikante hypothesenkonsistente Zusammenhänge zwischen ***Verstehen*** (auf beiden Dimensionen) und Vergnügen in der Gruppe ohne Anreiz. Außerdem wurden die Korrelationen (ebenso hypothesenkonsistent) von ***Verstehen***^{V} mit Flow und Valenz signifikant. Darüber hinaus zeigten sich Moderatorseffekte von Anreiz auf den Zusammenhang zwischen ***Verstehen***^{A} und Vergnügen sowie ***Verstehen***^{V} und Flow, Vergnügen und PA. Für diesen Teil der Studie kann somit die Hypothese bestätigt werden, dass die Passung von ***Verstehen*** mit der Aufgabe zu einer erhöhten Motivationsintensität führt. Darüber hinaus wurde festgestellt, dass ein monetärer Anreiz einen negativen Effekt auf das Motivationsempfinden in Abhängigkeit der Ausprägung des kongruenten Wertesystems haben kann. Diese Befunde können als Erklärung dafür gesehen werden, wann bzw. welchen Effekt ein monetärer Anreiz auf die Motivation haben kann und damit auch einen Ansatz für die Integration von sich widersprechenden Studienergebnisse fungieren. Darüber hinaus eröffnen diese Ergebnisse Raum für weitere Untersuchungen, in denen zum einen z.B. unterschiedlich hohe Geldbeträge als Anreize gesetzt werden könnten oder gänzlich andere Anreize verwendet werden könnten (z.B. Sachpreise wie die Chance auf einen Urlaub oder die Teilnahme an einer Fortbildung).

Insgesamt kann aus den Ergebnissen abgeleitet werden, dass die Passung zwischen Wertesystempräferenz und Aufgabe einen signifikanten Einfluss auf die Intensität der Motivation haben kann. Dies hat sowohl Auswirkungen auf die psychologische Forschung, da in experimentellen

Studien häufig Aufgaben präsentiert werden, jedoch nicht berücksichtigt wird, ob die Aufgabe kongruent zu einem bestimmten Wertesystem ist und dadurch die Generalisierbarkeit von Ergebnissen einschränkt. Diese Schlussfolgerung gilt insbesondere auch deshalb, da es sich bei der hier verwendeten Stichprobe wie in den meisten Experimentalstudien um ein studentisches Sample handelt. In zukünftigen experimentellen müsste auf Unabhängigkeit der Aufgaben von Wertesystemen getestet werden.

Nimmt man an, dass sich die Ergebnisse auch auf den Teil der Bevölkerung verallgemeinern lässt, der nicht studiert, dann haben sie Konsequenzen für die Frage nach der Passung zwischen Mitarbeiter (und seiner Wertesysteme) mit einer Aufgabe. Wertesysteme könnten demnach hilfreiche Indikatoren dafür sein, welche Art von Aufgabe eher zu wem passt und ob ein monetärer Anreiz zu mehr Motivation führt. Diese Funktion ist wiederum für die Führung von Mitarbeitern relevant, denn ein elementarer Bestandteil von Mitarbeiterführung besteht in der Delegation von Aufgaben und dem Setzen von Anreizen.

Bzgl. der Orthogonalitätshypothese der Wertesysteme kann gesagt werden, dass die Befunde bei allen drei Aufgaben tendenziell entgegengesetzt sind. Am deutlichsten ist die Gegenläufigkeit von Annäherung und Vermeidung bei ***Verstehen*** in der KE-Aufgabe zu erkennen, aber auch bei ***Erfolg*** und der GM-Aufgabe können entgegengesetzte Tendenzen der Kennwerte festgestellt werden. Zwar ohne Signifikanzen, jedoch zumindest auch den Vorzeichen nach zu urteilen scheinen auch die Ergebnisse bei ***Gewissheit*** und der SB-Aufgabe gegenläufig. Es kann deshalb gesagt werden, das die Ergebnisse dieser Studie eher für die Bipolarität der Wertesysteme sprechen.

Kritisch ist zu sehen, dass es sich einerseits um eine studentische Stichprobe handelt und Verallgemeinerungen der gerade getroffenen Art deshalb mit Vorsicht zu handhaben sind. Andererseits muss bedacht werden, dass die Beschaffenheit der Aufgaben nicht empirisch überprüft wurde. Sie wurden lediglich auf Grundlage theoretischer Überlegungen hin konstruiert und es kann nicht gesagt werden, in welchem Ausmaß sie zu den entsprechenden Wertesystemen kongruent waren. Beide Extreme sind denkbar. Im Falle, dass die Aufgaben sehr kongruent zu den Wertesystemen waren, würde dies den Einfluss von Wertesystemen auf Motivationsempfinden verringern. Falls sie wenig kongruent waren, würde dies eher für einen robusten Zusammenhang zwischen Wertesystemen und Motivationsempfinden sprechen. Finale Aussagen über das Ausmaß des Zusammenhangs zwischen Wertesystemen und Intensität der Motivation wären nur dann zulässig, wenn man das Ausmaß der Passung zwischen Wertesystem und Aufgabe quantifizieren könnte. Für die zukünftige Forschung wäre die Entwicklung eines solchen Maßes wünschenswert, das darüber hinaus den Vergleich zwischen unterschiedlichen Studien ermöglichen würde. Zwar waren die Effektstärken und Bestimmtheitsmaße tendenziell gering, dennoch kann aufgrund der eben dargelegten Sachlage nicht gesagt werden, ob der Grund dafür methodische Artefakte oder in der Natur der Beziehung zwischen Wertesystemen und Motivationsempfinden liegt. Abgesehen davon wurden in vergleichbaren Studien zum Zu-

sammenhang zwischen Werte-Kongruenz im Kontext der Person-Organisation-Fit-Forschung Effektstärken in ähnlichen Größenordnungen gefunden (vgl. Verquer et al., 2003).

Insgesamt deutet diese Studie an, wie komplex man den Zusammenhang von der Richtungs- und Intensitätsdimensionen der Motivation im Zusammenhang mit Werten als Indikator für die Richtung der Motivation modellieren kann. Damit eröffnet diese Studie ein breites Feld an unerforschten Fragestellungen. Welche Wertesysteme sind mit welchen Aufgaben kongruent? Wie kann die Kongruenz zwischen Aufgaben und Wertesystemen quantifiziert werden? Hat die Kongruenz von Wertesystemen und Führungsstil Auswirkungen auf die Motivation? Um nur ein paar Fragen zu nennen. Darüber hinaus können einen Vielzahl von Incentives in das Versuchsdesign aufgenommen werden und deren mögliche moderierende Wirkung untersucht werden.

10.3 Studie zur prädiktiven und inkrementellen Validität

In dieser Untersuchung wird neben der prädiktiven auch die inkrementelle Validität der Wertesysteme untersucht, indem diese in Bezug zu den Verkaufsdaten einer Stichprobe von Vertriebsmitarbeitern gesetzt werden. Das Außenkriterium stellt in dieser Untersuchung die mittlere monatliche Verkaufszahl der Vertriebsmitarbeiter über deren Zeitraum der Betriebszugehörigkeit dar. Inkrementelle Validität eines Merkmals ist dann gegeben, wenn dieses zusätzlich zu bekannten Kriteriumsvariablen Varianz aufklären kann (Hunsley & Meyer, 2003; Sechrest, 1963). In der vorliegenden Untersuchung liegen als zusätzliche Kriteriumsvariablen bei jedem Vertriebsmitarbeiter Alter, Geschlecht, Berufserfahrung und Dauer der Betriebszugehörigkeit vor.

Es wird vermutet, dass es einen linearen Zusammenhang zwischen den Ausprägungen bestimmter Wertesysteme und der Leistung (Verkaufserfolg) im Vertrieb gemessen an den mittleren Verkäufen gibt. Die Experten, die in der Untersuchung zur konkurrenten Validität herangezogen wurden (Kapitel 10.1), vermuteten einen Zusammenhang zwischen der durchschnittlichen Ausprägung von ***Erfolg*** von Vertriebsmitarbeitern. Diese Hypothese wurde für $\boldsymbol{Erfolg}^{V}$ bestätigt. Es liegt nahe, daraus zu schließen, dass $\boldsymbol{Erfolg}^{V}$ in signifikantem Zusammenhang mit dem Verkaufserfolg steht, insbesondere dann, wenn man davon ausgeht, dass sich Menschen Aufgaben suchen, in denen sie erfolgreicher sind als der Durchschnitt. Darüber hinaus hatten Vertriebsmitarbeiter höhere Ausprägungen auf $\boldsymbol{Macht}^{A}$ als zwei Vergleichsgruppen, niedrigere Ausprägungen auf $\boldsymbol{Verstehen}^{A}$ und $\boldsymbol{Nachhaltigkeit}^{A}$ sowie höhere Ausprägungen auf $\boldsymbol{Verstehen}^{A}$ und $\boldsymbol{Nachhaltigkeit}^{A}$ als mehrere Vergleichsgruppen. Bemüht man dieselbe Schlussfolgerung, kann auch mit diesen Wertesystemen deshalb ein Zusammenhang mit Verkaufserfolg vermutet werden.

Auch in der Literatur gibt es Studien, die einen Zusammenhang zwischen Persönlichkeitsfaktoren und Erfolg von Vertriebsmitarbeitern herstellen. Der Überblickartikel von Vinchur et al. (1998) zeigt, dass neben Fähigkeitsdispositionen auch insbesondere die Faktoren Leistungsorientierung (*Achievement*) und Entscheidungsfreude (*Assertiveness*) signifikante Prädiktoren von Verkaufserfolg waren. Zwar waren diese beiden Faktoren den Big Five Dimensionen Extraversion bzw. Gewissenhaftigkeit untergeordnet, erinnern in dieser Form jedoch an die Wertesysteme ***Erfolg*** bzw. ***Macht***.

Für privatwirtschaftlich agierende Unternehmen ist die Fragestellung dieser Untersuchung von höchster Relevanz, da der Vertriebserfolg den langfristigen Erfolg oder Misserfolg einer Unternehmung häufig maßgeblich beeinflusst.

10.3.1 Methode

Zur Analyse der prädiktiven Validität werden häufig Korrelationen eingesetzt (Groth-Marnat, 2003). Um das Ausmaß der inkrementellen Validität zu bestimmen, eignet sich die Durchführung

einer hierarchischen Regressionsanalyse (auch Mehrebenenanalyse genannt, Eid et al., 2015), da darin über die Differenz der aufgeklärten Varianz der direkt auf die eingefügte Prädiktorvariable zurückgehende Einfluss – das sogenannte Inkrement (Kersting, 2001) – beziffert werden kann (Haynes & Lench, 2003).

Die Stichprobe der Vertriebsmitarbeiter, die den Verkaufszahlen zugrunde liegt, wurde bereits in Kapitel 3.7 als Stichprobe IIa beschrieben. Bei der Firma handelte es sich um ein Unternehmen mit ca. 100 Mitarbeitern aus dem IT-Sektor. Die Dauer der Betriebszugehörigkeit lag im Mittel bei 9.9 Monaten ($SD = 8.39$) und schwankte zwischen 1 und 28 Monaten.

Die in dieser Studie verwendeten Materialien waren zum einen der MVSQ, den die Versuchspersonen entweder zum Beginn der Studie oder zum Beginn der Tätigkeit beim Unternehmen bearbeitet haben. Die Wertesysteme wurden somit zwar zu unterschiedlichen Zeitpunkten gemessen, die Bedingung für Untersuchungen der prädiktiven Validität, dass die Merkmalsmessung zeitlich *vor* der Erhebung der abhängigen Variable erfolgt, war damit jedoch erfüllt. Alle Teilnehmer haben ihr Einverständnis erklärt, dass ihre Daten für wissenschaftliche Zwecke verwendet werden dürfen. Außerdem erhielt jede Versuchsperson einen schriftlichen Feedback-Bericht.

Als abhängige Variable wurde die durchschnittliche monatliche Anzahl von Verkäufen pro Mitarbeiter verwendet. Diese wurden nach Vereinbarung mit der Leitung der Vertriebsabteilung von eben dieser in pseudonymisierter Form zum Ende der Studie übermittelt. Dem Stichprobenumfang entsprechend lagen 31 durchschnittliche Verkaufszahlen vor.

Neben der Berechnung der bivariaten Korrelationen zwischen Wertesystemen und dem durchschnittlichen Verkaufserfolg als direkte Zusammenhangsmaße, werden hierarchische Regressionsanalysen ähnlich der Vorgehensweise bei Day und Silverman (1989) und Elhai et al. (2008) berechnet. Darin kann in mehreren geschachtelten Regressionsmodellen überprüft werden, ob Wertesysteme zusätzliche Varianz erklären, also inkrementelle Validität besitzen. Als Index der inkrementellen Validität wird die Differenz der Bestimmtheitsmaße R^2 zwischen dem Modell mit und ohne inkrementellem Prädiktor herangezogen (Haynes & Lench, 2003).

Bei der Wahl eines geeigneten Regressionsmodells war zunächst zu beachten, dass es sich bei den Verkaufszahlen um positive Zähldaten handelte. Gewöhnliche Regressionmodelle (Ordinary Least Squares (OLS)-Modelle) sind in der Regel zur Analyse solcher Daten ungeeignet, da Zähldaten erstens häufig nicht normalverteilt sind und zweitens per Definition nicht negativ werden können (Gardner et al., 1995). Letzteres kann in OLS Modellen nicht berücksichtigt und deshalb zu falschen Schlussfolgerungen führen (Gardner et al., 1995). Alternativ stehen dafür generalisierte lineare Regressionsmodelle (GLM) mit geeigneten Verteilungsfunktionen zur Verfügung (Poisson, negativ Binomial, Gamma), die speziell für positive Zielvariablen konzipiert wurden (Fahrmeir et al., 2009; Gelman & Hill, 2006). Bei der vorliegenden Zielvariable (durchschnittliche monatliche Verkaufszahl) handelt es sich um eine positive und stetige Variable. Die Annahme der Poisson- und negativ-Binomialverteilung trifft deshalb nicht zu, da

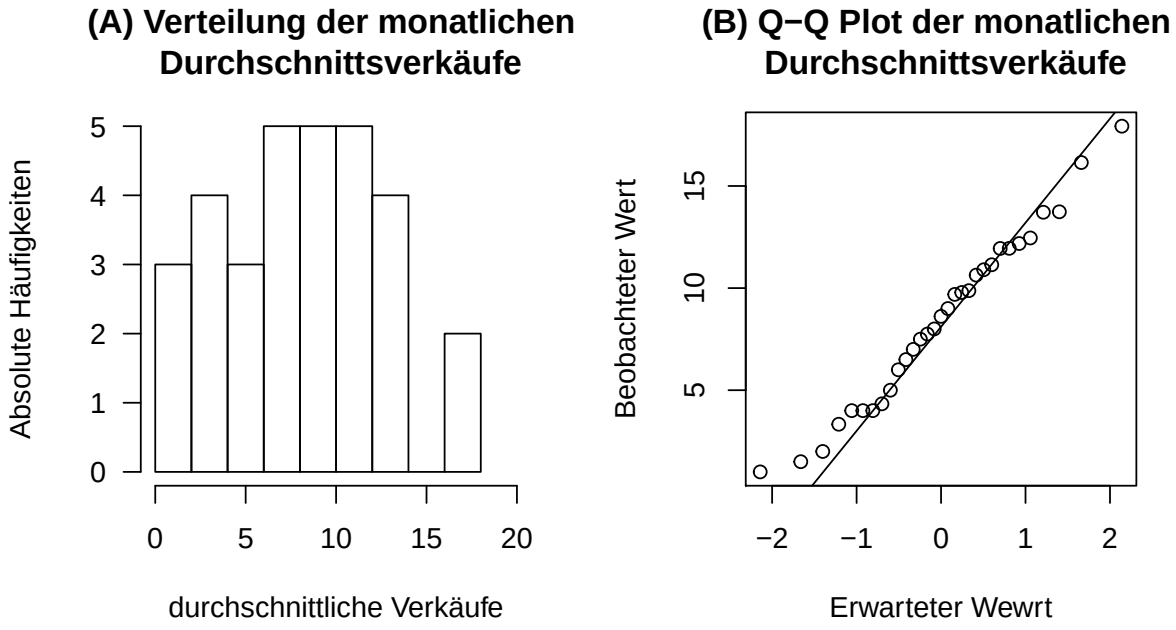

Abbildung 12. Verteilung (A) und Quantil-Quantil-Plot (B) der durchschnittlichen monatlichen Verkäufe pro Mitarbeiter.

diese Verteilungsfunktionen nur für ganzzahlige Variablen gelten (Gelman & Hill, 2006). Passend für positive stetige Zielvariablen ist die Annahme der Gamma-Verteilung (Fahrmeir et al., 2009). Deshalb wurden die GLMs zur Bestimmung der Inkremente unter Annahme der Gamma-Verteilung spezifiziert. Abbildung 12 zeigt ein Histogramm und Quantil-Quantil-Diagramm der Verkaufsdaten.

Da bei GLMs kein exaktes Äquivalent zum R^2 der linearen Regression berechnet werden kann, müssen zur Beurteilung der inkrementellen Validität stattdessen sogenannte Pseudo-R^2-Maße, wie McFadden's ρ (McFadden, 1974) oder Nagelkerke's R^2 (Nagelkerke, 1991)[2] verwendet werden (Elhai et al., 2008; McGrath et al., 2002). Bei der Interpretation von Pseudo-R^2-Maßen ist allerdings zu beachten, dass diese nicht eins zu eins wie klassische R^2 interpretiert werden können (Long & Freese, 2001) und niedriger ausfallen können (Smith & McKenna, 2013). Eine Monte-Carlo-Vergleichsstudie von sieben Pseudo-R^2-Maßen von Smith und McKenna (2013) hat jedoch gezeigt, dass Nagelkerke's R^2_N relativ nahe und nur leicht unter dem OLS-R^2 lag. McFadden's R^2 unterschritt das normale R^2 hingegen stark. Deshalb wird in dieser Untersuchung auf Nagelkerke's R^2_N zurückgegriffen.

10.3.2 Ergebnisse

Zunächst wurde aufgrund des geringen Stichprobenumfangs ein Überblick über die Beschaffenheit der Wertesystemausprägungen in der Stichprobe ermittelt (Tabelle 71). Außerdem wurde

[2]Nagelkerke's R^2 wird teilweise auch als Cragg-Uhler's R^2 bezeichnet (vgl. Windzio, 2013).

Mardia's (1970) Test auf multivariate Normalverteilung mit einer Funktion aus dem R-Paket MVN (Korkmaz et al., 2014) berechnet. Dieser ergab, dass die Daten multivariat normalverteilt sind, da die p-Werte für Kurtosis und Schiefe beide $> .05$ sind. Bei den Wertesystemen gibt es somit keine Ausreißer und sowohl die Schiefen als auch Kurtosi befanden sich relativ nahe an der Wölbung der Normalverteilung (DeCarlo, 1997). Außerdem enthält Tabelle 71 die Kenndaten der Variablen Alter, Berufserfahrung und Betriebszugehörigkeit. Von diesen ist vor allem die Betriebszugehörigkeit nicht mehr normalverteilt ($< .01$ bei Shapiro-Wilk-Tests) und deshalb bei der Interpretation mit Vorsicht zu behandeln. Auch eine Logarithmierung konnte diese Variable nicht in eine Normalverteilung überführen ($p = .03$), jedoch war sie der Normalverteilung nach Transformation trotz Zweigipfligkeit etwas ähnlicher (Tabelle 13). Sie wurde deshalb im weiteren Verlauf dieser Studie als logarithmierte Variable verwendet.

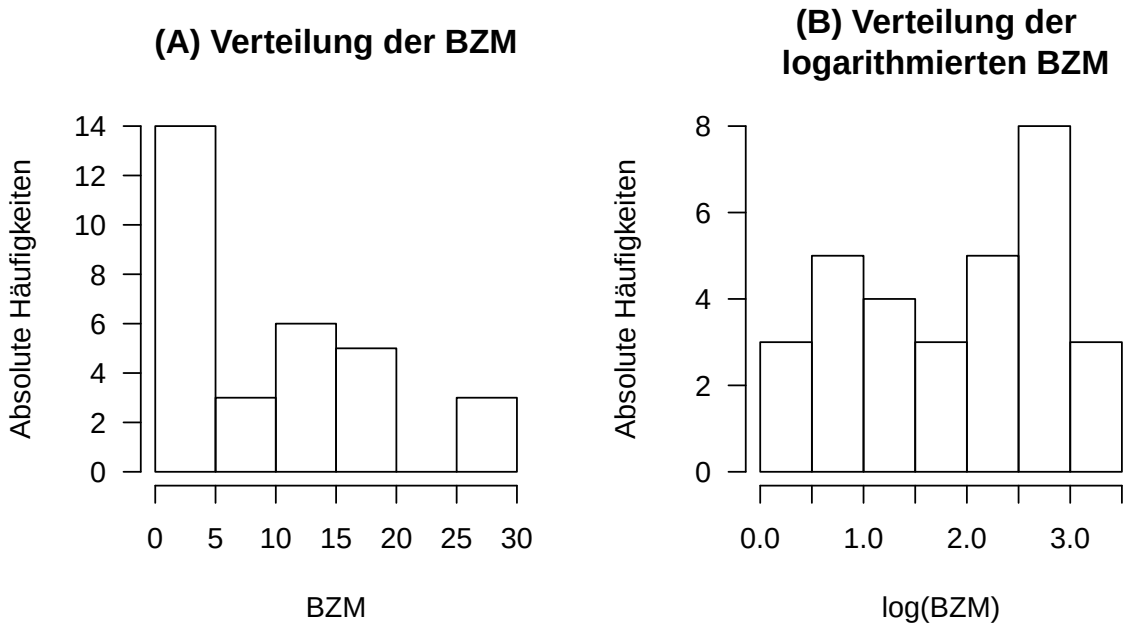

Abbildung 13. Verteilung der Betriebszugehörigkeit in Monaten (BZM) normal (A) und logarithmiert (B).

10.3.2.1 Prädiktive Validität

Zum Bericht der Korrelationen der Prädiktoren (UVs) mit den durchschnittlichen Verkaufszahlen (AV) wurde eine Darstellung in Anlehnung an Haynes und Lench (2003) gewählt, in der neben den Korrelationen zwischen AV und UVs auch die Interkorrelationen der Prädiktoren aufgeführt werden. Dies hat den Vorteil, dass der Anteil gemeinsamer Varianz der Prädiktoren und dadurch der Grad ihrer Unabhängigkeit auf einen Blick begutachtet werden können. Letzteres ist im Hinblick auf die mögliche inkrementelle Validität von Nutzen, da gegebene Unabhängigkeit zweier UVs als Hinweis auf die inkrementelle Validität einer der beiden Prädiktoren

Tabelle 71. Deskriptivstatistiken der Wertesysteme.

Wertesystem	M	SD	Min	Max	Schiefe	Kurtosis
Geborgenheit^A	0.13	0.31	-0.63	0.79	-0.24	2.90
Macht^A	-0.02	0.36	-0.78	0.70	-0.20	2.54
Gewissheit^A	0.15	0.42	-0.76	0.93	-0.30	2.80
Erfolg^A	0.07	0.33	-0.90	0.62	-0.76	3.68
Gleichheit^A	0.03	0.37	-0.49	0.96	0.53	2.65
Verstehen^A	-0.31	0.35	-0.94	0.34	-0.13	2.27
Nachhaltigkeit^A	-0.13	0.33	-0.89	0.62	0.13	3.12
Geborgenheit^V	-0.13	0.27	-0.53	0.54	0.73	2.68
Macht^V	-0.02	0.32	-0.82	0.43	-0.58	2.75
Gewissheit^V	-0.13	0.45	-0.94	0.93	0.35	3.10
Erfolg^V	-0.16	0.27	-0.66	0.48	0.16	2.44
Gleichheit^V	0.06	0.39	-0.81	0.90	-0.01	2.72
Verstehen^V	0.12	0.38	-0.63	0.71	-0.33	2.01
Nachhaltigkeit^V	0.28	0.45	-0.62	1.11	-0.16	2.71
Alter	33.06	6.86	23.00	48.00	0.57	2.36
Berufserfahrung	7.67	6.00	0.00	26.00	1.00	4.29
log(Betriebszugehörigkeit)	1.84	1.06	0.00	3.33	-0.28	1.80

Anmerkung: A = Annäherung, V = Vermeidung.

gewertet werden kann. Bezogen auf die Vorhersage der Kriteriumsvariable werden zwei Variablen umso redundanter, je stärker sie zusammenhängen. Tabelle 72 zeigt die entsprechenden Korrelationen.

Es ist zu sehen, dass die drei Wertesysteme ***Macht***^{A}, ***Erfolg***^{A}, und ***Gewissheit***^{V} mit r = .38, .41 und .36, sowie die logarithmierte Betriebszugehörigkeit (BZM) mit r =.85 signifikant mit den durchschnittlichen Verkaufszahlen korrelierten. Damit können für die Annäherungsdimension die Hypothesen bzgl. ***Macht*** und ***Erfolg*** bestätigt werden, dass diese Wertesysteme signifikant mit dem Verkaufserfolg zusammenhängen. Der Zusammenhang mit ***Gewissheit***^{V} war so nicht erwartet und die hohe Korrelation zwischen Betriebszugehörigkeit und durchschnittlicher Verkaufszahl lässt sich so erklären, dass die VPn mit zusätzlicher Erfahrung auch mehr verkaufen, wobei allerdings zu bedenken ist, dass die BZM nicht normalverteilt ist und

Tabelle 72. Bivariate Korrelationen zwischen Prädiktoren und durchschnittlichen monatlichen Verkaufszahlen.

		Wertesysteme								
Prädiktor	VK	GB	MA	GW	ER	GL	VE	NA	Alter	BE
GB^A	-.03									
MA^A	.38*	-.16								
GW^A	-.31	.27	-.21							
ER^A	.41*	-.25	.22	-.01						
GL^A	.02	.42*	-.08	-.09	-.30					
VE^A	.18	-.38*	.35	-.23	.04	.03				
NA^A	-.09	.43*	-.12	.11	-.45*	.14	-.10			
Alter	-.30	-.03	.11	.28	-.34	.18	.00	.03		
BE	.03	.01	.28	.26	-.05	.14	.11	.05	.33	
log(BZM)	.85***	.07	.34	-.16	.52**	.06	-.03	-.21	-.26	.01
GB^V	.26									
MA^V	-.14	-.30								
GW^V	.36*	.24	-.01							
ER^V	-.07	-.32	.14	.09						
GL^V	.29	.19	.06	.31	.00					
VE^V	.00	-.14	.22	-.18	.14	.11				
NA^V	.32	.26	-.17	-.26	-.16	.03	.33			
Alter		-.26	.12	-.26	-.04	-.21	.05			
BE		.03	-.25	-.03	-.25	-.03	-.09			
log(BZM)		.12	.00	.37*	-.03	.51**	.28			

Anmerkung. N = 31; VK = durchschnittliche Verkaufszahl; GB = Geborgenheit; MA = Macht; GW = Gewissheit; ER = Erfolg; GL = Gleichheit; VE = Verstehen; NA = Nachhaltigkeit; A = Annäherung; V = Vermeidung; BE = Berufserfahrung; BZM = Betriebszugehörigkeit in Monaten; * $p < .05$; ** $p < .01$; *** $p < .001$.

sich die Frage stellt, ob dieser Zusammenhang im Falle der Normalverteilung ein ähnlich hohes Ausmaß hätte.

Des Weiteren ist zu sagen, dass ***Macht***^{A} mit keiner weiteren Variable, insbesondere nicht mit den anderen Kriteriumsvariablen (Alter, BE und BZM) signifikant korreliert. Die Voraus-

setzungen dafür, dass dieses Wertesystem auch inkrementelle Validität besitzt, ist demnach gut. ***Erfolg***^{A} und ***Gewissheit***^{V} hingegen korrelieren signifikant mit BZM ($r = .51$ bzw. .37), weshalb hier fraglich ist, ob diese Wertesysteme über BZM hinaus zusätzliche Varianz aufklären. Ein Mediatoreffekt von BZM auf den Zusammenhang zwischen diesen Wertesystemen und Verkaufserfolg ist hier denkbar.

Außerdem gibt es noch weitere Wertesysteme (***Gewissheit***^{A}, ***Geborgenheit***^{V}, ***Gleichheit***^{V} und ***Nachhaltigkeit***^{V}), die zwar nicht signifikant, aber in durchaus bemerkenswerten betragsmäßigen Größenordnung um .30 mit Verkaufserfolg korrelieren. Auch diese könnten inkrementelle Validität besitzen. Von den weiteren Kritierumsvariablen korrelierte auch Alter, zwar nicht signifikant, aber merklich mit Verkaufserfolg $r = -.30$. Die allgemeine Berufserfahrung spielte dagegen scheinbar keine Rolle.

Um einen möglichen Einfluss von Geschlecht auf die Verkaufszahlen zu untersuchen, wurde ein t-Test für unabhängige Stichproben gerechnet. Dieser ergab keinen signifikanten Unterschied zwischen Frauen und Männern ($t(19) = 0.52$, $p = .61$).

10.3.2.2 Inkrementelle Validität

Für die Mehrebenenanalyse zur inkrementellen Validität wurde in einem ersten Schritt das sogenannte Basismodell bestimmt. In dieses Modell wurden diejenigen der bekannten Kriteriumsvariablen aufgenommen, die einen signifikanten Effekt auf die Verkaufszahlen hatten. Dazu wurde in einer Schrittweise Vorwärtsprozedur (Tabelle 73) ein Modell spezifiziert, dessen Koeffizienten in Tabelle 74 zusammengefasst sind. Auf Grundlage dieses Modells wird nachfolgend die inkrementelle Validität der Wertesysteme untersucht.

Zur Bestimmung des Basismodells wurden zuerst fünf Modelle mit verschiedenen Transformationen der *Betriebszugehörigkeit in Monaten* (BZM) spezifiziert und jeweils gegen das Nullmodell getestet. Modell 5 wurde gegen Modell 4 getestet, da diese Modelle ineinander geschachtelt waren. Alle F-Tests gegen das Nullmodell wurden dabei signifikant, Modell 5 machte jedoch keinen Unterschied zu Modell 4 (siehe Tabelle 73). Die Hinzunahme des Prädiktors BZM, egal in welcher Form, hat demnach stets signifikant mehr Devianz als das Nullmodell erklärt. Insgesamt hat sich gezeigt, dass diejenigen Modelle, in denen die logarithmierte Transformation der BZM enthalten ist, sowohl die höchsten R^2_N, als auch die niedrigsten AIC erreichten (Modell 4 und 5). Zwar waren in Modell 5 beide Maße etwas besser als in Modell 4, der direkte Vergleich der Modelle in einem F-Test, zeigt jedoch, dass die Hypothese verworfen werden muss, dass Modell 5 signifikant mehr Devianz als Modell 4 erklären würde. In den Modellen 6 bis 8 wurden jeweils weitere Prädiktoren aufgenommen. Doch weder Alter (Modell 6), noch Berufserfahrung (Modell 7) oder Geschlecht (Modell 8) konnten zusätzlich zur logarithmierten Betriebszugehörigkeit Varianz erklären. Insbesondere im Hinblick auf das Kriterium der Sparsamkeit unter Berücksichtigung des geringen Stichprobenumfangs, wurde deshalb Modell 4 als Basismodell festgelegt.

Tabelle 73. Entwicklung des Basismodells für die hierarchischen Regressionsanalysen.

Modell	Prädiktor(en)	R^2_N	AIC	F	df	p
1	BZM	.46	166.3	29.61	1	<.001
2	BZM^2	.27	175.5	13.39	1	<.01
3	$BZM + BZM^2$	.63	156.7	26.73	2	<.001
4	$log(BZM)$	.67	151.3	62.12	1	<.001
5	$BZM + log(BZM)$	.70	150.2	3.17	1	.09
6	$log(BZM) + Alter$	.68	152	1.33	1	.26
7	$log(BZM) + BE$	.69	151.5	1.93	1	.18
8	$log(BZM) + Geschlecht$	.68	152.4	1.03	1	.32

Anmerkung. N = 31; BZM = Betriebszugehörigkeit in Monaten; BE = Berufserfahrung; R^2_N = Nagelkerke's R^2; AIC = Akaike Information Criterion; *df* = Freiheitsgrade; Modelle 1 bis 4 wurden im F-Test gegen das Nullmodell getestet, Modelle 5 bis 8 gegen Modell 4.

Tabelle 74. Koeffizienten des Basismodells der hierarchischen Regressionsanalyse zur Bestimmung der inkrementellen Validität von Wertesystemen.

Prädiktor	B	SE	t	p
$(Konstante)$	1.09	0.125	8.7	<.001
$log(BZM)$	0.51	0.059	8.56	<.001

Anmerkung. N = 31; Nagelkerke's R^2 = .67; F-Test gegen das Nullmodell $F(1) = 62.12, p < .001$; BZM = Betriebszugehörigkeit.

Die Modellprämissen wurden mit entsprechenden Tests überprüft und bestätigt. Ein Breusch-Pagan-Test ($p > .05$) ergab, dass keine Heteroskedastizität vorlag, die Residuen waren approximativ normalverteilt (Shapiro-Wilk-Test $p > .05$) und wiesen keine Autokorrelation auf (Durbin-Watson-Test $p > .05$)

Der Regressionskoeffizient des Basismodells kann folgendermaßen interpretiert werden (vgl. Wooldridge, 2013, S. 44): Wenn sich die Zahl des BZM um 100% erhöht, dann nimmt die durchschnittliche Verkaufszahl um 51% zu, oder anders gesagt, die Verkaufszahl verdoppelt (Steigerung um 100%) sich, wenn sich die Betriebszugehörigkeit verdreifacht (Steigerung um knapp 200%). Die Modellierung des BZM durch einen Logarithmus kann auch insofern als plausibel gesehen werden, da demnach die Lernkurve mit zunehmender Betriebszugehörigkeit abflacht.

Tabelle 75. Hierarchische Regressionsmodelle zu den univariaten inkrementellen Validitäten der Wertesysteme.

Prädiktoren	Modellkoeffizienten				Änderungsstatistiken		
Basismodell	B	SE	p	R^2_N	ΔR^2_N	$F(1)$	p
+ GW^A	-0.31	0.14	<.05	.71	.041	4.26	<.05
+ VE^A	0.4	0.16	<.05	.72	.055	6.58	<.05
+ NA^A	0.44	0.18	<.05	.70	.037	4.01	.06
+ VE^V	-0.49	0.14	<.01	.75	.085	12.47	<.01

Anmerkung. N = 31; R^2_N = Nagelkerke's R^2; GW = Gewissheit; VE = Verstehen; NA = Nachhaltigkeit; A = Annäherung; V = Vermeidung.

Tabelle 75 beinhaltet die Zusammenfassung der Koeffizienten der geschachtelten GLMs inklusive Änderungsstatistiken. Berichtet werden dabei nur diejenigen Modelle, in denen ein Wertesystem als Prädiktor signifikant wurde. Dies war bei den Wertesystemen ***Gewissheit***^{A}, ***Verstehen***^{A}, ***Nachhaltigkeit***^{A} und ***Verstehen***^{V} der Fall, wobei zu sagen ist, dass der F-Test des Modells mit ***Nachhaltigkeit***^{A} gegen das Basismodell knapp nicht signifikant wurde. Bei allen Modellen wurden Breusch-Pagan-Tests auf Heteroskedastizität, Durbin-Watson-Tests auf Autokorrelation und die Varianzinflationsfaktoren (VIF) berechnet. In den Modellen mit ***Verstehen***^{A} und ***Verstehen***^{V} wurden die Breusch-Pagan-Tests mit $p < .05$ signifikant. Im Modell mit ***Nachhaltigkeit***^{A} zeigte ein Durbin-Watson-Test Abhängigkeiten der Fehler durch einen p-Wert < .05 an. Beide Verletzungen der Annahmen liefern zwar nach wie vor erwartungstreue Schätzer der Regressoren, allerdings werden die Standardfehler und Irrtumswahrscheinlichkeiten verzerrt (Eid et al., 2015).

Zusammenfassend kann festgestellt werden, dass die signifikanten Wertesysteme zu einem Zuwachs des R^2_N zwischen 3.7% und 8.5% führten und die dazugehörigen Regressionsgewichte betragsmäßig zwischen 0.31 und 0.49 schwankten (siehe Tabelle 75). Der Effekt eines Wertesystems kann so interpretiert werden, dass mit der Erhöhung der Ausprägung dieses Wertesystems um eine Einheit[3] die Verkaufszahl ceteris paribus, d.h. unter Kontrolle der Betriebszugehörigkeit um $(100 \cdot e^B - 1)$ % steigt. Im Beispiel ***Gewissheit***^{A} liegt diese Steigung damit bei -26%. Beim Wertesystem ***Verstehen***^{A} z.B. geht laut dem Modell die um eine Einheit höhere Ausprägung dieses Wertesystems mit einer Steigerung der Verkaufszahlen um 49% einher.

[3]Die Einheit ist hier Logit.

10.3.2.3 Mediatoranalyse

Vergleicht man die eben berichteten signifikanten Regressionsgewichte mit den Korrelationen in Tabelle 72, fällt auf, dass die signifikanten Korrelationen der Wertesysteme ***Macht***^{A}, ***Erfolg***^{A} und ***Gewissheit***^{V} mit den Verkäufen unter Kontrolle der Betriebszugehörigkeit nicht signifikant wurden. Da eben diese Korrelationen verhältnismäßig hoch (vor allem ***Erfolg***^{A}) mit Betriebszugehörigkeit korrelieren, kann an dieser Stelle ein Mediatoreffekt von Betriebszugehörigkeit vermutet werden. Um diesen zu überprüfen wurde eine Mediatoranalyse durchgeführt. Nach Baron und Kenny (1986) und Holmbeck (1997) sind dazu vier Bedingungen zu überprüfen:

1. Der Prädiktor (Wertesystem) muss einen signifikanten Effekt auf die abhängige Variable (mittlere Verkaufszahl, VK) ausüben.
2. Der Prädiktor muss einen signifikanten Effekt auf den Mediator (Betriebszugehörigkeit in Monaten, BZM) haben.
3. Auch der Mediator muss einen signifikanten Effekt auf die abhängige Variable haben.
4. In einer multiplen Regression muss sich der Effekt des Prädiktors auf die abhängige Variable verringern, wenn der Mediator kontrolliert wird.

Tabelle 76. Modelle der Mediatoranalyse.

Schritt	AV	UV	B	SE	t	p	F-Test	R^2_N
1	BZM	ER^A	0.79	0.26	3.03	<.01	$F(1)$ = 7.5*	.16
2	VK	ER^A	1.69	0.46	3.67	<.001	$F(1)$ = 9.07**	.25
3	VK	log(BZM)	0.51	0.06	8.56	<.001	$F(1)$ = 62.12***	.67
4	VK	log(BZM)	0.52	0.07	7.38	<.001	$F(2)$ = 29.92***	.67
		+ ER^A	-0.08	0.22	0.35	.73		

Anmerkung. N = 31; ER = Erfolg; A = Annäherung; BZM = Betriebszugehörigkeit in Monaten; R^2_N = Nagelkerke's R^2; * $p < .05$; ** $p < .01$; *** $p < .001$.

Diese vier Bedingungen wurden für die Wertesysteme ***Macht***^{A}, ***Erfolg***^{A} und ***Gewissheit***^{V} durchgeführt, wobei sich einzig für ***Erfolg***^{A} ein Mediatoreffekt mit BZM zeigte. Die Ergebnisse der vier entsprechenden Schritte sind in Tabelle 76 zusammengefasst. Dabei ist zunächst global festzustellen, dass alle vier Modelle signifikante F-Tests aufweisen, also die Nullhypothese abgelehnt werden kann, dass sie nicht mehr als die Nullmodelle erklären. Heteroskedastizität und Autokorrelation der Residuen wurden in allen Modellen mit Breusch-Pagan- und Durbin-Watson-Tests überprüft. Keiner der Tests wurde signifikant. In Schritt vier wurden zudem die VIFs berechnet, die nahe bei 1 liegen, weshalb davon ausgegangen werden kann, dass keine Multikollinearität vorlag. Insgesamt konnten die Voraussetzungen für Regression somit

als erfüllt angenommen werden. Auch die Bedingungen zur Feststellung einer Mediation sind erfüllt, da ***Erfolg***^{A} einen signifikanten Einfluss auf die Verkaufszahlen (Bedingung 1) und die Dauer der Betriebszugehörigkeit (Bedingung 2) hat, die Betriebszugehörigkeit einen signifikanten Effekt auf die Verkaufszahlen aufweist (Bedingung 3) und der Effekt von ***Erfolg***^{A} verschwindet, wenn auf Betriebszugehörigkeit kontrolliert wird (Bedingung 4). Somit konnte die Hypothese bestätigt werden, dass es einen Mediatoreffekt von BZM auf den Zusammenhang zwischen ***Erfolg***^{A} und Verkaufserfolg gibt.

10.3.3 Diskussion

In dieser Untersuchung wurde sowohl die prädiktive, wie auch die inkrementelle Validität der Wertesysteme anhand von Felddaten untersucht. Bei den Felddaten handelt es sich um die durchschnittlichen Verkaufszahlen einer Stichprobe von 31 Vertriebsmitarbeitern eines Unternehmens der IT-Branche. Die prädiktive Validität wurde anhand von Korrelationen zwischen Wertesystemen und Verkaufserfolg bestimmt. Es zeigten sich signifikante Zusammenhänge zwischen Verkaufserfolg und ***Macht***^{A}, ***Erfolg***^{A} und ***Gewissheit***^{V}. Die Zusammenhänge mit den beiden erstgenannten Wertesystemen waren hypothesenkonsistent, der Zusammenhang mit ***Gewissheit***^{V} jedoch nicht erwartet.

Des Weiteren hat sich in einer Mehrebenenanalyse gezeigt, dass die vier Wertesysteme ***Gewissheit***^{A}, ***Verstehen***^{A}, ***Nachhaltigkeit***^{A} und ***Verstehen***^{V} einen zusätzlichen Teil der Varianz des Verkaufserfolgs über die Dauer der Betriebszugehörigkeit hinaus erklären können. Dabei fällt auf, dass es sich bei diesen Wertesystemen, um andere Wertesysteme handelt, als diejenigen, die mit der Zielvariable korrelieren. Es ist fraglich, ob die Verletzung der Annahmen der der Homoskedastizität in den ***Verstehen***-Modellen und der Unabhängigkeit der Fehler im ***Nachhaltigkeit***^{A}-Modell der Grund für die Signifikanzen war. Darüber hinaus wurde in einer Mediatoranalyse festgestellt, dass die Dauer der Betriebszugehörigkeit eine Mediatorrolle im Zusammenhang von ***Erfolg***^{A} und Verkaufszahlen spielt.

Die Ergebnisse bedeuten, dass manche Wertesysteme Anteile des Verkaufserfolgs von Vertriebsmitarbeitern erklären können. Darüber hinaus suggerieren die Ergebnisse der Mediatoranalyse, dass dem Wertesystem ***Erfolg***^{A} eine besondere Rolle dahingehen zukommt, dass es eine Ursache dafür sein kann, dass Mitarbeiter länger bei der Firma bleiben und als Folge daraus höhere durchschnittliche Verkaufszahlen erreichen. Dieser Zusammenhang passt auch zur weitläufigen Auffassung, dass Wertesysteme Handlungen bzw. Handlungsergebnisse nicht direkt sondern indirekt über den im Rahmen der Person-Job-Fit-Theorie Zusammenhang der Kongruenz beeinflussen (Roe & Ester, 1999). Demnach wäre die Kongruenz des ***Erfolg***-Wertesystems mit dem Vertriebsberuf die Ursache für eine längere Betriebszugehörigkeit, was mit mehr Erfahrung und einem Aufbau von Verkaufsfähigkeiten einhergehen würde und dadurch die höheren durchschnittlichen Verkaufszahlen erklären könnte.

Bedenkt man darüber hinaus die mögliche Anzahl weiterer intervenierender Variablen auf die Höhe der Verkaufszahlen, wie Fähigkeitsdispositionen, Temperamentsdispositionen, den Einfluss von Führungskräften, Gruppenprozesse innerhalb des Teams oder Branche und Standort der Firma, um nur einige wenige mögliche Einflussfaktoren zu nennen, dann können auch die tendenziell niedrigen Varianzaufklärungen im einstelligen Prozentbereich als durchaus bemerkenswert gesehen werden.

Bezogen auf die Orthogonalitätshypothese der Wertesysteme kann gesagt werden, dass zwar einige der korrelativen Maße tendenziell entgegengesetzt sind, bei den Ergebnissen der Regressionanalysen jedoch fast keine spiegelbildlichen Zusammenhänge zwischen Annäherungs- und Vermeidungsdimension der Wertesysteme aufgetreten sind. Insgesamt sprechen die Ergebnisse somit eher für Orthogonalität und gegen Bipolarität der Wertesystemdimensionen.

Bedenklich sind bei dieser Untersuchung mehrere Aspekte. Erstens kann von Streuungsrestriktionen ausgegangen werden, da die Stichprobe sehr klein und vorselektiert war (Bühner, 2011). Zum einen ergibt sich Selbstselektion daraus, dass sich Mitarbeiter vermutlich meist bewusst für diesen Job beworben haben und außerdem in einem Auswahlprozess ausgewählt wurden, von dem nicht bekannt ist, wie objektiv er war. Untersuchungen zeigen, dass die subjektive Einschätzung der Recruiter eine nicht zu unterschätzende Rolle bei der Personalauswahl spielen können (Cable & Judge, 1997).

Der zweite Punkt betrifft die Beschaffenheit der Daten. Vor allem die Variable der Dauer der Betriebszugehörigkeit war nicht normalverteilt, was zu verzerrten Ergebnissen führen kann. Auch handelte es sich bei der AV um einen Mittelwert, der nicht bei allen VPn über denselben Zeitraum gemittelt wurde. Zwar floss dies durch die Inklusion der Betriebszugehörigkeit als Prädiktor in den Regressionsanalysen zur inkrementellen Validität ein, kann jedoch nicht bei den Korrelationen berücksichtigt werden. Außerdem gibt es weitere potenzielle Störgrößen auf die Daten, wie z.B. saisonale Effekte, wie sie häufig in Vertriebsberufen auftreten.

Als dritte Einschränkung ist der geringe Umfang der Stichprobe, vor allem im Hinblick auf die multiplen Regressionsmodelle zu nennen. Im Falle zu weniger Beobachtungen können die Schätzungen der Regressionsgewichte ungenau sein (Tabachnick & Fidell, 2007). Die Ergebnisse dieser Studie sind deshalb nur mit äußerster Vorsicht zu interpretieren. Andererseits kann der Untersuchung zu Gute gehalten werden, dass in keinem Modell mehr als drei Prädiktoren spezifiziert wurden und das Verhältnis von Prädiktor zu Beobachtungen zumindest bei 1:10 lag. Auch die Tatsache, dass es sich bei den Verkaufsdaten um für Unternehmen sensible Daten handelt, kann als Argument dafür aufgeführt werden, dass trotz der geringen Stichprobe die Regressionsanalysen durchgeführt wurden. Denn der Zugriff auf solche Felddaten ist häufig schwierig, was für den vorliegenden Fall als Aufwertung der Untersuchung gewertet werden kann, da dadurch die ökologische Validität erhöht wird.

Viertens können die Daten im Hinblick auf die Zusammenhänge mit Wertesystemen durch Besonderheiten der Unternehmenskultur oder Spezifitäten der Branche geprägt sein, d.h. je

nach verkauftem Produkt können unterschiedliche Wertesysteme in unterschiedlichem Ausmaß kongruent zu den Verkaufsaufgaben sein und dementsprechend anders mit Verkaufserfolg zusammenhängen. Insbesondere in Kombination mit dem geringen Stichprobenumfang kann die Stichprobe deshalb nicht auf andere Branchen generalisiert werden.

Darüber hinaus konnte in der vorliegenden Untersuchung nicht überprüft werden, ob die Voraussetzungen für die einzelnen Mitarbeiter vergleichbar waren (z.B. Größe der Verkaufsgebiete, Anzahl der Leads, etc.). Es ist denkbar und durchaus wahrscheinlich, dass ein gewisses Maß an Kriteriumskontamination vorliegt (Bühner, 2011). Auch ist das Kriterium Verkaufserfolg gemessen an mittleren monatlichen Verkaufszahlen ein eingeschränktes Außenkriterium für den langfristigen Verkaufserfolg. Größen wie Kundenzufriedenheit, Mitarbeiterzufriedenheit oder Commitment wurden nicht erhoben. Dies wären Variablen, die in Folgeuntersuchungen untersucht werden sollten, denn auch sie könnten signifikante Effekte auf den Verkaufserfolg haben.

Abschließend können mehrere Schlussfolgerungen für zukünftige Untersuchungen gezogen werden. Zum einen können die Ergebnisse als vielversprechend gesehen werden und sollten weitere Untersuchungen in diese Richtung motivieren. Dabei erscheint es jedoch empfehlenswert einerseits unterschiedliche Branchen und Unternehmensgrößen abzudecken und andererseits größere Stichprobenumfänge anzustreben. Außerdem sollte das Außenkriterium Verkaufserfolg durch weitere Variablen wie Kunden- und Mitarbeiterzufriedenheit erweitert werden. Auch die Rolle intervenierender Variablen, wie z.B. Führungsstil und Anreizsysteme stellen eine interessante Erweiterung des Forschungsdesigns dar.

Unter Berücksichtigung des Ergebnisses, dass die Betriebszugehörigkeit als Mediatorvariable zwischen Verkaufserfolg und dem Wertesystem $\boldsymbol{Erfolg}^{A}$ fungierte, bietet es sich des Weiteren an, Komponenten wie Erfahrung und Fähigkeit zu erheben, da diese vermutlich einen hohen Einfluss auf den Verkaufserfolg haben. Dadurch könnte zudem empirisch überprüft werden, in welchem Ausmaß eine erbrachte Leistung vom Faktor Motivation (d.h. Werte-Kongruenz) und Fähigkeit abhängt und möglicherweise weitere Rückschlüsse auf deren Zusammenspiel gezogen werden (Rosenstiel, 2007). Insgesamt kann somit gesagt werden, dass es sich um einen vielversprechendes Forschungsdesign handelt.

Kapitel 11

Diskussion

Das Hauptziel dieser Arbeit lag in der Untersuchung der psychometrischen Güte des MVSQ anhand der Gütekriterien der Objektivität, Reliabilität und Validität. Zudem ergaben sich aufgrund von Format und Konzeptualisierung des Instruments weitere Untersuchungsbereiche. Diese umfassen erstens die Schätzung von Thurstonian-IRT-Modellen an Skalen mit sieben Items pro Block und die damit einhergehende Suche nach einer passenden Einstellung des Schätzalgorithmus'. Zweitens trat die Frage nach der Sinnhaftigkeit der Orthogonalitätshypothese, also der Konzeptualisierung der Wertesysteme auf den zwei Dimensionen Annäherung und Vermeidung auf. Im Folgenden werden die Hauptergebnisse zu den einzelnen Fragestellungen kurz zusammengefasst und hinsichtlich ihrer praktischen Relevanz diskutiert. Darüber hinaus werden methodische Grenzen der Arbeit aufgezeigt und ein Ausblick auf zukünftige Forschung gegeben.

11.1 Objektivität und Reliabilität

Die Untersuchung zur Objektivität zeigte, dass Durchführungs, Auswertungs- und Interpretationsobjektivität als gegeben angesehen werden können. Einzig bei der Interpretationsobjektivität kann bemängelt werden, dass keine veröffentlichten Informationen zur empirischen Basis der Interpretationsrichtlinien vorliegen. Die Befunde konnten dennoch als hinreichende Voraussetzung zur Untersuchung der Reliabilität gewertet werden.

Um die Reliabilität des MVSQ bestimmen zu können, erwiesen sich die gängigen klassischen wie probabilisitischen Testmodelle als ungeeignet, da das Forced-Choice-Format des MVSQ ipsative Daten produziert, wodurch die Grundannahme der Unabhängigkeit der Fehler verletzt ist. Zum Zeitpunkt der Arbeit gab es lediglich einen testtheoretisch korrekten Ansatz, mit dem *dominant* formulierte Forced-Choice-Fragebögen – wie der MVSQ – modelliert werden können (Brown, 2014). Dabei handelt es sich um die Thurstonian Item-Response-Theorie (Brown & Maydeu-Olivares, 2011).

Entsprechend der TIRT wurden zwei Modelle an die beiden Subskalen des MVSQ angepasst und in einer simulationsbasierten Herangehensweise die empirischen Reliabilitäten (ρ) bestimmt (Kapitel 7). Diese lagen im Mittel bei ρ = .69 und bewegten sich zwischen .60 und .77. Da es sich bei Wertesystemen um konzeptuell vergleichsweise breite Konstrukte handelt (vgl. Kapitel 2.1) und sich die Reliabilitäten vergleichbarer Instrumente wie dem Schwartz Value Survey in ähnlichen Größenordnungen bewegen (vgl. Kapitel 3.6.2.2), können zwölf der 14 vorliegenden Reliabilitäten als akzeptabel (> .65) eingestuft werden. Die verbleibenden zwei Reliabilitäten zeigen den Verbesserungsbedarf des Fragebogens an.

Zudem ist zur Vorgehensweise bei der Bestimmung der empirischen Reliabilitäten zu sagen, dass es sich dabei um ein vergleichsweise aufwändiges Verfahren handelt. Problematisch daran könnte sein, dass fortgeschrittene Kenntnisse in der Bedienung von entsprechenden Statistikprogrammen (M*plus* oder R) erforderlich sind, ebenso wie ein vertieftes Verständnis der zugehörigen Testtheorie, was eine Hürde für die Anwendung bzw. Verbreitung des Verfahrens darstellen könnte. Da es sich dabei allerdings um die einzige Möglichkeit handelt, einen Reliabilitätskoeffizienten für Forced-Choice Fragebögen zu berechnen, bei dem *eine* Testadministration ausreicht, erscheint es ein lohnenswertes Ziel, die Anwendung durch Neu- oder Weiterentwicklung entsprechender Programme zu vereinfachen.

Im Verlauf der Anfertigung dieser Arbeit wurde der MVSQ von den Entwicklern überarbeitet und für beide Versionen die zugehörigen empirischen Reliabilitäten bestimmt und miteinander verglichen (vgl. Kapitel 8). Dieser Vergleich hat gezeigt, dass die Überarbeitung im Durchschnitt zu einer Verbesserung der Messgenauigkeit geführt hat. Dabei haben sich vor allem die Wertesysteme ***Erfolg***^{A} ($\Delta\rho = .19$), ***Erfolg***^{V} ($\Delta\rho = .09$) und ***Verstehen***^{V} ($\Delta\rho = .12$) deutlich verbessert, allerdings auch die Reliabilitäten von ***Gleichheit***^{A} ($\Delta\rho = -.09$), ***Geborgenheit***^{A} und ***Geborgenheit***^{V} (beide $\Delta\rho = -.05$) merklich verschlechtert. Diese Veränderungen zeigen Verschiedenes: Erstens kann daraus geschlossen werden, dass deutliche Verbesserungen einzelner Reliabilitäten in Forced-Choice-Fragebögen möglich sind. Da sich jedoch auch Verschlechterungen einzelner Reliabilitäten ergeben haben, kann daraus als zweites abgeleitet werden, dass die klassische Itemanalyse (vgl. Kapitel 5), nur bedingt geeignet war, um darauf basierend die verbesserungswürdigen Items auszuwählen. Drittens zeigt diese Entwicklung, dass es sich bei der Revision von Forced-Choice-Fragebögen grundsätzlich – auf Grund der zahlreichen Interdependenzen zwischen den Items innerhalb der Blöcke – um ein komplexes Unterfangen handelt.

In einer weiteren Untersuchung wurden akzeptabel bis gute Test-Retest-Reliabilitäten (r_{tt}) festgestellt, die im Durchschnitt bei r_{tt} = .76 lagen und zwischen r_{tt} = .65 und r_{tt} = .87 schwankten. Da das Intervall zwischen erster und zweiter Testung mehr als zehn Wochen betrug, kann gesagt werden, dass Wertesysteme als zumindest mittelfristige stabil gesehen werden können. Diese Aussage steht in Einklang mit bestehender Forschung, wonach es sich bei Werten um zeitlich stabile Konstrukte handelt (Jin & Rounds, 2012).

Bzgl. der Untersuchung der Retest-Reliabilitäten kann kritisiert werden, dass es sich um eine relativ kleine (N = 62), studentische Stichprobe handelte und Verallgemeinerbarkeit der Ergebnisse deshalb nur eingeschränkt gelten kann. Zudem erhielten alle Teilnehmer jeweils kurz nach der ersten Testung Feedback zu ihren Ergebnissen. Zwar können Erinnerungseffekte bzgl. der Items aufgrund des langen Intervalls nahezu ausgeschlossen werden, allerdings kann nicht gesagt werden, ob das Wissen über das Wertemodell und das eigene Ergebnis einen Einfluss auf die zweite Testung hatte. Damit die Möglichkeit besteht, dass letztgenanntes zutrifft, müsste der MVSQ eine erhöhte Augenscheinvalidität besitzen. Für zukünftige Untersuchungen der Retest-Reliabilität können deshalb mehrere Empfehlungen abgegeben werden: größere Stichproben, unterschiedliche Zeitintervalle sowie eine Untersuchung ohne Ausgabe des Feedbacks. Auch eine Langzeitstudie im Hinblick auf die Veränderbarkeit von Wertesystemen im Lebensverlauf stellt eine interessante Fragestellung dar.

11.2 Validität

Die Validität des MVSQ wurde aus unterschiedlichen Perspektiven beleuchtet. In Kapitel 9.1 wurde zunächst die faktorielle Validität begutachtet, die zumindest für sieben Wertesysteme unter Vernachlässigung der Aufteilung in Annäherungs- und Vermeidungsdimension als gegeben, wenngleich mit Verbesserungsbedarf gewertet werden kann. Danach wurde die Konstruktvalidität durch Untersuchungen zur konvergenten (Kapitel 9.2) und divergenten (Kapitel 9.3) analysiert. In der Studie zur konvergenten Validität wurde verifiziert, dass es sich bei Wertesystemen um inhaltlich ähnliche Konstrukte zu einigen der Werte-Typen in der viel beforschten und weit verbreiteten Wertetheorie von Schwartz handelt. Diese Ergebnisse können als affirmativ bzgl. der Konzeptualisierung der Wertesysteme gewertet werden. Weiterhin kann aus einer integrativen Perspektive an dieser Stelle empfohlen werden, die Wertesysteme in Bezug zu weiteren ähnlichen Konstrukten zu setzen. Zum Beispiel sind inhaltliche Überschneidungen der Wertesysteme mit den Motiven Macht, Leistung und Anschluss alleine schon aufgrund der verwendeten Begrifflichkeiten zu erwarten, zumal Motive und Werte auch konzeptuell als sehr ähnlich gesehen werden können (Locke & Henne, 1986). Hierfür könnte das Multitrait-Multimethod-Design verwendet werden und Ergebnisse nicht nur im Zusammenhang der Konstruktvalidierung des MVSQ, sondern vor allem auch im Sinne der Weiterentwicklung einer integrativen Theorie motivationspsychologischer Konstrukte gesehen werden (Locke & Latham, 2004). Auch im Zusammenhang mit Klassifizierungen von Interessen (Larson et al., 2002) wären weiterführende Untersuchungen erstrebenswert, auch um Gemeinsamkeiten und Unterschiede dieser Konstruktarten auf inhaltlicher Basis voranzubringen. In der Untersuchung zur divergenten Validität wurde zudem die Hypothese bekräftigt, dass es sich bei Wertesystemen um motivationale Konstrukte handelt, die sich gemäß der Eigenschaftstheorie von Cattell (1973) von kognitiven Dispositionen und Temperamtensdispositionen unterscheiden.

In einer anschließenden Reihe von Untersuchungen zur Kriteriumsvalidität wurden schließlich solide Hinweise für die kriteriumsbezogene Validität festgestellt. In der Untersuchung zur konkurrenten Validität (Kapitel 10.1) erwiesen sich Wertesysteme als charakteristische Kriterien dafür, Gruppen von Personen je nach Zugehörigkeit zu einem Studiengang, Studiengangsschwerpunkt, Aufgabenbereich und Hierarchieebene zu beschreiben. Abgesehen davon, dass diese Befunde für die kriteriumsbezogene Validität der Wertesysteme sprechen, sind diese Ergebnisse in vielerlei Hinsicht von Relevanz. Zum einen bestätigen sie die Bedeutung von Wertesystemen in der Konzeptualisierung von Person-Organisation-Fit und Person-Job-Fit (Kristof-Brown et al., 2005). Zum anderen leisten sie dahingehend einen relevanten Beitrag, als dass sie einem relativ unerforschten Wertemodell Eingang in dieses Forschungsgebiet geben. Bestehende Forschung gestaltet sich vor allem dahingehend, den Einfluss von Wertekongruenz auf andere Konstrukte wie z.B. Commitment, Zufriedenheit und Leistung (vgl. Meta-Analyse von Verquer et al., 2003), Stress (Borg et al., 2011) oder Wohlbefinden (Sagiv & Schwartz, 2000) zu untersuchen. Die Unterscheidung von (Gruppen von) Personen anhand charakteristischer Wertesysteme scheint dabei neu, wobei das hier verwendete Wertemodell dafür einen vielversprechenden Rahmen darstellt.

Auch für praktische Fragestellungen können diese Befunde von Bedeutung sein. So kann z.B. der Bezug zur Studiengangsberatung hergestellt werden, indem Wertesystem-Studiengangs-Kongruenz als denkbares Kriterium dafür verwendet werden könnte, um Empfehlungen zur Studiengangswahl abzuleiten. Ebenso könnte logischerweise Empfehlungen zu passenden Studiengangsschwerpunkten und Aufgabenbereichen in Fragen der Berufsorientierung hergeleitet werden. Voraussetzung dafür ist allerdings, dass die Stichprobe deutlich vergrößert wird. Wünschenswert wäre des Weiteren die Verknüpfung Wertesystem-Kongruenzen mit Daten über Zufriedenheit und Wohlbefinden in den entsprechenden Umfeldern. Zwar sprechen empirische Befunde für diesen Zusammenhang (Sagiv & Schwartz, 2000; Verplanken, 2004), allerdings sollte dieser für den MVSQ gesondert untersucht werden. Daraus könnten des Weiteren auch die Interpretationsrichtlinien erweitert und validiert werden.

Bzgl. der prädiktiven Validität (Kapitel 10.2) wurde in einer Experimentalstudie untersucht, ob die Kongruenz von Wertesystemen mit einer Aufgabe die Intensität der Motivation – operationalisiert als Flow, intrinsische Motivation und Aktivierung – beeinflusst. Dazu wurden drei Wertesystem-Aufgaben-Kongruenzen untersucht, von denen sich in zweien signifikant und in einer zumindest tendenziell hypothesenkonsistente Ergebnisse zeigten. Es kann deshalb gesagt werden, dass die Studie wertvolle Hinweise auf die prädiktive Validität dreier Wertesysteme lieferte. Darüber hinaus können die Ergebnisse als Bestätigung der Hypothese gesehen werden, dass Wertesysteme die Richtung der Motivation steuern können und dadurch die Intensität der Motivation beeinflussen.

Kritisch zu sehen ist, dass fraglich ist, ob bzw. in welchem Maß die vom Autor konzipierten Aufgaben kongruent zu den entsprechenden Wertesystemen waren und ob die Ergebnisse

der Studie folglich generalisierbar sind, d.h. externe Validität besitzen (Eid et al., 2015). Zwar kann die Tatsache, dass hypothesenkonsistente Befunde erzielt wurden, im Sinne der Generalisierbarkeit gewertet werden, dennoch stellt sich die Frage, inwiefern 15 Minuten der Aufgabenbearbeitung ausreichen, um eine Übertragung der Befunde auf natürliche Situationen zu erlauben. Um diese Frage zu beantworten, empfiehlt es sich in einer Folgestudie längere Bearbeitungszeiten der Aufgaben zu implementieren, um dadurch auch die externe Validität des Designs zu erhöhen. Die Generalisierbarkeit betreffend eröffnen sich ferner Fragen danach, welche Aufgaben mehrheitlich in einer Erwerbsbevölkerung bearbeitet werden, wie diese adäquat im Labor nachgebildet werden können und ob diese durch das Wertemodell angemessen charakterisiert und kategorisiert werden können?

Ferner hat sich in dieser Studie gezeigt, dass ein monetärer Anreiz in Abhängigkeit des Wertesystems ***Verstehen*** einen negativen Einfluss auf die Motivationsintensität haben kann. Dieses Ergebnis steht zwar im Einklang mit bestehender Forschung (Frey & Osterloh, 2005), allerdings muss dem der Zusammenhang beim ***Erfolg***-Wertesystem und der entsprechenden kongruenten Aufgabe gegenüber gestellt werden. Hier zeigten sich nur in der experimentellen Bedingung mit monetärem Anreiz die hypothetisierten Zusammenhänge, was bedeutet, dass der monetäre Anreiz eine Art Hygiene-Faktor darstellte und der herrschenden Meinung, dass monetäre Anreize vor allem in negativem Zusammenhang mit Motivation stehen (Furnham, 2014), kollidiert. Weitere Studien sind hier unbedingt notwendig, um dieser Fragestellung nachzugehen. Kritisiert werden kann in diesem Zusammenhang, dass der monetäre Anreiz als experimentelle Bedingung nicht in einer Vorstudie untersucht wurde. Zwar waren die zu gewinnenden Geldbeträge (50, 25 und 15 Euro) in einer für Studierende relevanten Größenordnung, wenn man zum Vergleich ein durchschnittliches monatlich verfügbares Nettoeinkommen von 215 Euro nimmt (Quelle www.statista.de). Es bleibt jedoch unklar, ob auch Effekte bei anderen Geldbeträgen zu erwarten sind.

Des Weiteren kann an dieser Untersuchung kritisch gesehen werden, dass direkt von der Wertesystem-Aufgaben-Kongruenz auf die Motivationsintensität geschlossen wurde, obgleich unter Motivationswissenschaftlern die Ansicht etabliert ist, dass die Beziehung zwischen Werten und Handlung (inklusive Empfindungen) von Zielen moderiert wird (Locke, 1991, 1997). Es ist ferner denkbar, dass auch Einstellungen und Interessen eine interagierende Wirkung haben können. Für zukünftige Untersuchungen sollten diese Konstrukte deshalb bedacht und in einer geeigneten Form in Experimentaldesigns berücksichtigt werden, indem z.B. Zielspezifität und -Komplexität als Varibalen einbezogen werden und abgefragt wird, wie die Einstellung gegenüber den entsprechenden Einstellungsobjekten (z.B. Aufgaben oder Anreiz) ist. Dennoch kann aus der methodischen Perspektive gesagt werden, dass die wesentlichen Bedingungen der internen Validität der Untersuchung erfüllt waren, denn die Erhebung der unabhängigen Variablen ging der Messung der abhängigen Variablen zeitlich voraus, die verwendeten Messinstrumente waren hinreichend reliabel, die Bedingungszuweisungen waren randomisiert und der kausale

Zusammenhang zwischen Wertesystem-Aufgaben-Kongruenz und Motivationsintensität kann nicht plausibel auf andere Variablen zurückgeführt werden (Shadish et al., 2002).

Darüber hinaus kann bei der Operationalisierung der Motivationsintensität hinterfragt werden, ob die gewählten Konstrukte (Flow, intrinsische Motivation und Affekt/Aktivierung) adäquat waren. Zwar besteht Klarheit, dass es sich bei Motivationsintensität um eine Erlebensqualität handelt, unklar ist jedoch, welches Konstrukt diesen psychologischen Zustand am adäquatesten bzw. welche Facetten davon abbildet. Eventuell ist an dieser Stelle auch die Entwicklung einer neuen Skala angebracht, in der gezielter die Intensität oder Qualität des Erlebens abgefragt wird. Wie dem auch sei hat sich die Verwendung von Kurzskalen zur Erhebung der Konstrukte bewährt, da diese erstens hohe Reliabilitäten bei den unterschiedlichen Aufgaben aufwiesen und zweitens signifikante Unterschiede bei den Aufgaben anzeigten.

Zum Abschluss sei gesagt, dass bezogen auf die verwendete Konzeptualisierung der Motivation die dritte Dimension, nämlich die Frage nach der Persistenz nicht untersucht wurde. Dies kann damit gerechtfertigt werden, dass es den Umfang dieser Arbeit überstiegen hätte. In zukünftigen Untersuchungen könnte diese Dimension jedoch zweifelsohne zu weiteren aufschlussreichen Erkenntnissen führen und stellt vor allem im Hinblick auf die praktische Relevanz einen großen Anreiz dar, berücksichtigt zu werden.

Bei der letzte Studie dieser Arbeit (Kapitel 10.3) handelte es sich um eine Feldstudie, in der sowohl die prädiktive als auch die inkrementelle Validität untersucht wurde. Die Wertesystemausprägungen wurden dabei in Zusammenhang mit der Leistung von Vertriebsmitarbeitern in Form des Verkaufserfolgs gesetzt, wobei drei Wertesysteme einen moderaten positiven Zusammenhang (zwischen $r = .36$ und $.41$) mit den durchschnittlichen Verkaufszahlen zeigten. Da die Wertesysteme vor Erhebung der Verkaufszahlen gemessen wurden, sind dieses Ergebnisse der prädiktiven Validität zuzuordnen.

Ferner wurden mehrere hierarchische Regressionsanalysen durchgeführt, in denen sich gezeigt hat, dass vier der Wertesysteme unter Kontrolle der Dauer der Betriebszugehörigkeit einen zusätzlichen Anteil der Varianz der Verkaufszahlen erklärten, also inkrementelle Validität besitzen. Die zusätzlichen Varianzanteile lagen dabei zwischen 3.7% und 8.5%. Zu dieser Untersuchung sei zudem gesagt, dass weitere Prädiktoren untersucht wurden, die jedoch keinen signifikanten Einfluss auf den Verkaufserfolg hatten. Diese waren Alter, Geschlecht und Berufserfahrung.

Des Weiteren wurde mit den Verkaufsdaten eine Mediatoranalyse durchgeführt, in der ein Mediatoreffekt zwischen der Betriebszugehörigkeit und dem Wertesystem ***Erfolg***^{A} festgestellt wurde. Je höher dieses Wertesystem ausgeprägt war, umso höher war der Mediator Betriebszugehörigkeit, der einen hochsignifikanten Effekt auf den Verkaufserfolg hatte und unter Kontrolle den Zusammenhang zwischen ***Erfolg***^{A} und Verkaufserfolg verringerte. Dieser Befund lässt sich durch den Person-Job-Fit-Ansatz erklären, wobei die logische Argumentationskette wie folgt lautet: Die Kongruenz von ***Erfolg***^{A} und Vertriebsaufgaben (siehe Hypothesen

Kapitel 10.1) führt zu einer erhöhten Dauer der Betriebszugehörigkeit, da Zufriedenheit und Wohlbefinden von der Wertesystem-Kongruenz beeinflusst werden. Die dadurch steigende Erfahrung geht mit einem Aufbau von Fähigkeiten und Fertigkeiten einher, wodurch wiederum die Verkaufszahlen steigen, also die Leistung größer wird.

Zusammenfassend kann zu den Ergebnissen der Feldstudie gesagt werden, dass die Effekte zwar absolut gesehen nur im niedrigen bis moderaten Bereich lagen. Unter Berücksichtigung der Tatsache, dass zahlreiche Variablen den Zusammenhang zwischen Wertesystemen und Verkaufserfolg beeinflussen können, sind die beobachteten Koeffizienten als durchaus gewichtig einzustufen. Zwar handelte es sich um eine Stichprobe geringen Umfangs, dennoch können die Ergebnisse als vielversprechend gesehen werden und ermutigen dazu, weitere Untersuchungen im Feld durchzuführen. Denn gerade die Forschung im Feld stellt die Basis für die Bestätigung der ökologischen Validität dar. Diesbezüglich können neben den Ergebnissen der Feldstudie auch die Befunde der Analyse der konkurrenten Validität aufgeführt werden, denn auch darin wurden die Kriteriumsvariablen als Felddaten erhoben, die somit die Lebensumwelt als Ganzes repräsentieren (Lewin, 1939) und den Rückschluss erlauben, dass die gefundenen Zusammenhänge zwischen Wertesystemen und Kriteriumsvariablen auch im Alltagsgeschehen Gültigkeit besitzen (Eid et al., 2015).

Um die Ergebnisse zur Validität in einem Fazit zusammenzufassen, kann gesagt werden, dass in mehreren Untersuchungen zur Konstrukt- sowie Kriteriumsvalidität vielversprechende und empirisch begründete Hinweise auf die Gültigkeit der im MVSQ gemessenen Wertesysteme gefunden werden konnten. Insbesondere die Befunde zur Kriteriumsvalidität zeugen von hoher praktischer Relevanz, wohingegen die Resultate der Konstruktvalidierung für die inhaltliche Validität der Wertesysteme sprechen. Dennoch sollten in allen Bereichen weiterführende Studien durchgeführt werden, einerseits, um die hier berichteten Ergebnisse zu replizieren und andererseits, die Bandbreite der Gültigkeit zu erhöhen.

11.3 Orthogonalitätshypothese der Wertesysteme

In diesem Abschnitt erfolgt nun eine Zusammenfassung der Befunde zur Frage, ob es sich bei den Wertesystemdimensionen Annäherung und Vermeidung um orthogonal oder bipolar zueinander stehende Dimensionen handelt. Tabelle 77 gibt einen Überblick der in dieser Arbeit gesammelten Befunde.

Insgesamt können acht Analysen dazu herangezogen werden können, um Rückschlüsse auf die Dimensionalität der Wertesysteme zu ziehen. Zwei dieser Analysen ergaben ein relativ eindeutiges Bild, nämlich die Ergebnisse der explorativen Faktorenanalyse und die Laborstudie zur Motivationsintensität. Beide Studien zeigen recht eindeutig, dass es sich bei Annäherung und Vermeidung um die entgegengesetzten Enden *einer* Dimension handelt. Weniger eindeutig, aber noch tendenziell zu Gunsten der Bipolaritätsannahme können die Skaleninterkorrelationen,

Tabelle 77. Überblick der Ergebnisse zur Orthogonalitätshypothese.

Analyse	Ergebnis	Tendenz	Kapitel
Skaleninterkorrelationen	4 Wertesysteme bipolar, 2 orthogonal	Bipolarität	9.1.2.1
Explorative Faktorenanalyse	7 Wertesysteme bipolar	Bipolarität	9.1.2.2
bipolare TIRT-Modelle und Korrelationen der Scores	4 bipolar, 2 orthogonal	Bipolarität	9.1.2.3
Geschlechterunterschiede	5 bipolar, 2 orthogonal	Bipolarität	10.1.2.1
Wertesystem-Kongruenz zu Studiengängen und Aufgabenbereichen	uneinheitlich	Orthogonalität	10.1.2.2
Wertesystem-Kongruenz zu Aufgabenbereichen	uneinheitlich	Orthogonalität	10.1.2.3
Motivationsintensität	7 bipolar	Bipolarität	10.2
Verkaufserfolg	uneinheitlich	Orthogonalität	10.3

die Korrelationen zwischen Annäherungs-, Vermeidungs und die bipolaren Scores sowie die Ergebnisse der Untersuchung der Geschlechterunterschiede gesehen werden. Dabei wirken jeweils eine Mehrheit der Wertesysteme, bipolar anstatt orthogonal zu sein. Uneinheitlich sind die Befunde der Untersuchung der konkurrenten Validität in Form der Wertesystem-Kongruenzen zu Studiengängen und Aufgabenbereichen, ebenso wie die Ergebnisse zum Zusammenhang von Wertesystemen mit Verkaufserfolg.

In Summe kann somit zwar eine leichte Tendenz hin zur Bipolarität konstatiert werden, ohne dass jedoch ein abschließendes Urteil über Annahme oder Verwerfung der Orthogonalitätshypothese gefällt werden kann. Deshalb besteht für diese Fragestellung weiterer Forschungsbedarf, wobei als erster Schritt eine Verbesserung der Messgenauigkeit des Instruments angestrebt werden sollte, da die Interkorrelationen der Dimensionen vom Messfehler abhängen (Russell & Carroll, 1999). Dadurch können mögliche Verzerrungen, die darauf zurückgehen, reduziert werden. Zum Schluss sei angemerkt, dass auch denkbar ist, dass Annäherungs- und Vermeidungsdimension in Wahrheit weder bipolar noch orthogonal, sondern multipolar sind. Diese These bedarf jedoch neben empirischer Überprüfung auch die Entwicklung theoretischer Konzeptualisierungen über die Natur der Annäherungs- und Vermeidungsdimension in der Motivation.

11.4 Methodische Aspekte

Bei den methodischen Aspekten sind zwei Punkte zu adressieren. Zum einen die Anwendung der Thurstonian Item-Response-Theorie inklusive Auswirkungen und zum anderen die Beschaffenheit der in dieser Arbeit verwendeten Stichproben.

Thurstonian Item-Response-Theorie

Wie beschrieben wurden in dieser Arbeit mehrere Thurstonian Item-Response-Theorie (TIRT) Modelle an die ipsativen Daten des MVSQ angepasst. Dazu ist zu sagen, dass es sich beim TIRT-Ansatz um eine sehr junge Theorie handelt, die erst 2011 von Brown und Maydeu-Olivares veröffentlicht wurde und zu der deshalb wenige Anwendungsbeispiele existieren. Der Hauptgrund könnte darin liegen, dass es sich dabei um ein theoretisch, wie praktisch anspruchsvolles Verfahren handelt, das nicht nur ein vertieftes Verständnis der klassischen und probabilistischen Testtheorie erfordert, sondern auch Zugang und Anwendung entsprechender Statistik-Programme. Bei der Anwendung mit den im M*plus* implementierten Schätzern (DWLS und MAP, vgl. Kapitel 6.1.1) lag das Problem darin, dass nur Modelle mit Blockgrößen von maximal vier Items pro Block verzerrungsfrei geschätzt werden können. Bei der Umsetzung mit dem R-Paket bestand die Schwierigkeit darin, geeignete Tuning-Parameter für den Shrinkage-Mechanismus des Schätzers (MSS, vgl. Kapitel 6.1.5) zu ermitteln. Dies war jedoch erforderlich, um ein TIRT-Modell mit sieben Items pro Block zu schätzen. Das High Performance Computing Cluster der Universität Regensburg[1] hat sich dabei als äußerst hilfreiches Werkzeug erwiesen, da dadurch bis zu 64 Schätzvorgänge parallel durchgeführt werden konnten und somit die reine Rechenzeit (von mehr als 500 Tagen für die Ergebnisse dieser Arbeit) um den Faktor 64 verringert werden konnte.

Des Weiteren können die in dieser Arbeit vorgenommenen Modellschätzungen als wertvoller Beitrag für die Anwendbarkeit der TIRT gesehen werden, da bis kurz vor Druck dieser Arbeit keine vergleichbaren Studien gefunden werden konnten. An dieser Stelle ist es erstens angebracht, sowohl den Nutzen der von Brown und Maydeu-Olivares entwickelten Thurstonian Item-Response-Theorie hervorzuheben, die es ermöglicht, Forced-Choice-Daten von den Restriktionen der Ipsativität zu befreien und interindividuell vergleichbare Scores zu schätzen. Dies stellt wiederum die unbedingte Voraussetzung dafür dar, die Reliabilität und Validität von Forced-Choice-Fragebögen wissenschaftlich fundiert zu untersuchen. Zweitens sei auch die Praktikabilität des `kcirt`-Pakets zu betonen, das sich als hilfreiche und fähige Software erwiesen hat, um ein komplexes Testmodell, wie das für den MVSQ notwendige, zu schätzen.

[1] `http://www.uni-regensburg.de/rechenzentrum/it-services/scientific-computing/index.html`

Stichproben

Insgesamt flossen in dieser Arbeit Daten von 2045 Personen ein, die in drei große Stichproben aufgeteilt wurden und von denen knapp 58% Berufstätige waren. Der Grund für die Dreiteilung lag darin, dass die Schätzung von TIRT-Modellen mit zunehmender Stichprobengröße zuverlässiger wird und Stichproben von wenigen Hundert Personen nicht geeignet waren, um TIRT-Modelle inklusive der dazugehörigen Scores zu schätzen. Als Konsequenz kann in Frage gestellt werden, inwiefern die Scores je innerhalb der drei großen Stichproben tatsächlich unabhängig voneinander waren. Für die Unabhängigkeit der Scores sprechen allerdings die Ergebnisse der Untersuchung zur Plausibilität der TIRT-Scores aus Stichprobe II (vgl. Kapitel 6.2.3), die gezeigt haben, dass die Eigenschaften (Verteilung, Zusammenhänge mit KTT-Scores und Skaleninterkorrelationen) der TIRT-Scores ähnlich zu Eigenschaften von TIRT-Scores waren, die Brown und Maydeu-Olivares (2011) berichteten.

Kritisiert werden kann des Weiteren, dass dieselben Personen in mehreren Untersuchungen einflossen und somit nicht für jede Untersuchung eine unabhängige Zufallsstichprobe verwendet wurde. Dem kann gegenübergestellt werden, dass die es sich bei der Mehrheit der Stichprobe um Berufstätige handelte, die allgemein schwieriger zu akquirieren sind als Studierende, weswegen an dieser Stelle die üblicherweise in psychologischen Untersuchungen herrschende Überrepräsentation von Studierenden (Henrich et al., 2010) nicht zutrifft. Die Generalisierbarkeit ist zwar aufgrund der Mehrfach-Verwendung der Daten eingeschränkt, jedoch sind einige der Befunde nicht auf die Grundgesamtheit der Studierenden beschränkt, sondern betreffen ebenso den berufstätigen Teil der Bevölkerung. In zukünftigen Untersuchungen sollten die hier ermittelten Ergebnisse in unabhängigen Stichproben repliziert werden.

Zum Schluss sei darauf hingewiesen, dass auch für die in dieser Arbeit verwendeten Stichproben gilt, dass es sich darin um Testpersonen handelt, die der sogenannten WEIRD-Population (Western, Educated, Industrialized, Rich and Democratic, Henrich et al., 2010) zugeordnet werden müssen. Zwar gibt es Hinweise darauf, dass es universell gültige Aspekte in der Natur von Wertesystemen gibt (Schwartz, 1992, 1994), jedoch muss diese Universalitätshypothese für den MVSQ separat untersucht werden. Dies könnte mittelfristig in entsprechenden internationalen Forschungsprojekten untersucht werden.

11.5 Theoretischer und praktischer Nutzen

Zunächst kann gesagt werden, dass in dieser Arbeit das Wertesystem-Konstrukt einer tiefgehenden theoretischen Untersuchung unterzogen wurde. Darin wurden theoriegeleitet die Unterschiede zu und Zusammenhänge mit zentralen Konstrukten der Motivationspsychologie, wie Bedürfnissen, Interessen, Zielen, Einstellungen und Motiven erarbeitet. Diese Abhandlung hat zum Einen die Bedeutung von Wertesystemen für die Beantwortung von Fragen

der Motivation gezeigt und kann zum Anderen als Beitrag für der Weiterentwicklung der Theorie von Wertesystemen gesehen werden. Abgesehen davon können die theoretischen formulierten Zusammenhänge dazu verwendet werden, um weitere Fragestellungen für die empirische Forschung abzuleiten. So könnten in zukünftigen Experimentalstudien z.B. die unterschiedlichen zeitlichen Stabilitäten sowie die unterschiedlichen Abstraktionsgrade der genannten Konstrukte empirisch untersucht werden.

Darüber hinaus haben einige der Ergebnisse dieser Arbeit hohe Relevanz für die tägliche Berufspraxis. Alle drei Untersuchungen der Kriteriumsvalidität haben gezeigt, dass es signifikante Zusammenhänge zwischen Wertesystemen und Bereichen gibt, die für die Arbeit von Personalern relevant sind. So wurden einerseits Zusammenhänge zwischen Wertesystempräferenz und Aufgabenbereichen sowie Hierarchieebenen festgestellt. Andererseits erklären Wertesysteme Motivationsempfinden und Leistung. Sie können deshalb als psychologische Merkmale gesehen werden, dessen Berücksichtigung bei der Personalauswahl und Personalentwicklung lohnende Effekte haben kann. Denn wenn die Wertesysteme der Mitarbeiter zum Job passen, erhöht sich nicht nur die Wahrscheinlichkeit, dass diese Mitarbeiter motivierter bei der Arbeit sind, sondern auch, dass sie länger in den entsprechenden Jobs bleiben. Je länger ein Mitarbeiter in einem Job arbeitet, umso eher wird er entsprechende Fähigkeiten entwickeln um die Leistung zu verbessern. Dadurch kann nicht nur insgesamt die Produktivität erhöht, sondern auch die Mitarbeiterfluktuation verringert werden.

Über Personalauswahl und -entwicklung hinaus kann die Berücksichtigung der Wertesysteme aus denselben Überlegungen auch für die unternehmensinterne Neuorganisation von Aufgaben und Zusammensetzung von Teams verwendet werden. Insofern sind die Ergebnisse nicht nur für Personaler relevant, sondern auch für Führungskräfte. Auch das Ergebnis, dass der monetäre Anreiz den Zusammenhang von Motivationsempfinden und Wertesystemausprägung moderiert, ist von hoher Relevanz. Insbesondere für die Führung von Teams in Forschung und Entwicklung, die überdurchschnittlich durch das Wertesystem ***Verstehen*** charakterisiert werden können, ist dieser Zusammenhang von Bedeutung, denn Geld als Incentive zeigt hier einen kontraproduktiven Effekt. Für Aufgaben der Gewinnmaximierung, wie z.B. in Vertriebsaufgaben, die besonders durch das Wertesystem ***Erfolg*** gekennzeichnet sind, gilt das Gegenteil. Hier stellt der monetäre Anreiz eine Grundvoraussetzung für Motivation dar.

Ganz allgemein zeigen die Befunde, dass Menschen je nach Wertesystemausprägungen unterschiedliche Empfindungen bei der Bearbeitung derselben Aufgabe haben. Aus einer humanistischen Haltung heraus kann somit geschlussfolgert werden, dass die Beachtung von Wertesystempräferenzen allein aus dem Grund anstrebenswert ist, da dadurch die motivationale Entfaltung von Menschen gefördert werden kann.

Das Hauptziel dieser Arbeit lag in der Untersuchung der psychometrischen Güte des MVSQ. Dazu kann gesagt werden, dass alle Untersuchungen dafür sprechen, dass es sich beim MVSQ um ein valides Instrument handelt, wenngleich vor allem die Messgenauigkeit betreffend noch

Verbesserungsbedarf besteht. Ein weiterer praktischer Nutzen dieser Arbeit liegt deshalb darin, dass einige der Inhalte als Basis für die Entwicklung eines Testmanuals verwendet werden können.

11.6 Ausblick

Diese Arbeit kann als Ausgangspunkt für umfangreiche psychologische Wertesystem-Forschung mit Hilfe des MVSQ gesehen werden, denn es wurden darin Objektivität und Reliabilität als Grundvoraussetzung untersucht und bestätigt. Die Untersuchungen zu den unterschiedlichen Facetten der Validität haben ferner gezeigt, dass es sich bei dem Wertemodell um einen vielversprechenden Ansatz handelt, die Richtung von Motivation zu operationalisieren.

Insbesondere das Design der Studie zur prädiktiven Validität kann als integratives Studien-Design verwendetet werden, in dem die drei Dimensionen der Motivation (Richtung, Intensität und Persistenz) in Verbindung miteinander untersucht werden können. Davon ausgehend könnte ein ausgedehntes Forschungsprogramm entwickelt werden, indem erstens die Kongruenz von unterschiedlichen Aufgaben zu Wertesystemen erforscht werden kann, zweitens Moderator- und Mediatorvariablen eingebunden werden können und drittens das Zusammenwirken von Wertesystem-Aufgaben-Kongruenz unter Berücksichtigung von intervenierenden Variablen (wie monetärem Anreiz) auf die Intensität und Persistenz der Motivation untersucht werden kann. Ergebnisse daraus könnten Anwendung in allen Bereichen finden, in denen das Thema Arbeitsmotivation relevant ist. Beginnend bei Fragen der Studien- und Berufswahl, über Recruiting, Teambesetzung, Mitarbeiterführung bis hin zur Organisationsentwicklung sind Verknüpfungen mit dem Studiendesign denkbar. Auch auf Untersuchungen zu Gruppenprozessen kann es angewendet werden, da es sich bei den untersuchten Aufgaben auch um Gruppenaufgaben handeln kann.

Bevor ein solches Forschungsprogramm realisiert werden kann, sollten jedoch folgende Schritte erfolgen. Erstens sollte das Instrument mit dem Ziel der Verbesserung der empirischen Reliabilitäten überarbeitet werden. Zweitens sollte ein Testmanual erstellt werden, das neben den in dieser Arbeit behandelten Elementen der Hauptgütekriterien auch Untersuchungen der Nebengütekriterien enthalten sollte. Allen voran ist hier die Frage der Skalierung (d.h. Entwicklung einer Verrechnungsregel) zu nennen (Moosbrugger & Kelava, 2012), da es unvorteilhaft ist, bei jeder neuen Testperson ein TIRT-Modell schätzen zu müssen. Des Weiteren ist es empfehlenswert, den MVSQ zu eichen. Dazu sollten für unterschiedliche Vergleichsstichproben wie Studierende und Berufstätige, möglicherweise nach Alter gestuft, Normtabellen erstellt werden. Darüber hinaus können aus den in dieser Arbeit erforschten Zusammenhängen Rückschlüsse auf die Nützlichkeit des Instruments gezogen werden, die ebenfalls Eingang ins Testmanual finden sollten. Auch die Aspekte der Zumutbarkeit, Unverfälschbarkeit und Fairness sollten dabei

bedacht werden. Drittens sollten weitere Studien zur Test-Retest-Reliabilität der Wertesysteme durchgeführt werden, um die Stabilität in unterschiedlichen Zeitintervallen zu untersuchen.

Zu guter Letzt sei gesagt, dass mit dieser Arbeit das formulierte Ziel erreicht wurde, die psychometrische Güte des MVSQ gemessen an den Hauptgütekriterien der Objektivität, Reliabilität und Validität zu untersuchen. Die Untersuchungen wurden dabei in einen inhalts- und testtheoretischen Rahmen eingebettet und die Hauptgütekriterien anhand aller wesentlichen Facetten untersucht. Ferner wurde die praktische Relevanz des Konstrukts *Wertesystem* aufgezeigt, sowie ein flexibel anpassbares Experimentaldesign als Vorlage für weitere experimentelle Untersuchungen entworfen und erprobt. Diese Arbeit leistet somit einen Beitrag zur Erforschung des Einflusses von Wertesystemen in motivationspsychologischen Fragestellungen und kann gleichzeitig als Anreiz und Grundlage für weitere Wertesystem-Forschung gesehen werden.

Literatur

Abbott, G. N., White, F. A. & Charles, M. A. (2005). „Linking values and organizational commitment: A correlational and experimental investigation in two organizations“. In: *Journal of Occupational and Organizational Psychology* 78 (4), S. 531–551.

Achtziger, A. & Gollwitzer, P. M. (2010). „Motivation und Volition im Handlungsverlauf“. In: *Motivation und Handeln*. Hrsg. von Heckhausen, J. & Heckhausen, H. 4. Aufl. Berlin: Springer, S. 309–336.

Adams, J. S. (1963). „Towards an understanding of inequity“. In: *The Journal of Abnormal and Social Psychology* 67 (5), S. 422–436.

Adkins, C. L., Russell, C. J. & Werbel, J. D. (1994). „Judgement of Fit in the Selection Process: The Role of Work Value Congruence“. In: *Personnel Psychology* 47, S. 605–623.

Ajzen, I. (1991). „The Theory of Planned Behavior“. In: *Organizational Behavior and Human Decision Processes* 50, S. 179–211.

Ajzen, I. (2005). *Attitudes, Personality, and Behavior*. 2. Aufl. Maidenhead: Open University Press.

Allport, G. W. (1961). *Pattern and growth in personality*. New York: Holt, Rinehart & Winston.

Allport, G., Vernon, P. & Lindzey, G. (1960). *Study of Values: A Scale for Measuring the Dominant Interests in Personality ; Manual. Supplement*. Houghton Mifflin.

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

Andrich, D. (1988). *Rasch models for measurement*. Newbury Park: Sage Publications.

Asendorpf, J. B. (2015). *Persönlichkeitspsychologie für Bachelor*. 3. Aufl. Berlin: Springer.

Atkinson, J. W. (1957). „Motivational determinants of risk-taking behavior“. In: *Psychological Review* 64 (6), S. 359–372.

Ayala, R. J. d. (2013). *The Theory and Practice of Item Response Theory*. New York: Guilford Press.

Bagozzi, R. P., Bergami, M. & Leone, L. (2003). „Hierarchical representation of motives in goal setting". In: *Journal of Applied Psychology* 88 (5), S. 915–943.

Ball-Rokeach, S. J. & Loges, W. E. (1994). „Choosing Equality: The Correspondence Between Attitudes About Race and the Value of Equality". In: *Journal of Social Issues* 50 (4), S. 9–18.

Bandura, A. (1977). „Self-efficacy: Toward a Unifying Theory of Behavioral Change". In: *Psychological Review* 84 (2), S. 191–215.

Bandura, A. (1982). „Self-efficacy mechanism in human agency". In: *American Psychologist* 37 (2), S. 122–147.

Bandura, A. (1991). „Social cognitive theory of self-regulation". In: *Organizational Behavior and Human Decision Processes* 50 (2), S. 248–287.

Bandura, A. & Cervone, D. (1983). „Self-evaluative and self-efficacy mechanisms governing the motivational effects of goal systems". In: *Journal of Personality and Social Psychology* 45 (5), S. 1017–1028.

Bär-Sieber, M., Krumm, R. & Wiehle, H. (2014). *Unternehmen verstehen, gestalten, verändern. Das Graves-Value-System in der Praxis*. Wiesbaden: Springer Gabler.

Bardi, A. & Schwartz, S. H. (2003). „Values and behavior: strength and structure of relations". In: *Personality and Social Psychology Bulletin* 29 (10), S. 1207–1220.

Bargh, J. A. (1990). „Auto-Motives: Preconscious Determinants of Social Interaction". In: *Handbook of motivation and cognition. Foundations of social behavior*. Hrsg. von Higgins, E. T. & Sorrentino, R. M. New York: Guilford Press.

Bargh, J. A. & Ferguson, M. J. (2000). „Beyond behaviorism: On the automaticity of higher mental processes". In: *Psychological Bulletin* 126 (6), S. 925–945.

Baron, H. (1996). „Strengths and limitations of ipsative measurement". In: *Journal of Occupational and Organizational Psychology* (69), S. 49–56.

Baron, R. M. & Kenny, D. A. (1986). „The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations". In: *Journal of Personality and Social Psychology* 51 (6), S. 1173–1182.

Barreto, H. & Howland, F. M. (2006). *Introductory econometrics. Using Monte Carlo Simulation with Microsoft Excel*. Cambridge und New York: Cambridge University Press.

Barrick, M. R. & Mount, M. K. (1991). „The Big Five Personality Dimensions and Job Performance: A Meta-Analysis". In: *Personnel Psychology* 44 (1), S. 1–26.

Bartram, D. (1996). „The relationship between ipsatized and normative measures of personality". In: *Journal of Occupational and Organizational Psychology* 69, S. 25–39.

Bartram, D. (2005). „The Great Eight Competencies: A Criterion-Centric Approach to Validation". In: *Journal of Applied Psychology* 90 (6), S. 1185–1203.

Bartram, D. (2007). „Increasing Validity with Forced-Choice Criterion Measurement Formats". In: *International Journal of Selection and Assessment* 15 (3), S. 263–272.

Beck, D. E. & Cowan, C. C. (2007). *Spiral Dynamics. Leadership, Werte und Wandel*. Bielefeld: Kamphausen.

Beckmann, J. & Heckhausen, H. (2010). „Motivation durch Erwartung und Anreiz". In: *Motivation und Handeln*. Hrsg. von Heckhausen, J. & Heckhausen, H. 4. Aufl. Berlin: Springer, S. 105–144.

Bonett, D. G. & Wright, T. A. (2000). „Sample size requirements for estimating pearson, kendall and spearman correlations". In: *Psychometrika* 65 (1), S. 23–28.

Borg, I., Groenen, P. J. F., Jehn, K. A., Bilsky, W. & Schwartz, S. H. (2011). „Embedding the Organizational Culture Profile Into Schwartz's Theory of Universals in Values". In: *Journal of Personnel Psychology* 10 (1), S. 1–12.

Borsboom, D., Mellenbergh, G. J. & Heerden, J. van (2004). „The Concept of Validity". In: *Psychological Review* 111 (4), S. 1061–1071.

Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler*. 4. Aufl. Heidelberg: Springer.

Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler*. 7. Aufl. Springer-Lehrbuch. Berlin: Springer.

Box, G. E. P. (1953). „Non-Normality and Tests on Variances". In: *Biometrika* 40 (3/4), S. 318.

Boxx, W. R., Odom, R. Y. & Dunn, M. G. (1991). „Organizational Values and Value Congruency and Their Impact on Satisfaction, Commitment, and Cohesion: An Empirical Examination within the Public Sector“. In: *Public Personnel Management* 20 (2), S. 195–205.

Braithwaite, V. A. & Law, H. G. (1985). „Structure of human values: Testing the adequacy of the Rokeach Value Survey“. In: *Journal of Personality and Social Psychology* 49 (1), S. 250–263.

Brandstätter, V., Schüler, J., Puca, R. M. & Lozo, L. (2013). *Motivation und Emotion*. Berlin und Heidelberg: Springer.

Brosch, T. (2013). „Comment: On the Role of Appraisal Processes in the Construction of Emotion“. In: *Emotion Review* 5 (4), S. 369–373.

Brown, A. (2010). „How Item Response Theory can solve problems of ipsative data“. Facultad de Psicología - Departamento de Personalidad, Evaluación y Tratamientos Psicológicos. Dissertation. Barcelona: Universidad de Barcelona.

Brown, A. (2014). „Item Response Models for Forced-Choice Questionnaires: A Common Framework“. In: *Psychometrika*, S. 1–26.

Brown, A. & Croudace, T. J. (2015). „Scoring and Estimating Score Precision Using Multidimensional IRT Models“. In: *Handbook of Item Response Theory Modeling. Applications to Typical Performance Assessment*. Hrsg. von Reise, S. P. & Revicki, D. A. New York: Routledge, S. 307–333.

Brown, A. & Maydeu-Olivares, A. (2011). „Item Response Modeling of Forced-Choice Questionnaires“. In: *Educational and Psychological Measurement* 71 (3), S. 460–502.

Brown, A. & Maydeu-Olivares, A. (2012). „Fitting a Thurstonian IRT model to forced-choice data using Mplus“. In: *Behavior Research Methods* 44 (4), S. 1135–1147.

Brown, A. & Maydeu-Olivares, A. (2013). „How IRT can solve problems of ipsative data in forced-choice questionnaires“. In: *Psychological Methods* 18 (1), S. 36–52.

Brunstein, J. C. (2010a). „Implizite und explizite Motive“. In: *Motivation und Handeln*. Hrsg. von Heckhausen, J. & Heckhausen, H. 4. Aufl. Berlin: Springer, S. 237–256.

Brunstein, J. C. (2010b). „Leistungsmotivation“. In: *Motivation und Handeln*. Hrsg. von Heckhausen, J. & Heckhausen, H. 4. Aufl. Berlin: Springer, S. 145–192.

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion.* 3. Aufl. München: Pearson Studium.

Cable, D. M. & Judge, T. A. (1997). „Interviewers' perceptions of person–organization fit and organizational selection decisions". In: *Journal of Applied Psychology* 82 (4), S. 546–561.

Calogero, R. M., Bardi, A. & Sutton, R. M. (2009). „A need basis for values: Associations between the need for cognitive closure and value priorities". In: *Personality and Individual Differences* 46 (2), S. 154–159.

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

Carter, S. M. & Greer, C. R. (2013). „Strategic Leadership: Values, Styles, and Organizational Performance". In: *Journal of Leadership & Organizational Studies* 20 (4), S. 375–393.

Carver, C. S. (2006). „Approach, Avoidance, and the Self-Regulation of Affect and Action". In: *Motivation and Emotion* 30 (2), S. 105–110.

Cattell, R. B. (1944). „Psychological Measurement: Normative, Ipsative, Interactive". In: *Psychological Review* 51 (5), S. 292–303.

Cattell, R. B. (1946). „Personality structure and measurement II: The determination and utility of trait modality". In: *The British Journal of Psychology* (36), S. 159–174.

Cattell, R. B. (1973). *Die empirische Erforschung der Persönlichkeit.* Weinheim: Beltz.

Cattell, R. B. (1987). *Intelligence. Its structure, growth, and action.* Amsterdam und New York: North-Holland.

Cattell, R. B. & Brennan, J. (1994). „Finding Personality Structure When Ipsative Measurements Are the Unavoidable Basis of the Variables". In: *The American Journal of Psychology* 107 (2), S. 261.

Chan, W. (2003). „Analyzing Ipsative Data in Psychological Research". In: *Behaviormetrika* 30 (1), S. 99–121.

Chatman, J. (1989). „Improving Interactional Organizational Research: A Model of Person-Organization Fit". In: *Academy of Management Review* 14 (3), S. 333–349.

Cherrington, D. J., Conde, S. J. & England, J. L. (1979). „Age Work Values“. In: *Academy of Management Journal* 22 (3), S. 617–623.

Cheung, M. W.-L. & Chan, W. (2002). „Reducing Uniform Response Bias With Ipsative Measurement in Multiple-Group Confirmatory Factor Analysis“. In: *Structural Equation Modeling: A Multidisciplinary Journal* 9 (1), S. 55–77.

Christiansen, N. D., Burns, G. N. & Montgomery, G. E. (2005). „Reconsidering Forced-Choice Item Formats for Applicant Personality Assessment“. In: *Human Performance* 18 (3), S. 267–307.

Clemans, W. V. (1966). „An analytical and empirical examination of some properties of ipsative measures.“ In: *Psychometric Monographs* 14.

Closs, S. J. (1996). „On the factoring and interpretation of ipsative data“. In: *Journal of Occupational and Organizational Psychology* 69 (1), S. 41–47.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York: Psychology Press.

Cornwell, J. M., Manfredo, P. A. & Dunlap, W. P. (1991). „Factor Analysis of the 1985 Revision of Kolb's Learning Style Inventory“. In: *Educational and Psychological Measurement* 51 (2), S. 455–462.

Costello, A. B. & Osborne, J. (2005). „Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis“. In: *Practical Assessment Research & Evaluation* 10 (7).

Crawford, J. T., Jussim, L. & Pilanski, J. M. (2014). „How (Not) To Interpret and Report Main Effects and Interactions in Multiple Regression: Why Crawford and Pilanski Did Not Actually Replicate Lindner and Nosek (2009)“. In: *Political Psychology* 35 (6), S. 857–862.

Cronbach, L. J. & Meehl, P. E. (1955). „Construct validity in psychological tests“. In: *Psychological Bulletin* 52 (4), S. 281–302.

Csikszentmihalyi, M. (1975). *Beyond boredom and anxiety*. San Francisco: Jossey-Bass.

Csikszentmihalyi, M. (1990). *Flow. The psychology of optimal experience*. 1. ed. New York: Harper & Row. XII, 303 S.

Csikszentmihalyi, M. (2002). *Flow. The Classic Work on How to Achieve Happiness*. London: Rider.

Day, D. V. & Silverman, S. B. (1989). „Personality and Job Performance: Evidence of Incremental Validity". In: *Personnel Psychology* 42 (1), S. 25–36.

DeCarlo, L. T. (1997). „On the meaning and use of kurtosis". In: *Psychological Methods* 2 (3), S. 292–307.

Deci, E. L. & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York: Plenum Press.

Dekking, F. M., Kraaikamp, C., Lopuhaä, H. P. & Meester, L. E. (2005). *A Modern Introduction to Probability and Statistics. Understanding Why and How*. London: Springer.

Diedenhofen, B., Musch, J. & (Keine Angabe) (2015). „cocor: A Comprehensive Solution for the Statistical Comparison of Correlations". In: *PLOS ONE* 10 (4), e0121945.

Diekmann, A. (2010). *Empirische Sozialforschung. Grundlagen, Methoden, Anwendungen*. 21. Aufl. Reinbek bei Hamburg: Rowohlt-Taschenbuch-Verl.

Diener, E. (1984). „Subjective well-being". In: *Psychological Bulletin* 95 (3), S. 542–575.

Digman, J. M. (1990). „Personality Structure: Emergence of the Five-Factor Model". In: *Annual Review of Psychology* 41 (1), S. 417–440.

Dose, J. J. (1997). „Work Values: An integrative framework and illustrative application to organizational socialization". In: *Journal of Occupational and Organizational Psychology* 70 (3), S. 219–240.

Dubinsky, A. J., Kotabe, M., Lim, C. U. & Wagner, W. (1997). „The impact of values on salespeople's job responses: A cross-national investigation". In: *Journal of Business Research* 39 (3), S. 195–208.

Dunlap, W. P. & Cornwell, J. M. (1994). „Factor Analysis of Ipsative Measures". In: *Multivariate Behavioral Research* 29 (1), S. 115–126.

Eagly, A. H. & Chaiken, S. (1993). *The Psychology of Attitudes*. Fort Worth: Harcourt Brace Jovanovich.

Eccles, J. S. & Wigfield, A. (2002). „Motivational Beliefs, Values, and Goals“. In: *Annual Review of Psychology* 53 (1), S. 109–132.

Eid, M., Gollwitzer, M. & Schmitt, M. (2015). *Statistik und Forschungsmethoden. Lehrbuch. Mit Online-Material*. Weinheim: Beltz.

Elhai, J. D., Schweinle, W. & Anderson, S. M. (2008). „Reliability and validity of the Attitudes Toward Seeking Professional Psychological Help Scale-Short Form“. In: *Psychiatry Research* 159 (3), S. 320–329.

Elkins, T. & Keller, R. T. (2003). „Leadership in research and development organizations: A literature review and conceptual framework“. In: *The Leadership Quarterly* 14 (4-5), S. 587–606.

Elliot, A. J., Hrsg. (2008). *Handbook of approach and avoidance motivation*. New York: Psychology Press.

Elliot, A. J. & Devine, P. G. (1994). „On the motivational nature of cognitive dissonance: Dissonance as psychological discomfort“. In: *Journal of Personality and Social Psychology* 67 (3), S. 382–394.

Embretson, S. E. & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Multivariate applications books series. Mahwah und NJ: Erlbaum.

England, G. W. & Lee, R. (1974). „The relationship between managerial values and managerial success in the United States, Japan, India, and Australia“. In: *Journal of Applied Psychology* 59 (4), S. 411–419.

Eysenck, H. (1991). „Dimensions of personality: 16, 5 or 3?—Criteria for a taxonomic paradigm“. In: *Personality and Individual Differences* 12 (8), S. 773–790.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C. & Strahan, E. J. (1999). „Evaluating the use of exploratory factor analysis in psychological research“. In: *Psychological Methods* 4 (3), S. 272–299.

Fahrmeir, L., Kneib, T. & Lang, S. (2009). *Regression*. Berlin und Heidelberg: Springer.

Feather, N. T. (1977). „Value importance, conservatism, and age“. In: *European Journal of Social Psychology* 7 (2), S. 241–245.

Feather, N. T. (2002). „Values and Value Dilemmas in Relation to Judgments Concerning Outcomes of an Industrial Conflict“. In: *Personality and Social Psychology Bulletin* 28 (4), S. 446–459.

Feather, N. T. (1975). *Values in education and society*. New York: Free Press.

Feather, N. T. (1995). „Values, Valences, and Choice: The Influence of Values on the Perceived Attractiveness and Choice of Alternatives“. In: *Journal of Personality and Social Psychology* 68 (6), S. 1135–1151.

Feather, N. T. & Peay, E. R. (1975). „The Structure of Terminal and Instrumental Values: Dimensions and Clusters“. In: *Australian Journal of Psychology* 27 (2), S. 151–164.

Festinger, L. (1962). *A theory of cognitive dissonance*. Stanford: Stanford University Press.

Fischer, P., Frey, D. & Niedernhuber, J. (2013a). „Führung und Werte: humanistische Führung in Theorie und Praxis“. In: *Führungskompetenzen lernen. Eignung, Entwicklung, Aufstieg*. Hrsg. von Häring, K. & Litzcke, S. Stuttgart: Schäffer-Poeschel, S. 161–180.

Fischer, P., Asal, K. & Krueger, J. I. (2013b). *Sozialpsychologie für Bachelor: Lesen, Hören, Lernen im Web*. Berlin und Heidelberg: Springer.

Fischer, R. & Schwartz, S. (2011). „Whence Differences in Value Priorities?: Individual, Cultural, or Artifactual Sources“. In: *Journal of Cross-Cultural Psychology* 42 (7), S. 1127–1144.

Fishbein, M. & Ajzen, I. (2010). *Predicting and changing behavior. The reasoned action approach*. New York: Psychology Press.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Genesis Publishing Pvt Ltd.

Fletcher, T. D. (2010). *psychometric: Applied Psychometric Theory*. Version 2.2.

Frey, B. S. & Osterloh, M. (2005). „Yes, Managers Should Be Paid Like Bureaucrats“. In: *Journal of Management Inquiry* 14 (1), S. 96–111.

Furnham, A. (2014). *The New Psychology of Money*. Taylor & Francis.

Gardner, W., Mulvey, E. P. & Shaw, E. C. (1995). „Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models“. In: *Psychological Bulletin* 118 (3), S. 392–404.

Gelman, A. & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Leiden: Cambridge University Press.

Glöckner-Rist, A. (2012). *Der Schwartz Value Survey (SVS)*. Hrsg. von Glöckner-Rist, A. Bonn.

Goldberg, L. R. (1990). „An alternative "description of personality": The Big-Five factor structure". In: *Journal of Personality and Social Psychology* 59 (6), S. 1216–1229.

Goldberg, L. R. (1992). „The development of markers for the Big-Five factor structure". In: *Psychological Assessment* 4 (1), S. 26–42.

Goldberg, L. R. (1993). „The Structure of Phenotypic Personality Traits". In: *American Psychologist* 48 (1), S. 26–34.

Gollwitzer, P. M. (1996). „The Volitional Benefits of Planning". In: *The Psychology of Action. Linking Cognition and Motivation to Behavior*. Hrsg. von Gollwitzer, P. M. & Bargh, J. A. New York: Guilford Press, S. 287–312.

Gollwitzer, P. M. (1999). „Implementation intentions: Strong effects of simple plans". In: *American Psychologist* 54 (7), S. 493–503.

Gollwitzer, P. M. & Sheeran, P. (2006). „Implementation Intentions and Goal Achievement: A Meta-Analysis of Effects and Processes". In: *Advances in Experimental Social Psychology* 38, S. 69–119.

Gollwitzer, P. M., Barry, H. & Oettingen, G. (2011). „Needs and incentives as sources of goals". In: *Goal-Directed Behavior*. Hrsg. von Aarts, H. & Elliot, A. New York: Psychology Press, S. 115–149.

Graf, M. M., Quaquebeke, N. van & Dick, R. van (2011). „Two Independent Value Orientations: Ideal and Counter-Ideal Leader Values and Their Impact on Followers' Respect for and Identification with Their Leaders". In: *Journal of Business Ethics* 104 (2), S. 185–195.

Graumann, C. F. & Willig, R. (1983). „Wert, Wertung, Werthaltung". In: *Theorien und Formen der Motivation*. Hrsg. von Thomae, H. Göttingen: Hogrefe.

Graves, C. W. (1966). „Deterioration of Work Standards". In: *Harvard Business Review* 44 (5), S. 117–128.

Graves, C. W. (1969). *A Systems View of Values Problems*. Unter Mitarb. von Cybernetic Corporation. Philadelphia und PA.

Graves, C. W. (1970). „Levels of Existence: an Open System Theory of Values“. In: *Journal of Humanistic Psychology*, S. 131–155.

Graves, C. W. (1971a). *How Should Who Lead Whom To Do What? Paper delivered at the YMCA Management Forum of 1971-1972.*

Graves, C. W. (1971b). „Levels of Existence Related to Learning Systems. Paper read at the Ninth Annual Conference of the National Society for Programmed Instruction, Rochester, New York, March 31, 1971“. Rochester und New York.

Graves, C. W. (1971c). *Levels of human existence : transcription of a seminar at the Washington School of Psychiatry, October 16, 1971.* Hrsg. von Lee, W. R. Version 3. Santa Barbara und CA.

Graves, C. W. (1974). „Human Nature Prepares for a Momentous Leap“. In: *The Futurist*, S. 72–87.

Graves, C. W. (2005). *The Never Ending Quest: Dr. Clare W. Graves Explores Human Nature. A Treatise on an Emergent Cyclical Conception of Adult Behavioral Systems and Their Development.* Unter Mitarb. von Cowan, C. C. & Todorovic, N. Santa Barbara und CA: ECLECT Publishing.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L. & Reckase, M. D. (1984). „Technical Guidelines for Assessing Computerized Adaptive Tests“. In: *Journal of Educational Measurement)* 21 (4), S. 347–360.

Groth-Marnat, G. (2003). *Handbook of Psychological Assessment.* 4. Aufl. Hoboken und N.J: John Wiley & Sons.

Gruber, T. (2011). *Gedächtnis.* Wiesbaden: VS, Verl. für Sozialwiss.

Hagemeister, C., Lang, F. & Kersting, M. (2010). „Einstellung von Psychologinnen und Psychologen in Deutschland zu Tests“. In: *Report Psychologie* 35, S. 428–439.

Hartig, J., Frey, A. & Jude, N. (2012). „Validität“. In: *Testtheorie und Fragebogenkonstruktion.* Hrsg. von Moosbrugger, H. & Kelava, A. Berlin: Springer, S. 143–172.

Harwell, M., Stone, C. A., Hsu, T.-C & Kirisci, L. (1996). „Monte Carlo Studies in Item Response Theory“. In: *Applied Psychological Measurement* 20 (2), S. 101–125.

Harzing, A.-W., Baldueza, J., Barner-Rasmussen, W., Barzantny, C., Canabal, A., Davila, A., Espejo, A., Ferreira, R., Giroud, A., Koester, K., Liang, Y.-K., Mockaitis, A., Morley, M. J., Myloni, B., Odusanya, J. O., O'Sullivan, S. L., Palaniappan, A. K., Prochno, P., Choudhury,

S. R., Saka-Helmhout, A., Siengthai, S., Viswat, L., Soydas, A. U. & Zander, L. (2009). „Rating versus ranking: What is the best way to reduce response and language bias in cross-national research?“ In: *International Business Review* 18 (4), S. 417–432.

Haslam, S. A. (2009). *Psychology in organizations. The social identity approach*. London: Sage. 306 S.

Hastie, T., Tibshirani, R. & Friedman, J. H. (2013). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. New York: Springer.

Haynes, S. N. & Lench, H. C. (2003). „Incremental Validity of New Clinical Assessment Measures“. In: *Psychological Assessment* 15 (4), S. 456–466.

Haynes, S. N., Richard, D. C. S. & Kubany, E. S. (1995). „Content validity in psychological assessment: A functional approach to concepts and methods“. In: *Psychological Assessment* 7 (3), S. 238–247.

Heckhausen, J. & Heckhausen, H. (2010). „Motivation und Handeln: Einführung und Überblick“. In: *Motivation und Handeln*. Hrsg. von Heckhausen, J. & Heckhausen, H. 4. Aufl. Berlin: Springer, S. 1–10.

Heider, F. (1958). *The psychology of interpersonal relations*. Hoboken und N.J: John Wiley & Sons.

Henrich, J., Heine, S. J. & Norenzayan, A. (2010). „The weirdest people in the world?“ In: *Behavioral and Brain Sciences* 33 (2-3), S. 61–83.

Herzberg, P. Y. & Roth, M. (2014). *Persönlichkeitspsychologie*. Wiesbaden: Springer VS.

Heywood, H. B. (1931). „On Finite Sequences of Real Numbers“. In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 134 (824), S. 486–501.

Hicks, L. E. (1970). „Some Properties of Ipsative, Normative, and Forced-Choice Normative Measures“. In: *Psychological Bulletin* 74 (3), S. 167–184.

Higgins, E. T. (1997). „Beyond pleasure and pain“. In: *American Psychologist* 52 (12), S. 1280–1300.

Hitlin, S. (2003). „Values As the Core of Personal Identity: Drawing Links Between Two Theories of Self“. In: *Social Psychology Quarterly* 66 (2), S. 118–137.

Hitlin, S. & Piliavin, J. A. (2004). „Values: Reviving a Dormant Concept“. In: *Annual Review of Sociology* 30, S. 359–393.

Hofstede, G. (1993). „Cultural constraints in management theories“. In: *Academy of Management Perspectives* 7 (1), S. 81–94.

Hofstede, G. (1980). „Motivation, Leadership, and Organization: Do American Theories Apply Abroad?“ In: *Organizational Dynamics* 9 (1), S. 42–63.

Hofstede, G. (1984). „The Cultural Relativity of the Quality of Life Concept“. In: *The Academy of Management Review* 9 (3), S. 389–398.

Holland, J. L. (1997). *Making vocational choices: A theory of vocational personalities and work environments*. Psychology Assessment Resources.

Holmbeck, G. N. (1997). „Toward terminological, conceptual, and statistical clarity in the study of mediators and moderators: Examples from the child-clinical and pediatric psychology literatures“. In: *Journal of Consulting and Clinical Psychology* 65 (4), S. 599–610.

Horn, J. L. (1965). „A rationale and test for the number of factors in factor analysis“. In: *Psychometrika* 30 (2), S. 179–185.

Hu, L.-t. & Bentler, P. M. (1999). „Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives“. In: *Structural Equation Modeling: A Multidisciplinary Journal* 6 (1), S. 1–55.

Hunsley, J. & Meyer, G. J. (2003). „The Incremental Validity of Psychological Testing and Assessment: Conceptual, Methodological, and Statistical Issues“. In: *Psychological Assessment* 15 (4), S. 446–455.

Ihme, J. M., Lemke, F., Lieder, K., Martin, F., Müller, J. C. & Schmidt, S. (2009). „Comparison of ability tests administered online and in the laboratory“. In: *Behavior Research Methods* 41 (4), S. 1183–1189.

Irtel, H. (1996). *Entscheidungs- und testtheoretische Grundlagen der Psychologischen Diagnostik*. Frankfurt am Main: P. Lang.

Jaccard, J. & Turrisi, R. (2003). *Interaction effects in multiple regression*. Thousand Oaks: Sage Publications. vii, 92.

Jackson, D. N., Wroblewski, V. R. & Ashton, M. C. (2000). „The Impact of Faking on Employment Tests: Does Forced Choice Offer a Solution?“ In: *Human Performance* 13 (4), S. 371–388.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2014). *An Introduction to Statistical Learning. with Applications in R*. New York: Springer.

Jin, J. & Rounds, J. (2012). „Stability and change in work values: A meta-analysis of longitudinal studies“. In: *Journal of Vocational Behavior* 80 (2), S. 326–339.

Johnson, C. E., Wood, R. & Blinkhorn, S. F. (1988). „Spuriouser and spuriouser: The use of ipsative personality tests“. In: *Journal of Occupational and Organizational Psychology* (61), S. 153–162.

Jolliffe, I. T. (1986). *Principal Component Analysis*. New York und NY: Springer.

Jonkisz, E., Moosbrugger, H. & Brandt, H. (2012). „Planung und Entwicklung von Tests und Fragebogen“. In: *Testtheorie und Fragebogenkonstruktion*. Hrsg. von Moosbrugger, H. & Kelava, A. Berlin: Springer, S. 27–74.

Joreskog, K. G. & Goldberger, A. S. (1972). „Factor analysis by generalized least squares“. In: *Psychometrika* 37 (3), S. 243–260.

Judge, T. A. & Bretz, R. D. (1992). „Effects of work values on job choice decisions“. In: *Journal of Applied Psychology* 77 (3), S. 261–271.

Judge, W. Q., Fryxell, G. E. & Dooley, R. S. (1997). „The New Task of R&D Management: Creating Goal-Directed Communities for Innovation“. In: *California Management Review* 39 (3), S. 72–85.

Kahneman, D. & Tversky, A. (1979). „Prospect Theory: An Analysis of Decision under Risk“. In: *Econometrica* 47 (2), S. 263–292.

Kaiser, H. F. (1960). „The Application of Electronic Computers to Factor Analysis“. In: *Educational and Psychological Measurement* 20 (1), S. 141–151.

Kamakura, W. A. & Mazzon, J. A. (1991). „Value Segmentation: A Model for the Measurement of Values and Value Systems“. In: *Journal of Consumer Research* 18 (2), S. 208–218.

Kehr, H. M. (2004a). „Implicit/Explicit Motive Discrepancies and Volitional Depletion among Managers“. In: *Personality and Social Psychology Bulletin* 30 (3), S. 315–327.

Kehr, H. M. (2004b). „Integrating implicit motives. explicit motives, and perceived abilities - The compensatory model of work motivation and volition“. In: *Academy of Management Review* 29 (3), S. 479–499.

Keijser, C. & Vat, S. van der (2009). *Management drives field book. For managers, teams and their coaches.* Amsterdam: FT Prentice Hall.

Kelava, A. & Moosbrugger, H. (2012). „Deskriptivstatistische Evaluation von Items (Itemanalyse) und Testwertverteilungen“. In: *Testtheorie und Fragebogenkonstruktion.* Hrsg. von Moosbrugger, H. & Kelava, A. Berlin: Springer, S. 75–102.

Kersting, M. (2001). „Zur Konstrukt- und Kriteriumsvalidität von Problemlöseszenarien anhand der Vorhersage von Vorgesetztenurteilen über die berufliche Bewährung“. In: *Diagnostica* 47 (2), S. 67–76.

Kim, S. (2012a). *ppcor: Partial and Semi-partial (Part) correlation.* Version 1.0.

Kim, S. (2015). „ppcor: An R Package for a Fast Calculation to Semi-partial Correlation Coefficients“. In: *Communications for Statistical Applications and Methods* 22 (6), S. 665–674.

Kim, S. (2012b). „A Note on the Reliability Coefficients for Item Response Model-Based Ability Estimates“. In: *Psychometrika* 77 (1), S. 153–162.

Kleinbeck, U. (2010). „Handlungsziele“. In: *Motivation und Handeln.* Hrsg. von Heckhausen, J. & Heckhausen, H. 4. Aufl. Berlin: Springer, S. 285–307.

Kluckhohn, C. (1951). „Values and value-orientations in the theory of action: An exploration in definition and classification.“ In: *Toward a general theory of action.* Hrsg. von Parsons, T. & Shils, E. Cambridge und MA: Harvard University Press.

Köbler, R. (2009). *Neue Wege im Recruiting. Mehr Effektivität mit Gravesmodell und Metaprogrammen ; ein praxisorientiertes Handbuch.* Paderborn: Junfermann.

Korkmaz, S., Goksuluk, D. & Zararsiz, G. (2014). „MVN: An R Package for Assessing Multivariate Normality“. In: *The R Journal* 6 (2), S. 151–162.

Krapp, A. (1999). „Interest, motivation and learning: An educational-psychological perspective“. In: *European Journal of Psychology of Education* 14 (1), S. 23–40.

Kristof-Brown, A. L., Zimmerman, R. D. & Johnson, E. C. (2005). „Consequences of Individuals' Fit at Work: A Meta-Analysis of Person-Job, Person-Organization, Person-Group, and Person-Supervisor Fit". In: *Personnel Psychology* 58, S. 281–342.

Krohne, H. W., Egloff, B., Kohlmann, C.-W. & Tausch, A. (1996). „Untersuchungen mit einer deutschen Version der Positive and Negative Affect Schedule (PANAS)". In: *Diagnostica* 42 (2), S. 139–156.

Krumm, R. (2012). *9 levels of value systems*. Haiger: Werdewelt Verl.- und Medienhaus-GmbH.

Kuhl, J. (2010). „Individuelle Unterschiede in der Selbststeuerung". In: *Motivation und Handeln*. Hrsg. von Heckhausen, J. & Heckhausen, H. 4. Aufl. Berlin: Springer, S. 337–364.

Kurz, R. & Bartram, D. (2002). „Competency and Individual Performance: Modelling the World of Work". In: *Organizational effectiveness. The Role of Psychology*. Hrsg. von Robertson, I., Callinan, M. & Bartram, D. Chichester und New York: Wiley, S. 227–255.

Lang, F. R., Lüdtke, O. & Asendorpf, J. B. (2001). „Testgüte und psychometrische Äquivalenz der deutschen Version des Big Five Inventory (BFI) bei jungen, mittelalten und alten Erwachsenen". In: *Diagnostica* 47 (3), S. 111–121.

Larson, L. M., Rottinghaus, P. J. & Borgen, F. H. (2002). „Meta-analyses of Big Six Interests and Big Five Personality Factors". In: *Journal of Vocational Behavior* 61 (2), S. 217–239.

Latham, G. P. (2012). *Work motivation. History, theory, research, and practice*. Thousand Oaks: Sage.

Latham, G. P. & Locke, E. A. (1991). „Self-Regulation Through Goal Setting". In: *Organizational Behavior and Human Decision Processes* 50, S. 212–247.

Latham, G. P. & Pinder, C. C. (2005). „Work Motivation Theory and Research at the Dawn of the Twenty-First Century". In: *Annual Review of Psychology* 56, S. 485–516.

Lawler, E. E. (1968). „Equity theory as a predictor of productivity and work quality". In: *Psychological Bulletin* 70 (6), S. 596–610.

Lawrence, M. A. (2015). *ez: Easy Analysis and Visualization of Factorial Experiments*. Version 4.3.

Lea, S. E. G. & Webley, P. (2006). „Money as tool, money as drug: The biological psychology of a strong incentive". In: *Behavioral and Brain Sciences* 29 (2), S. 161–209.

LeBreton, J. M. & Senter, J. L. (2007). „Answers to 20 Questions About Interrater Reliability and Interrater Agreement“. In: *Organizational Research Methods* 11 (4), S. 815–852.

Lewin, K. (1939). „Field Theory and Experiment in Social Psychology: Concepts and Methods“. In: *American Journal of Sociology* 44 (6), S. 868–896.

Lewin, K. (1952). „Field theory in social science: Selected theoretical papers by Kurt Lewin“. In: *Field theory in social science: Selected theoreticalpapers*. Hrsg. von Cartwrigth, D. London: Tavistock.

Lienert, G. A. & Raatz, U. (1998). *Testaufbau und Testanalyse*. Weinheim: Beltz Verlagsgruppe.

Liepmann, D., Beauducel, A. & Brocke, B. A. R. (2007). *Intelligenz-Struktur-Test 2000 R (IST 2000 R). Manual*. Göttingen: Hogrefe.

Locke, E. A. (1968). „Toward a theory of task motivation and incentives“. In: *Organizational Behavior and Human Performance* 3 (2), S. 157–189.

Locke, E. A. (1969). „What is Job Satisfaction?“ In: *Organizational Behavior and Human Performance* 4, S. 309–336.

Locke, E. A. (1991). „The Motivation Sequence, the Motivation Hub, and the Motivation Core“. In: *Organizational Behavior and Human Decision Processes* 50, S. 288–299.

Locke, E. A. (1996). „Motivation through conscious goal setting“. In: *Applied and Preventive Psychology* 5 (2), S. 117–124.

Locke, E. A. (1997). „The Motivation to Work: What We Know“. In: *Advances in Motivation and Achievement*. Hrsg. von Maehr, M. L. & Pintrich, P. R. Greenwich und Conn: Jai Press, S. 375–412.

Locke, E. A. & Henne, D. (1986). „Work Motivation Theories“. In: *International Review of Industrial and Organizational Psychology*.

Locke, E. A. & Kristof, A. L. (1996). „Volitional Choices in the Goal Achievement Process“. In: *The Psychology of Action. Linking Cognition and Motivation to Behavior*. Hrsg. von Gollwitzer, P. M. & Bargh, J. A. New York: Guilford Press, S. 365–403.

Locke, E. A. & Latham, G. P. (1990). „Work Motivation: The High Performance Cycle“. In: *Work motivation*. Hrsg. von Kleinbeck, U., Quast, H.-H., Thierry, H. & Häcker, H. Hillsdale und N.J: Lawrence Erlbaum Associates, S. 3–25.

Locke, E. A. & Latham, G. P. (2002). „Building a practically useful theory of goal setting and task motivation: A 35-year odyssey“. In: *American Psychologist* 57 (9), S. 705–717.

Locke, E. A. & Latham, G. P. (2004). „What Should We Do about Motivation Theory? Six Recommendations for the Twenty-First Century“. In: *The Academy of Management Review* 29 (3), S. 388–403.

Locke, E. A., Frederick, E., Lee, C. & Bobko, P. (1984). „Effect of self-efficacy, goals, and task strategies on task performance“. In: *Journal of Applied Psychology* 69 (2), S. 241–251.

Long, J. S. & Freese, J. (2001). *Regression models for categorical dependent variables using Stata*. College Station: Stata Press.

Lord, F. M., Novick, M. R. & Birnbaum, A., Hrsg. (1968). *Statistical theories of mental test scores*. Reading und Mass.: Addison-Wesley.

Lord, R. G. & Brown, D. J. (2001). „Leadership, Values, and Subordinate Self-Concepts“. In: *The Leadership Quarterly* (12), S. 133–152.

Low, D. K. S., Yoon, M., Roberts, B. W. & Rounds, J. (2005). „The Stability of Vocational Interests From Early Adolescence to Middle Adulthood: A Quantitative Review of Longitudinal Studies“. In: *Psychological Bulletin* 131 (5), S. 713–737.

Maio, G. R., Roese, N. J., Seligman, C. & Katz, A. (1996). „Rankings, Ratings, and the Measurement of Values: Evidence for the Superior Validity of Ratings“. In: *Basic and Applied Social Psychology* 18 (2), S. 171–181.

Maltby, J., Day, L. & Macaskill, A. (2011). *Differentielle Psychologie, Persönlichkeit und Intelligenz*. München: Pearson Studium.

Mardia, K. V. (1970). „Measures of Multivariate Skewness and Kurtosis with Applications“. In: *Biometrika* 57 (3), S. 519.

Markus, H. (1977). „Self-schemata and processing information about the self“. In: *Journal of Personality and Social Psychology* 35 (2), S. 63–78.

Martin, B., Bowen, C.-C & Hunt, S. (2002). „How effective are people at faking on personality questionnaires?“ In: *Personality and Individual Differences* 32 (2), S. 247–256.

Maslow, A. H. (1943). „A theory of human motivation“. In: *Psychological Review* 50 (4), S. 370–396.

Mauchly, J. W. (1940). „Significance Test for Sphericity of a Normal n-Variate Distribution“. In: *The Annals of Mathematical Statistics* 11 (2), S. 204–209.

Maydeu-Olivares, A. & Brown, A. (2010). „Item Response Modeling of Paires Comparison and Ranking Data“. In: *Multivariate Behavioral Research* 45, S. 935–974.

McAuley, E., Duncan, T. & Tammen, V. V. (1989). „Psychometric Properties of the Intrinsic Motivation Inventory in a Competitive Sport Setting: A Confirmatory Factor Analysis“. In: *Research Quarterly for Exercise and Sport* 60 (1), S. 48–58.

McClelland, D. C. (1985). „How motives, skills, and values determine what people do“. In: *American Psychologist* 40 (7), S. 812–825.

McClelland, D. C. (1987). *Human motivation*. Cambridge: Cambridge University Press.

McClelland, D. C., Koestner, R. & Weinberger, J. (1989). „How do self-attributed and implicit motives differ?“ In: *Psychological Review* 96 (4), S. 690–702.

McCloy, R. A., Heggestad, E. D. & Reeve, C. L. (2005). „A Silk Purse From the Sow's Ear: Retrieving Normative Information From Multidimensional Forced-Choice Items“. In: *Organizational Research Methods* 8 (2), S. 222–248.

McDonald, R. P. (1985). *Factor analysis and related methods*. Hillsdale und N.J: Erlbaum. 259 S.

McFadden, D. (1974). „Conditional logit analysis of qualitative choice behavior“. In: *Frontiers in econometrics*. Hrsg. von Zarembka, P. Economic theory and mathematical economics. New York: Acad. Press, S. 105–142.

McGrath, R. E., Pogge, D. L. & Stokes, J. M. (2002). „Incremental validity of selected MMPI-A content scales in an inpatient setting“. In: *Psychological Assessment* 14 (4), S. 401–409.

Mead, A. D. & Drasgow, F. (1993). „Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis“. In: *Psychological Bulletin* 114 (3), S. 449–458.

Meade, A. W., Michels, L. C. & Lautenschlager, G. J. (2007). „Are Internet and Paper-and-Pencil Personality Tests Truly Comparable?: An Experimental Design Measurement Invariance Study“. In: *Organizational Research Methods* 10 (2), S. 322–345.

Meade, A. W. (2004). „Psychometric problems and issues involved with creating and using ipsative measures for selection“. In: *Journal of Occupational and Organizational Psychology* 77 (4), S. 531–551.

Meglino, B. M., Ravlin, E. C. & Adkins, C. L. (1989). „A Work Values Approach to Corporate Culture - A Field Test of the Value Congruence Process and Its Relationship to Individual Outcomes“. In: *Journal of Applied Psychology* 74 (3), S. 424–432.

Merritt, S. L. & Marshall, J. C. (1984). „Reliability and construct validity of alternate forms of the CLS Inventory“. In: *Advances in Nursing Science* 7 (1), S. 78–85.

Moors, A. (2013). „On the Causal Role of Appraisal in Emotion“. In: *Emotion Review* 5 (2), S. 132–140.

Moosbrugger, H. (2012). „Item-Response-Theorie (IRT)“. In: *Testtheorie und Fragebogenkonstruktion*. Hrsg. von Moosbrugger, H. & Kelava, A. Berlin: Springer, S. 227–274.

Moosbrugger, H. & Kelava, A. (2012). „Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien)“. In: *Testtheorie und Fragebogenkonstruktion*. Hrsg. von Moosbrugger, H. & Kelava, A. Berlin: Springer, S. 7–26.

Mount, M. K., Barrick, M. R., Scullen, S. M. & Rounds, J. (2005). „Higher-Order Dimensions of the Big Five Personality Traits and the Big Six Vocational Interest Types“. In: *Personnel Psychology* 58 (2), S. 447–478.

Murray, H. A. (1938). *Explorations in Personality. A Clinical And Experimental Study of Fifty Men of College Age*. New York: Oxford University Press.

Musek, J. (2007). „A general factor of personality: Evidence for the Big One in the five-factor model“. In: *Journal of Research in Personality* 41 (6), S. 1213–1233.

Muthén, L. K. & Muthén, B. O. (1998-2012). *Mplus User's Guide*. Version Seventh Edition. URL: www.statmodel.com.

Nagelkerke, N. J. (1991). „A note on a general definition on the coefficient of determination“. In: *Biometrika* 78 (3), S. 691–692.

Navarro, D. J. (2015). *Learning statistics with R: A tutorial for psychology students and other beginners*. Version 0.5. Adelaide und Australia.

O'Reilly, C. A. I., Chatman, J. & Caldwell, D. F. (1991). „People and organizational culture: A profile comparison approach to assessing person–organization fit.“ In: *Academy of Management Journal* 34 (3), S. 487–516.

Parks-Leduc, L., Feldman, G. & Bardi, A. (2015). „Personality Traits and Personal Values: A Meta-Analysis“. In: *Personality and Social Psychology Review* 19 (1), S. 3–29.

Perugini, M. & Bagozzi, R. P. (2001). „The role of desires and anticipated emotions in goal-directed behaviours: Broadening and deepening the theory of planned behaviour“. In: *British Journal of Social Psychology* 40 (1), S. 79–98.

Pervin, L. A., Cervone, D. & John, O. P. (2005). *Persönlichkeitstheorien*. München: Reinhardt.

Pinder, C. C. (2008). *Work motivation in organizational behavior*. New York: Psychology Press.

Ping, P. F., Tsui, A. S., Liu, J. & Li, L. (2010). „Pursuit of Whose Happiness? Executive Leaders' Transformational Behaviors and Personal Values“. In: *Administrative Science Quarterly* 55, S. 222–254.

Ponschab, R., Genius-Devime, B. & Schweizer, A. (2009). *Kooperation statt Konfrontation. Verhandeln in der Anwaltspraxis*. Köln: Schmidt.

Quaquebeke, N. v., Kerschreiter, R., Buxton, A. E. & Dick, R. v. (2010). „Two Lighthouses to Navigate: Effects of Ideal and Counter-Ideal Values on Follower Identification and Satisfaction with Their Leaders“. In: *Journal of Business Ethics* (93), S. 293–305.

Rajagopalan, N. & Datfa, D. K. (1996). „CEO Characteristics: Does Industry Matter?“ In: *Academy of Management Journal* 39 (1), S. 197–215.

Rammstedt, B. (2007). „The 10-Item Big Five Inventory“. In: *European Journal of Psychological Assessment* 23 (3), S. 193–201.

Rammstedt, B. & John, O. P. (2007). „Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German“. In: *Journal of Research in Personality* 41 (1), S. 203–212.

Rammstedt, B., Goldberg, L. R. & Borg, I. (2010). „The measurement equivalence of Big-Five factor markers for persons with different levels of education“. In: *Journal of Research in Personality* 44 (1), S. 53–61.

Rammstedt, B., Kemper, C. J., Klein, M. C., Beierlein, C. & Kovaleva, A. (2012). *Eine kurze Skala zur Messung der fünf Dimensionen der Persönlichkeit: Big-Five-Inventory-10 (BFI-10)*. Köln: GESIS.

Revelle, W. (2015). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Version 1.5.1. Online-Adresse: cran.r-project.org/package=psych. Evanston und Illinois: Northwestern University.

Rheinberg, F. (2010). „Intrinsische Motivation und Flow-Erleben". In: *Motivation und Handeln*. Hrsg. von Heckhausen, J. & Heckhausen, H. 4. Aufl. Berlin: Springer, S. 365–388.

Rheinberg, F. (2004). *Motivationsdiagnostik*. Göttingen: Hogrefe.

Rheinberg, F. & Vollmeyer, R. (2012). *Motivation*. Stuttgart: Kohlhammer.

Rheinberg, F., Vollmeyer, R. & Engeser, S. (2003). „Die Erfassung des Flow-Erlebens". In: *Diagnostik von Motivation und Selbstkonzept*. Hrsg. von Stiensmeier-Pelster, J. & Rheinberg, F. Göttingen: Hogrefe, S. 261–279.

Richins, M. L. (2004). „The Material Values Scale: Measurement Properties and Development of a Short Form". In: *Journal of Consumer Research* 31 (1), S. 209–219.

Roccas, S., Sagiv, L., Schwartz, S. H. & Knafo, A. (2002). „The Big Five Personality Factors and Personal Values". In: *Personality and Social Psychology Bulletin* 28, S. 789–801.

Roe, R. A. & Ester, P. (1999). „Values and Work Empirical Findings and Theoretical Perspective". In: *Applied Psychology: An International Review* 48 (1), S. 1–21.

Rohan, M. J. (2000). „A Rose by Any Name? The Values Construct". In: *Personality and Social Psychology Review* 4 (3), S. 255–277.

Rokeach, M. (1973). *The Nature of Human Values*. New York: Free Press.

Rosenstiel, L. v. (2007). *Grundlagen der Organisationspsychologie. Basiswissen und Anwendungshinweise*. Stuttgart: Schäffer-Poeschel.

Rothermund, K. & Eder, A. B. (2011). *Allgemeine Psychologie. Motivation und Emotion*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Russell, J. A. (2003). „Core Affect and the Psychological Construction of Emotion". In: *Psychological Review* 110 (1), S. 145–172.

Russell, J. A. & Carroll, J. M. (1999). „On the bipolarity of positive and negative affect". In: *Psychological Bulletin* 125 (1), S. 3–30.

Russell, J. A., Weiss, A. & Mendelsohn, G. A. (1989). „Affect Grid: A single-item scale of pleasure and arousal“. In: *Journal of Personality and Social Psychology* 57 (3), S. 493–502.

Ryan, R. M., Sheldon, K. M., Kasser, T. & Deci, E. L. (1996). „All Goals Are Not Created Equal. An Organismic Perspective on the Nature of Goals and Their Regulation“. In: *The Psychology of Action. Linking Cognition and Motivation to Behavior*. Hrsg. von Gollwitzer, P. M. & Bargh, J. A. New York: Guilford Press, S. 7–26.

Ryckman, R. M. (2008). *Theories of personality*. Belmont und CA: Thomson/Wadsworth.

Sagiv, L. & Schwartz, S. H. (2000). „Value priorities and subjective well-being: Direct relations and congruity effects“. In: *European Journal of Social Psychology* 30 (2), S. 177–198.

Saville, P. & Willson, E. (1991). „The reliability and validity of normative and ipsative approaches in the measurement of personality“. In: *Journal of Occupational and Organizational Psychology* 64 (3), S. 219–238.

Schallberger, U. (2005). *Kurzskalen zur Erfassung der Positiven Aktivierung, Negativen Aktivierung und Valenz in Experience Sampling Studien (PANAVA-KS). Theoretische und methodische Grundlagen, Konstruktvalidität und psychometrische Eigenschaften bei der Beschreibung intra- und interindividueller Unterschiede*. Zürich.

Scheffer, D. & Heckhausen, H. (2010). „Eigenschaftstheorien der Motivation“. In: *Motivation und Handeln*. Hrsg. von Heckhausen, J. & Heckhausen, H. 4. Aufl. Berlin: Springer, S. 43–72.

Schermelleh-Engel, K. & Werner, C. S. (2012). „Methoden der Reliabilitätsbestimmung“. In: *Testtheorie und Fragebogenkonstruktion*. Hrsg. von Moosbrugger, H. & Kelava, A. Berlin: Springer, S. 119–142.

Schmalt, H.-D. & Heckhausen, H. (2010). „Machtmotivation“. In: *Motivation und Handeln*. Hrsg. von Heckhausen, J. & Heckhausen, H. 4. Aufl. Berlin: Springer, S. 211–236.

Schmidt, P., Bamberg, S., Davidov, E., Herrmann, J. & Schwartz, S. H. (2007). „Die Messung von Werten mit dem “Portraits Value Questionnaire”“. In: *Zeitschrift für Sozialpsychologie* 38 (4), S. 261–275.

Schmitt, M. & Platzer, C. (2010). *Differentielle Psychologie und Persönlichkeitspsychologie kompakt*. 1. Aufl. Weinheim: Beltz.

Schmitt, N. & Stults, D. M. (1986). „Methodology Review: Analysis of Multitrait-Multimethod Matrices“. In: *Applied Psychological Measurement* 10 (1), S. 1–22.

Schumacker, R. E. (2014). *Learning statistics using R*. SAGE Publications Ltd.

Schwartz, S. H. (1992). „Universals in the Content and Structure of Values: Theoretical advances and Empirical Tests in 20 Countries“. In: *Advances in Experimental Social Psychology* 25.

Schwartz, S. H. (1994). „Are There Universal Aspects in the Structure and Contents of Human Values?“ In: *Journal of Social Issues* 50 (4), S. 19–45.

Schwartz, S. H. (1996). „Value Priorities and Behavior: Applying a Theory of Integrated Value Systems“. In: *The Psychology of Values. The Ontario Symposium, Volume 8*. Hrsg. von Seligman, C., Olson, J. M. & Zanna, M. P. Mahwah und N.J: L. Erlbaum Associates, S. 1–24.

Schwartz, S. H. (2003). *A Proposal for Measuring Value Orientations across Nations. Chapter 7 in the Questionnaire Development Package of the European Social Survey*. Hrsg. von European Social Survey. URL: www.europeansocialsurvey.org (besucht am 6. Okt. 2015).

Schwartz, S. H. (2005). „Robustness and fruitfulness of a theory of universals in individual human values“. In: *Valores e comportamentos nas organizações*. Hrsg. von Tamayo, Á. & Porto, J. B. Petrópolis: Vozes, S. 21–55.

Schwartz, S. H. (2012). „An Overview of the Schwartz Theory of Basic Values“. In: *Online Readings in Psychology and Culture* 2 (1).

Schwartz, S. H. & Bilsky, W. (1987). „Toward a universal psychological structure of human values“. In: *Journal of Personality and Social Psychology* 53 (3), S. 550–562.

Schwartz, S. H. & Bilsky, W. (1990). „Toward a Theory of the Universal Content and Structure of Values: Extensions and Cross-Cultural Replications“. In: *Journal of Personality and Social Psychology* 58 (5), S. 878–891.

Schwartz, S. H. & Rubel, T. (2005). „Sex differences in value priorities: Cross-cultural and multimethod studies“. In: *Journal of Personality and Social Psychology* 89 (6), S. 1010–1028.

Sechrest, L. (1963). „Incremental Validity: A Recommendation“. In: *Educational and Psychological Measurement* 23 (1), S. 153–158.

Shadish, W. R., Cook, T. D. & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Silvia, P. J. (2006). *Exploring the psychology of interest*. Oxford und New York: Oxford University Press.

Smith, C. A. & Ellsworth, P. C. (1985). „Patterns of Cognitive Appraisal in Emotion“. In: *Journal of Personality and Social Psychology* 48 (4), S. 813–838.

Smith, T. J. & McKenna, C. M. (2013). „A Comparison of Logistic Regression Pseudo R^2 Indices“. In: *Multiple Linear Regression Viewpoints* 39 (2), S. 17–26.

Sokolowski, K. & Heckhausen, H. (2010). „Soziale Bindung: Anschlussmotivation und Intimitätsmotivation“. In: *Motivation und Handeln*. Hrsg. von Heckhausen, J. & Heckhausen, H. 4. Aufl. Berlin: Springer, S. 193–210.

Spranger, E. (1921). *Lebensformen: Geisteswissenschaftliche Psychologie und Ethik der Persönlichkeit*. Halle/Saale: Max Niemeyer.

Stark, S., Chernyshenko, O. S. & Drasgow, F. (2005). „An IRT Approach to Constructing and Scoring Pairwise Preference Items Involving Stimuli on Different Dimensions: The Multi-Unidimensional Pairwise-Preference Model“. In: *Applied Psychological Measurement* 29 (3), S. 184–203.

Stiensmeier-Pelster, J. & Heckhausen, H. (2010). „Kausalattribution von Verhalten und Leistung“. In: *Motivation und Handeln*. Hrsg. von Heckhausen, J. & Heckhausen, H. 4. Aufl. Berlin: Springer.

Sturm, A., Gurt, J. & Opterbeck, I. (2011). *Organisationspsychologie*. Wiesbaden: VS-Verlag.

Tabachnick, B. G. & Fidell, L. S. (2007). *Using Multivariate Statistics*. Boston und Munich: Pearson. 980 S.

Tajfel, H. (1982). „Social Psychology of Intergroup Relations“. In: *Annual Review of Psychology* 33 (1), S. 1–39.

Tamir, P. & Lunetta, V. N. (1977). „A comparison of ipsative and normative procedures in the study of cognitive preferences“. In: *The Journal of Educational Research* 71 (2), S. 86–93.

Tenopyr, M. L. (1988). „Artifactual reliability of forced-choice scales“. In: *Journal of Applied Psychology* 73 (4), S. 749–751.

Thompson, B. (2007). „Effect sizes, confidence intervals, and confidence intervals for effect sizes“. In: *Psychology in the Schools* 44 (5), S. 423–432.

Thurstone, L. L. (1927). „Three psychological laws.“ In: *Psychological Review* 34 (6), S. 424–432.

Tversky, A. & Kahneman, D. (1981). „The framing of decisions and the psychology of choice“. In: *Science* 211 (4481), S. 453–458.

Vasilopoulos, N. L., Cucina, J. M., Dyomina, N. V., Morewitz, C. L. & Reilly, R. R. (2006). „Forced-Choice Personality Tests: A Measure of Personality and Cognitive Ability?“ In: *Human Performance* 19 (3), S. 175–199.

Vernon, P. E. & Allport, G. W. (1931). „A test for personal values“. In: *The Journal of Abnormal and Social Psychology* 26 (3), S. 231–248.

Verplanken, B. (2004). „Value congruence and job satisfaction among nurses: a human relations perspective“. In: *International Journal of Nursing Studies* 41 (6), S. 599–605.

Verplanken, B. & Holland, R. W. (2002). „Motivated Decision Making: Effects of Activation and Self-Centrality of Values on Choices and Behavior“. In: *Journal of Personality and Social Psychology* 82 (3), S. 434–447.

Verquer, M. L., Beehr, T. A. & Wagner, S. H. (2003). „A meta-analysis of relations between person–organization fit and work attitudes“. In: *Journal of Vocational Behavior* 63 (3), S. 473–489.

Versnel, H. & Koppenol, H. (2003). *Management Drives. The new approach for organisations and organisational issues*. Management Drives BV.

Versnel, H. & Koppenol, H. (2005). *The Values Matrix. The pattern we are trapped in*. Amsterdam: FT Prentice Hall. IX, 198.

Vinchur, A. J., Schippmann, J. S., Switzer, F. S. I. & Roth, P. L. (1998). „A meta-analytic review of predictors of job performance for salespeople“. In: *Journal of Applied Psychology* 83 (4), S. 586–597.

Viswesvaran, C. & Ones, D. S. (2000). „Measurement Error in "Big Five Factors" Personality Assessment: Reliability Generalization across Studies and Measures“. In: *Educational and Psychological Measurement* 60 (2), S. 224–235.

Vroom, V. H. (1964). *Work and motivation*. New York: Wiley.

Watson, D. & Tellegen, A. (1985). „Towards a Consensual Structure of Mood“. In: *Psychological Bulletin* 98 (2), S. 219–235.

Watson, D., Clark, L. A. & Tellegen, A. (1988). „Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales“. In: *Journal of Personality and Social Psychology* 54 (6), S. 1063–1070.

Welch, B. L. (1951). „On the Comparison of Several Mean Values: An Alternative Approach“. In: *Biometrika* 38 (3/4), S. 330.

Wilde, M., Bätz, K., Kovaleva, A. & Urhahne, D. (2009). „Überprüfung einer Kurzskala intrinsischer Motivation (KIM)“. In: *Zeitschrift für Didaktik der Naturwissenschaften* 15, S. 31–45.

Windzio, M. (2013). *Regressionsmodelle für Zustände und Ereignisse. Eine Einführung*. Wiesbaden: Springer Fachmedien.

Winter, D. G., John, O. P., Stewart, A. J., Klohnen, E. C. & Duncan, L. E. (1998). „Traits and Motives: Toward an Integration of Two Traditions in Personality Research“. In: *Psychological Review* 105 (2), S. 230–250.

Wooldridge, J. M. (2013). *Introductory Econometrics. A modern approach*. Mason und Ohio: South-Western Cengage Learning.

Zes, D., Lewis, J. & Landis, D. (2014). *kcirt: k-Cube Thurstonian IRT Models. Create, Simulate, Fit, Solve k-Cube Thurstonian IRT Models*. Version 0.6.0.

Zimmerman, B. J., Bandura, A. & Martinez-Pons, M. (1992). „Self-Motivation for Academic Attainment: The Role of Self-Efficacy Beliefs and Personal Goal Setting“. In: *American Educational Research Journal* 29 (3), S. 663–676.

Zinnes, J. L. & Griggs, R. A. (1974). „Probabilistic, Multidimensional Unfolding Analysis“. In: *Psychometrika* 39 (3), S. 327–350.

Zou, G. Y. (2007). „Toward Using Confidence Intervals to Compare Correlations“. In: *Psychological Methods* 12 (4), S. 399–413.

Abbildungsverzeichnis

Tabellenverzeichnis

Anhang A

Deskriptivstatistische Evaluation der Items

Tabelle 78. Paarweise Itemschwierigkeiten der MVSQ^{A}-Skala.

Item	GB^{A}	MA^{A}	GW^{A}	ER^{A}	GL^{A}	VE^{A}	NA^{A}
GB_1^A		.25	.22	.12	.09	.17	.38
MA_1^A	.75		.50	.28	.26	.33	.67
GW_1^A	.78	.50		.33	.30	.40	.69
ER_1^A	.88	.72	.67		.50	.55	.87
GL_1^A	.91	.74	.70	.50		.58	.86
VE_1^A	.83	.67	.60	.45	.42		.80
NA_1^A	.62	.33	.31	.13	.14	.20	
GB_2^A		.46	.34	.31	.16	.21	.35
MA_2^A	.54		.43	.36	.25	.21	.40
GW_2^A	.66	.57		.44	.26	.31	.49
ER_2^A	.69	.64	.56		.33	.33	.56
GL_2^A	.84	.75	.74	.67		.53	.74
VE_2^A	.79	.79	.69	.67	.47		.71
NA_2^A	.65	.60	.51	.44	.26	.29	
GB_3^A		.47	.42	.62	.29	.32	.65

MA_3^A	.53		.45	.70	.38	.34	.72
GW_3^A	.58	.55		.77	.40	.42	.77
ER_3^A	.38	.30	.23		.19	.19	.51
GL_3^A	.71	.62	.60	.81		.51	.85
VE_3^A	.68	.66	.58	.81	.49		.85
NA_3^A	.35	.28	.23	.49	.15	.15	
GB_4^A		.50	.37	.17	.20	.21	.16
MA_4^A	.50		.39	.13	.22	.16	.17
GW_4^A	.63	.61		.22	.29	.32	.28
ER_4^A	.83	.87	.78		.62	.56	.51
GL_4^A	.80	.78	.71	.38		.46	.41
VE_4^A	.79	.84	.68	.44	.54		.46
NA_4^A	.84	.83	.72	.49	.59	.54	
GB_5^A		.36	.52	.31	.18	.32	.51
MA_5^A	.64		.64	.45	.30	.44	.64
GW_5^A	.48	.36		.28	.15	.30	.48
ER_5^A	.69	.55	.72		.31	.53	.73
GL_5^A	.82	.70	.85	.69		.72	.87
VE_5^A	.68	.56	.70	.47	.28		.74
NA_5^A	.49	.36	.52	.27	.13	.26	
GB_6^A		.66	.60	.41	.30	.24	.89
MA_6^A	.34		.43	.26	.24	.18	.76
GW_6^A	.40	.57		.32	.28	.18	.85
ER_6^A	.59	.74	.68		.44	.29	.91
GL_6^A	.70	.76	.72	.56		.37	.91
VE_6^A	.76	.82	.82	.71	.63		.94
NA_6^A	.11	.24	.15	.09	.09	.06	
GB_7^A		.34	.15	.11	.14	.10	.22

MA_7^A	.66		.27	.16	.29	.14	.40
GW_7^A	.85	.73		.39	.55	.38	.63
ER_7^A	.89	.84	.61		.64	.40	.71
GL_7^A	.86	.71	.45	.36		.31	.60
VE_7^A	.90	.86	.62	.60	.69		.77
NA_7^A	.78	.60	.37	.29	.40	.23	
GB_8^A		.51	.23	.43	.13	.39	.72
MA_8^A	.49		.27	.41	.23	.38	.67
GW_8^A	.77	.73		.63	.49	.62	.86
ER_8^A	.57	.59	.37		.31	.51	.76
GL_8^A	.87	.77	.51	.69		.68	.90
VE_8^A	.61	.62	.38	.49	.32		.80
NA_8^A	.28	.33	.14	.24	.10	.20	
GB_9^A		.42	.48	.23	.15	.26	.43
MA_9^A	.58		.57	.30	.30	.33	.49
GW_9^A	.52	.43		.22	.20	.29	.45
ER_9^A	.77	.70	.78		.46	.52	.66
GL_9^A	.85	.70	.80	.54		.55	.72
VE_9^A	.74	.67	.71	.48	.45		.66
NA_9^A	.57	.51	.55	.34	.28	.34	
GB_{10}^A		.46	.43	.57	.16	.15	.41
MA_{10}^A	.54		.49	.63	.18	.15	.44
GW_{10}^A	.57	.51		.59	.16	.12	.46
ER_{10}^A	.43	.37	.41		.15	.11	.36
GL_{10}^A	.84	.82	.84	.85		.48	.83
VE_{10}^A	.85	.85	.88	.89	.52		.86
NA_{10}^A	.59	.56	.54	.64	.17	.14	

Anmerkung. Wertesysteme: GB = Geborgenheit; MA = Macht; GW = Gewissheit; ER = Erfolg; GL = Gleichheit; VE = Verstehen; NA = Nachhaltigkeit; A = Annäherung; Zahlen indizieren die Blocknummer.

Tabelle 79. Paarweise Itemschwierigkeiten der MVSQ^{V}-Skala.

Item	GB^{A}	MA^{A}	GW^{A}	ER^{A}	GL^{A}	VE^{A}	NA^{A}
GB_1^{V}		.70	.60	.58	.71	.75	.71
MA_1^{V}	.30		.35	.31	.48	.58	.46
GW_1^{V}	.40	.65		.50	.65	.72	.62
ER_1^{V}	.42	.69	.50		.62	.75	.62
GL_1^{V}	.29	.52	.35	.38		.61	.47
VE_1^{V}	.25	.42	.28	.25	.39		.38
NA_1^{V}	.29	.54	.38	.38	.53	.62	
GB_2^{V}		.44	.61	.47	.80	.84	.70
MA_2^{V}	.56		.64	.50	.77	.86	.73
GW_2^{V}	.39	.36		.39	.73	.82	.60
ER_2^{V}	.53	.50	.61		.78	.82	.67
GL_2^{V}	.20	.23	.27	.22		.58	.35
VE_2^{V}	.16	.14	.18	.18	.42		.26
NA_2^{V}	.30	.27	.40	.33	.65	.74	
GB_3^{V}		.36	.28	.52	.62	.58	.66
MA_3^{V}	.64		.50	.67	.70	.72	.77
GW_3^{V}	.72	.50		.72	.81	.76	.81
ER_3^{V}	.48	.33	.28		.59	.57	.62
GL_3^{V}	.38	.30	.19	.41		.47	.50
VE_3^{V}	.42	.28	.24	.43	.53		.57
NA_3^{V}	.34	.23	.19	.38	.50	.43	
GB_4^{V}		.40	.49	.26	.52	.42	.54

MA_4^V	.60		.66	.32	.61	.53	.63
GW_4^V	.51	.34		.23	.51	.41	.54
ER_4^V	.74	.68	.77		.73	.72	.78
GL_4^V	.48	.39	.49	.27		.41	.53
VE_4^V	.58	.47	.59	.28	.59		.65
NA_4^V	.46	.37	.46	.22	.47	.35	
GB_5^V		.41	.42	.66	.38	.51	.54
MA_5^V	.59		.52	.79	.47	.62	.65
GW_5^V	.58	.48		.74	.43	.57	.60
ER_5^V	.34	.21	.26		.23	.33	.35
GL_5^V	.62	.53	.57	.77		.65	.68
VE_5^V	.49	.38	.43	.67	.35		.54
NA_5^V	.46	.35	.40	.65	.32	.46	
GB_6^V		.73	.63	.78	.56	.60	.60
MA_6^V	.27		.35	.50	.31	.33	.32
GW_6^V	.37	.65		.66	.44	.47	.47
ER_6^V	.22	.50	.34		.29	.32	.30
GL_6^V	.44	.69	.56	.71		.52	.54
VE_6^V	.40	.67	.53	.68	.48		.49
NA_6^V	.40	.68	.53	.70	.46	.51	
GB_7^V		.43	.50	.43	.50	.55	.44
MA_7^V	.57		.60	.50	.55	.63	.53
GW_7^V	.50	.40		.39	.46	.54	.35
ER_7^V	.57	.50	.61		.59	.65	.52
GL_7^V	.50	.45	.54	.41		.55	.41
VE_7^V	.45	.37	.46	.35	.45		.34
NA_7^V	.56	.47	.65	.48	.59	.66	
GB_8^V		.17	.18	.36	.69	.13	.37

MA_8^V	.83		.54	.65	.86	.41	.74
GW_8^V	.82	.46		.62	.88	.42	.72
ER_8^V	.64	.35	.38		.73	.34	.55
GL_8^V	.31	.14	.12	.27		.10	.23
VE_8^V	.87	.59	.58	.66	.90		.79
NA_8^V	.63	.26	.28	.45	.77	.21	
GB_9^V		.29	.16	.41	.31	.46	.35
MA_9^V	.71		.37	.70	.60	.74	.58
GW_9^V	.84	.63		.78	.76	.80	.76
ER_9^V	.59	.30	.22		.47	.60	.44
GL_9^V	.69	.40	.24	.53		.58	.50
VE_9^V	.54	.26	.20	.40	.42		.42
NA_9^V	.65	.42	.24	.56	.50	.58	
GB_{10}^V		.61	.57	.52	.52	.70	.67
MA_{10}^V	.39		.44	.42	.39	.60	.55
GW_{10}^V	.43	.56		.46	.48	.65	.62
ER_{10}^V	.48	.58	.54		.49	.71	.64
GL_{10}^V	.48	.61	.52	.51		.74	.68
VE_{10}^V	.30	.40	.35	.29	.26		.44
NA_{10}^V	.33	.45	.38	.36	.32	.56	

Anmerkung. Wertesysteme: GB = Geborgenheit; MA = Macht; GW = Gewissheit; ER = Erfolg; GL = Gleichheit; VE = Verstehen; NA = Nachhaltigkeit; V = Vermeidung; Zahlen indizieren die Blocknummer.

Anhang B

Konvergenz des `kcirt`-Schätzungen

Die Abbildungen 14 und 15 zeigen das Verhalten der metaheuristischen stochastischen Suche bei der Schätzung der TIRT-Modellparameter des Annäherungs- und Vermeidungsmodells der Version 2 des Fragebogens. Es ist zu sehen, dass sich die Werte der Kostenfunktion, Faktorladungen und Varianzen der Merkmalsausprägungen bereits nach gut zehn Iterationen nicht mehr wesentlich verändern, ergo konvergieren. Auch die Utilities liegen nach gut zehn Schätzvorgängen nahe an den finalen Werten nach 50 Iterationen, dennoch sind hier größere Veränderungen zu beobachten als bei den anderen Modellparametern. Dieselben Muster können bei der Schätzung des Vermeidungsmodells beobachtet werden. Die Anzahl von 50 Iterationen kann somit als ausreichend hoch angesehen werden.

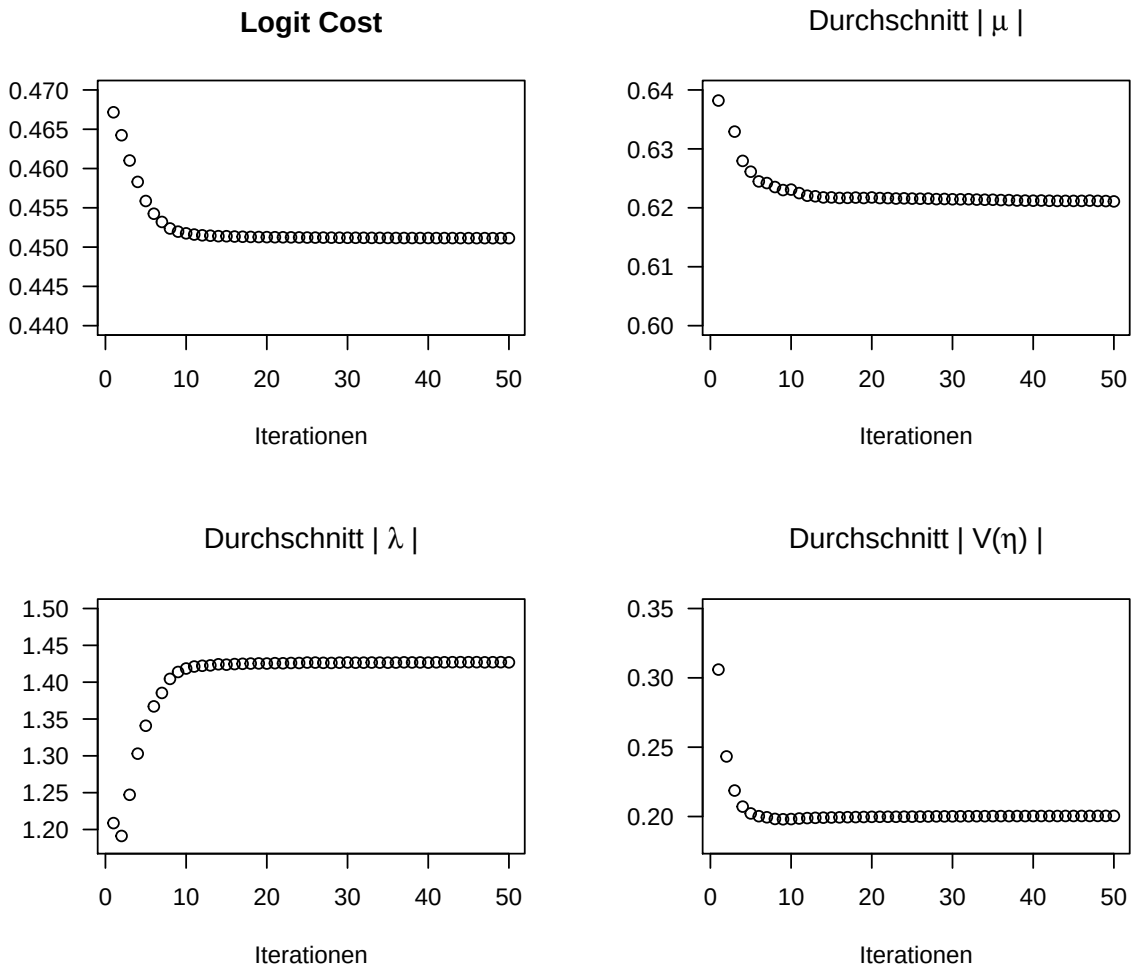

Abbildung 14. Konvergenz der metaheuristischen stochastischen Suche bei der Schätzung eines TIRT MVSQ-Annäherungs-Modells.

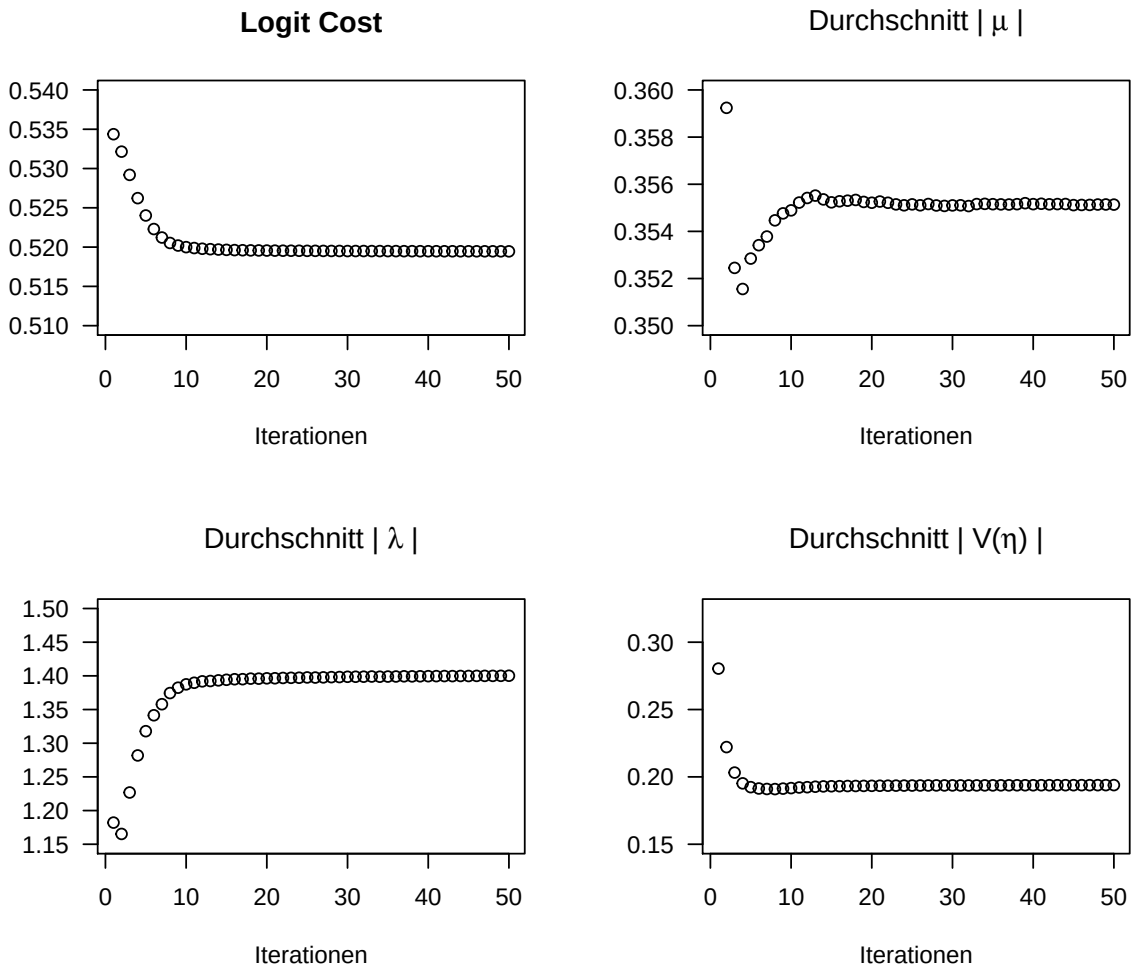

Abbildung 15. Konvergenz der metaheuristischen stochastischen Suche bei der Schätzung eines TIRT MVSQ-Vermeidungs-Modells.

Anhang C

Materialien der Studie zur prädiktiven Validität

Im Folgenden werden die in Kapitel 10.2 beschriebenen Materialien der experimentellen Studie aufgeführt. Das Deckblatt war dabei für alle Personen gleich. Danach folgt die Beschreibung des Szenarios, zuerst für die Gruppe der VPn, denen *kein* monetärer Anreiz geboten wurde, im Anschluss die Seite für die Gruppe *mit* Anreiz. Auf den nächsten drei Seiten folgt die *Sachbearbeitungsaufgabe*, danach die *Gewinnmaximierungsaufgabe* und zum Schluss die *Konzeptentwicklungsaufgabe*. Bei allen drei Aufgaben bestand der Unterschied zwischen den beiden Gruppen (mit und ohne Anreiz) jeweils in einem Satz am Ende der jeweiligen Aufgabenbeschreibung, der das jeweilige Leistungsmaß beschrieb. Hier sind nur die Aufgabenbeschreibungen der Gruppe *mit* Anreiz aufgeführt. In diesem Anhang nicht enthalten sind die Materialien der Skalen zur Erhebung der AVs. Für diese sei auf die entsprechenden Quellen verwiesen, die in Kapitel 10.2 beschrieben werden.

Regensburger Aufgabenstudie

Bearbeitungshinweise

Sehr geehrte Teilnehmerin, sehr geehrter Teilnehmer,

vielen Dank, dass Sie an der Regensburger Aufgabenstudie teilnehmen! Die Studie ist Teil einer Doktorarbeit zum Thema Arbeitsmotivation. In den kommenden 60 Minuten werden Sie drei Aufgaben aus den drei Bereichen *Gewinnmaximierung*, *Sachbearbeitung* und *Konzeptentwicklung* bearbeiten.

Für jede der Aufgaben haben Sie 20 Minuten Zeit. Diese 20 Minuten sollten ausreichend sein, um die jeweilige Aufgabe zu bearbeiten *und* Fragen zu den Aufgaben zu beantworten (ca. 4 Min). Es wird Ihnen bei jeder Aufgabe mitgeteilt, wann Sie mit der Bearbeitung beginnen sollen und wann Sie die Fragen *spätestens* beantworten sollen. D.h. wenn Sie mit der Aufgabe fertig sind, bevor die Versuchsleitung ein Signal gibt, dann beantworten Sie die Fragen zur Aufgabe gleich im Anschluss.

Alle Personen in diesem Raum werden dieselben Aufgaben bearbeiten, allerdings in unterschiedlichen Reihenfolgen. Wundern Sie sich also nicht, wenn Ihr Nachbar nicht dieselbe Aufgabe bearbeitet wie Sie.

Die Ergebnisse werden nur in pseudonymisierter Form ausgewertet und können in wissenschaftliche Veröffentlichungen nur in aggregierter Form einfließen. In keinem Fall wird es möglich sein, auf Einzelergebnisse rückzuschließen.

Bitte bearbeiten Sie die Aufgaben sorgfältig und beantworten Sie die Fragen gewissenhaft. Nehmen Sie Ihre Markierung bei den Fragen stets eindeutig vor. Markierungen zwischen zwei Kästchen können nicht ausgewertet werden. Haben Sie versehentlich eine falsche Markierung gesetzt, so kreisen Sie diese bitte ein und setzen Sie eine neue Markierung an einer anderen Stelle. Bei Verständnisproblemen melden Sie sich bitte und fragen Sie die Versuchsleitung.

Bitte tragen Sie zur Pseudonymisierung Ihren Zugangscode zum Online-Fragebogen ein und notieren Sie daneben das heutige Datum:

Zugangscode: ☐☐☐☐☐☐☐☐☐☐ *Datum:*_____

Beginnen Sie mit der Bearbeitung der ersten Aufgabe erst dann, wenn die Versuchsleitung das Startzeichen gibt.

Das Szenario für alle Aufgaben:

Stellen Sie sich vor, Sie haben einen Nebenjob in einem Café, das in der Nähe des Uni-und Hochschulgeländes gelegen ist und „Café am Campus“ heißt. Das Café ist bekannt für seine Auswahl an Kuchen und anderen Backwaren. Außerdem gibt es Kaffee, Tee und einige Erfrischungsgetränke. Es ist ein Catering-Service angegliedert und Ihre Tätigkeiten umfassen verschiedene Bereiche. Die drei Aufgaben haben alle mit diesem Nebenjob zu tun.

Hinweis

Dies ist keine Prüfung. Sie können nicht bestehen oder durchfallen. In dieser Studie geht es *nicht* um Ihre Leistung. Das Ziel der Studie liegt darin, die unterschiedlichen Aufgabentypen anhand der Fragen zu den Aufgaben zu vergleichen. Bitte bearbeiten Sie die Aufgaben trotzdem gewissenhaft und konzentriert. Bitte lassen Sie sich auf die Aufgaben ein, spüren Sie hin, wie es Ihnen bei der Bearbeitung geht und beantworten Sie die Fragen zu den Aufgaben ehrlich.

Vielen Dank!

Bitte blättern Sie jetzt um und beginnen Sie mit der ersten Aufgabe.

Das Szenario für alle Aufgaben:

Stellen Sie sich vor, Sie haben einen Nebenjob in einem Café, das in der Nähe des Uni- und Hochschulgeländes gelegen ist und „Café am Campus“ heißt. Das Café ist bekannt für seine Auswahl an Kuchen und anderen Backwaren. Außerdem gibt es Kaffee, Tee und einige Erfrischungsgetränke. Es ist ein Catering-Service angegliedert und Ihre Tätigkeiten umfassen verschiedene Bereiche. Die drei Aufgaben haben alle mit diesem Nebenjob zu tun.

Preise für die besten Studienteilnehmer

Die **drei besten** Teilnehmer der Studie erhalten folgende Geldgewinne:

1. Platz: 50 Euro

2. Platz: 25 Euro

3. Platz: 15 Euro

Bei jeder Aufgabe gibt es Kriterien für die Leistungsbewertung. Es gewinnt die/derjenige mit den meisten Gesamtpunkten. Die Aufgaben sind gleich gewichtet. Wie Sie die Leistungspunkte erreichen, wird je vor den Aufgaben erläutert.

Bitte tragen Sie in die folgende leere Zeile ein, wie wir Sie im Falle des Gewinns am besten erreichen können: _____.

E-Mail-Adresse *oder* Telefon-/Handynummer

Falls Sie zu den besten drei gehören, werden wir Sie kontaktieren und Ihnen den Gewinn in **bar** auszahlen.

Haben Sie Ihre Kontaktdaten eingetragen? Dann blättern Sie um und beginnen Sie mit der ersten Aufgabe!

Sachbearbeitungsaufgabe

Hinweise zur Bearbeitung der Aufgabe

Im Café am Campus werden in regelmäßigen Abständen Marketing-Aktionen durchgeführt, um den Umsatz zu steigern. In der Regel werden dazu kleine Zettel mit Sonderangeboten verteilt. Ihr Chef hat entschieden, dass nun wieder ein solche Marketingaktion durchzuführen ist.

Da der Drucker derzeit defekt ist, müssen die Zettel per Hand geschrieben werden. Auf den folgenden beiden Seiten finden Sie vorgefertigte Schablonen, um 50 Zettel zu beschreiben. Alle Zettel sollen so aussehen, wie hier dargestellt:

Café am Campus
Croissants für 99 Cent

Bitte achten Sie darauf, dass Ihre Schrift gut lesbar ist und schreiben Sie alle Wörter aus, d.h. keine Abkürzungen. Schreiben Sie nicht über die Linien, denn dann können die Zettel nicht verwendet werden. Am besten ist die Aufgabe ausgeführt, wenn alle Zettel exakt gleich aussehen. Die Felder werden danach ausgeschnitten und auf dem Campus-Gelände verteilt.

Leistungsmaß für diese Aufgabe: Anzahl der korrekt und sauber beschriebenen Zettel.

Haben Sie alle Felder fertig beschrieben? Dann blättern Sie jetzt bitte um und beantworten Sie die Fragen.

Gewinnmaximierungsaufgabe

Hinweise zur Bearbeitung der Aufgabe

Zu Ihren Tätigkeiten im „Café am Campus“ gehören:

- Einkauf und Verkauf von Waren,
- Planung und Durchführung von Cateringaufträgen.

Das Café erwirtschaftet im Monat durchschnittlich 5000 € Gewinn. Ihr monatliches Fixgehalt liegt bei 600 €. Zusätzlich haben Sie eine Gewinnbeteiligung von 10%. D.h. wenn Sie den Gewinn um z.B. 10% (= 500 €) steigern, erhalten Sie 50 € mehr Lohn. Der Gewinn, den Sie durch das Verkaufen von Waren vom Vortag erhalten, dürfen Sie behalten. Außerdem erhalten Sie einen Anteil der Einnahmen von Catering-Aufträgen, die Sie verantworten.

Auf den folgenden Seiten werden Ihnen 15 Entscheidungssituationen skizzenhaft beschrieben. Ihre Aufgabe ist es, in jeder Situation *eine* Entscheidung zu treffen. Bei den Entscheidungsmöglichkeiten sind Prozentzahlen angegeben, die jeweils Wahrscheinlichkeiten ausdrücken.

Entscheiden Sie so, dass Ihre persönlichen **Einnahmen maximal** werden!

Leistungsmaß für diese Aufgabe: Höhe der persönlichen Einnahmen. Je höher die persönlichen Einnahmen, umso besser.

Entscheidung 1 von 15
Vom Vortag sind noch 4 Croissants mit einem Verkaufswert von 7,20 € übrig. Die Einnahmen von alten Waren, die Sie verkaufen, werden komplett zu Ihren Einnahmen addiert. Was machen Sie mit den Croissants, wenn die Wahrscheinlichkeiten, dass die (als „alt" kegennzeichneten) Croissants gekauft werden, wie unten angegeben sind?

- ☐ Zum halben Preis anbieten. Gewinn: 3,60 € bei 90% Wahrscheinlichkeit. 0 € zu 10%.
- ☐ Zum vollen Preis anbieten. Gewinn: 7,20 € bei 60% Wahrscheinlichkeit. 0 € zu 40%.

Entscheidung 2 von 15
Bisher werden in Ihrem Café Premium-Kaffeebohnen verwendet. Sie finden aber, dass guter Kaffee nicht unbedingt teuer sein muss. Wenn die Kunden die neuen Kaffeebohnen annehmen, steigern Sie mit günstigeren Einkaufspreisen den Gewinn. Es besteht aber die Möglichkeit, dass die neuen Kaffeebohnen nicht gut bei Ihren Kunden ankommen und als Folge der Absatz sinkt. Wenn Sie die Wahl zwischen folgenden Möglichkeiten haben, für welche entscheiden Sie sich?

- ☐ Mittelklasse-Kaffee. Der Monatsgewinn steigt um 5%, d.h. Ihr Lohn erhöht sich um 25 € mit einer Wahrscheinlichkeit von 80%. Zu 20% sinkt der Monatsgewinn um 5% und Ihr Lohn um 25 €.
- ☐ Günstiger Kaffee. Der Monatsgewinn steigt um 8% , d.h. Ihr Lohn steigt um 40 € mit einer Wahrscheinlichkeit von 60%. Zu 40% sinkt der Monatsgewinn um 8% und Ihr Lohn um 40 €.

Entscheidung 3 von 15
Vom Vortag sind noch 8 Krapfen mit einem Verkaufswert von 12,80 € übrig. Was machen Sie damit, wenn die Wahrscheinlichkeiten dafür, dass die Krapfen gekauft werden, wie unten angegeben sind?

- ☐ Zum vollen Preis anbieten. Gewinn: 12,80 € bei 50% Wahrscheinlichkeit. 0 € zu 50%.
- ☐ Zum halben Preis anbieten. Gewinn: 6,40 € bei 90% Wahrscheinlichkeit. 0 € zu 10%.

Entscheidung 4 von 15
Erfrischungsgetränke wie Cola und Limonade werden in Gläsern serviert. Bisher kaufen Sie die Getränke von bekannten internationalen Marken, deren Namen allerdings nirgends zu sehen sind. Welche der folgenden Alternativen wählen Sie, wenn die Erfolgsaussichten wie unten angegeben sind?

- ☐ Discount-Getränke. Monatsgewinn steigt um 10% (und Ihr Lohn um 50 €) zu 75% Wahrscheinlichkeit. Zu 25% sinkt der Gewinn um 5% (Ihr Lohn um 25 €).
- ☐ Lokale Hersteller (mittlerer Preis). Monatsgewinn steigt um 6% (Ihr Lohn um 30 €) zu 95% Wahrscheinlichkeit. Zu 5% sinkt der Gewinn um 3% (Ihr Lohn um 15 €).

Entscheidung 5 von 15
Vom Vortag sind noch 6 Käse-Brötchen mit einem Verkaufswert von 10,20 € übrig. Was machen Sie damit, wenn die Wahrscheinlichkeiten dafür, dass die Käse-Brötchen zurückgehen wie unten angegeben sind?

- ☐ Zu einem leicht reduzierten Preis anbieten. Gewinn: 8.4 € bei 70% Wahrscheinlichkeit. 0 € zu 30%.
- ☐ Zum halben Preis anbieten. Gewinn: 5,10 € bei 90% Wahrscheinlichkeit. 0 € zu 10%.

Überschlagen Sie jetzt, wie viel Mehreinnahmen Sie mit Ihren Entscheidungen bisher generiert haben!

Entscheidung 6 von 15
Sie beziehen Ihre (handgemachten) Croissants aus einer regionalen Bäckerei, die nun die Preise erhöht hat. Wenn Sie weiterhin dort einkaufen, sinkt Ihr Monatsgewinn. Wenn Sie auf günstigere, qualitativ niederwertige Croissants umsteigen, wissen Sie nicht, wie Ihre Kunden darauf reagieren. Wie entscheiden Sie bei folgenden Möglichkeiten?

- ☐ Umstieg auf Großbäckerei. 50% Wahrscheinlichkeit für Gewinnsteigerung um 2% (Ihr Lohn steigt um 10 €) und 50% Wahrscheinlichkeit für Gewinnreduzierung um 8% (Ihr Lohn sinkt um 40 €).
- ☐ Weiterhin handgemachte Croissants beziehen. Der Monatsgewinn sinkt um 4% (20 € weniger Lohn) sicher.

Entscheidung 7 von 15
Vom Vortag sind noch 5 Butterbrezeln mit einem Verkaufswert von 9,50 € übrig. Was machen Sie damit, wenn die Wahrscheinlichkeiten, dass die Brezeln gekauft werden, wie unten angegeben sind?

- ☐ Zu einem leicht reduzierten Preis anbieten. Gewinn: 7,50 € bei 50% Wahrscheinlichkeit. 0 € zu 50%.
- ☐ Zum halben Preis anbieten. Gewinn: 4,75 € bei 90% Wahrscheinlichkeit. 0 € zu 10%.

Entscheidung 8 von 15
Die bei Ihnen verkauften Kuchen werden in einer regionalen Konditorei hergestellt. Nun wurden die Preise dafür erhöht. Wenn Sie weiterhin dort einkaufen, sinkt Ihr Monatsgewinn. Wie Ihre Kunden auf günstigere, qualitativ niederwertige Produkte reagieren, ist unsicher. Wie entscheiden Sie, wenn es folgenden Möglichkeiten gibt?

- ☐ Weiterhin regional hergestellte Kuchen beziehen. Der Monatsgewinn sinkt um 3% (Lohn sinkt um 15 €) sicher.
- ☐ Umstieg auf Großkonditorei. 50% Wahrscheinlichkeit für Gewinnsteigerung um 6% (Lohn steigt um 30 €) und 50% Wahrscheinlichkeit für Gewinnreduzierung um 12% (Lohn sinkt um 60 €).

Entscheidung 9 von 15
Die bei Ihnen verkauften Schokoladen-Muffins stammen aus einer regionalen Konditorei. Nun wurde die Preise dafür erhöht. Wenn Sie weiterhin dort einkaufen, sinkt Ihr Monatsgewinn. Wie Ihre Kunden auf günstigere, qualitativ niederwertige Produkte reagieren, ist unsicher. Wie entscheiden Sie, wenn es folgenden Möglichkeiten gibt?

- ☐ Weiterhin regional hergestellte Muffins beziehen. Der Monatsgewinn sinkt um 2% (= 10 €) sicher.
- ☐ Umstieg auf Großkonditorei. 50% Wahrscheinlichkeit für Gewinnsteigerung um 1% (= 5 €) und 50% Wahrscheinlichkeit für Gewinnreduzierung um 4% (= 20 €).

Entscheidung 10 von 15
Vom Vortag sind noch 10 Schokoladen-Muffins mit einem Verkaufswert von 14 € übrig. Was machen Sie damit, wenn die Wahrscheinlichkeiten, dass die Muffins zurückgehen, wie unten angegeben sind?

- ☐ Zum vollen Preis anbieten. Gewinn: 14 € bei 60% Wahrscheinlichkeit. 0 € zu 40%.
- ☐ Zu einem reduzierten Preis anbieten. Gewinn: 9 € bei 90% Wahrscheinlichkeit. 0 € zu 10%.

Überschlagen Sie jetzt wieder, wie viel Mehreinnahmen Sie mit Ihren Entscheidungen bisher generiert haben!

Entscheidung 11 von 15
In regemäßigen Abständen führen Sie Cateringaufträge durch, durch die Sie die Chance haben, Ihr persönliches Einkommen erheblich zu steigern. Ob Sie einen Auftrag erhalten, hängt maßgeblich vom Preis Ihres Angebots ab. Wenn Sie den Auftrag nicht erhalten, fallen Kosten für die Angebotserstellung an, an denen Sie ebenfalls beteiligt sind. Die in den Entscheidungen angegebenen Zahlen entsprechen Ihrem Anteil an Einnahmen und möglichen Kosten, die Sie direkt zu Ihrem persönlichen Einnahmen dazurechnen bzw. davon abziehen müssen. Für das Catering bei einer Uni-Tagung können Sie eins der folgenden zwei Angebote unterbreiten. Für welches entscheiden Sie sich?

- ☐ 80% Wahrscheinlichkeit für Erhalt des Auftrags bei einem Gewinn von 150 €. 20% Wahrscheinlichkeit für Verlust von 25 € (Angebotserstellungskosten).
- ☐ 70% Wahrscheinlichkeit für Erhalt des Auftrags bei einem Gewinn von 200 €. 30% Wahrscheinlichkeit für Verlust von 25 € (Angebotserstellungskosten).

Entscheidung 12 von 15
Für einen Catering-Auftrag an der Hochschule können Sie folgende Angebote unterbreiten. Wie entscheiden Sie?

- ☐ 25% Wahrscheinlichkeit für 200 € Gewinn. 75% Wahrscheinlichkeit für 25 € Verlust (Angebotskosten).
- ☐ 99% Wahrscheinlichkeit für 25 € Gewinn. 1% Wahrscheinlichkeit für 25 € Verlust (Angebotskosten).

Entscheidung 13 von 15
Ein Privatkunde bittet Sie um die Vorlage eines Angebots für ein privates Fest. Welche Option wählen Sie?

- ☐ 90% Wahrscheinlichkeit für 50 € Gewinn. 10% Wahrscheinlichkeit für 25 € Verlust.
- ☐ 80% Wahrscheinlichkeit für 100 € Gewinn. 20% Wahrscheinlichkeit für 25 € Verlust.

Entscheidung 14 von 15
Die Stadt Regensburg bittet Sie darum, dass Sie (als einer von vielen) ein Angebot für einen größeren Catering-Auftrag einreichen. Wie entscheiden Sie bei folgenden Möglichkeiten?

- ☐ 60% Wahrscheinlichkeit für 300 € Gewinn. 40% Wahrscheinlichkeit für 50 € Verlust.
- ☐ 40% Wahrscheinlichkeit für 400 € Gewinn. 60% Wahrscheinlichkeit für 50 € Verlust.

Entscheidung 15 von 15
Für einen weiteren Catering-Auftrag an der Hochschule können Sie folgende Angebote unterbreiten. Wie entscheiden Sie?

- ☐ 40% Wahrscheinlichkeit für 250 € Gewinn. 60% Wahrscheinlichkeit für 50 € Verlust.
- ☐ 50% Wahrscheinlichkeit für 150 € Gewinn. 50% Wahrscheinlichkeit für 50 € Verlust.

Überschlagen Sie bitte jetzt, wie viel Mehreinnahmen Sie insgesamt mit all Ihren Entscheidungen generiert haben! Tragen Sie das Ergebnis hier ein: _____

Sind Sie fertig? Dann blättern Sie bitte um und beantworten Sie die Fragen.

Konzeptentwicklungsaufgabe

Hinweise zur Bearbeitung der Aufgabe

Die Universität und Hochschule Regensburg wollen gegebenenfalls auf dem Campusgelände ein gemeinsames Studierendenhaus bauen. So ein „Haus der Studierenden“ soll eine Gastronomie, einen Lern- und Weiterbildungsbereich sowie einen Arbeitsbereich besitzen. An diesem Ort sollen Studierende die Möglichkeit erhalten, Wissen aufzubauen, an interdisziplinären Projekten zu arbeiten und ihre Persönlichkeit zu entwickeln. Zudem soll ein Gastronomie-Konzept integriert werden.

Zu diesem Zweck führen Universität und Hochschule eine erste Ausschreibung durch, an der Sie (mit Ihrem „Café am Campus“ im Rücken) teilnehmen.

Ihre Aufgabe ist es jetzt, einen ersten Entwurf für die Gestaltung eines innovativen Studierendenhauses zu erarbeiten. Zeichnen Sie keine Architekturpläne, sondern machen Sie sich Gedanken über Gestaltungsmöglichkeitungen und die vielfältige Nutzung der Räume.

Sie haben dazu eine leere DinA4 Seite (folgendes Blatt, wenn der Platz nicht ausreicht, können Sie auch die Rückseite verwenden). Überlegen Sie sich, wie ein zukunftsweisender Lern- und Entwicklungsort für Studierende aussehen könnte, indem Lernen, Bildung, Projektarbeit und Gastronomie miteinander verknüpft sind.

Entwickeln Sie ein innovatives Konzept für Studierende der Generation Z. Stellen Sie sich Fragen wie:

- Wie gestaltet man eine gute Lernumgebung, die Kreativität fördert?
- Wie könnte das „Studieren 4.0“ aussehen?
- Welche Rahmenbedingungen begünstigen die Weiterentwicklung von Studierenden?
- Wie würde der optimale Lernort des Jahres 2025 aussehen?
- Wie kann man die erarbeiteten Merkmale in ein Gastronomiekonzept integrieren?

Machen Sie sich in diesem Entwurf erst mal keine Sorgen um das Budget und beschränken Sie sich nicht durch irgendwelche Vorgaben, denn wenn Ihr Konzept in die nähere Auswahl kommt, werden in einer weiteren Untersuchung Machbarkeitsstudien ausgearbeitet. Auch die Umsetzung ist in diesem Stadium der Konzeptentwicklung also nicht von Bedeutung. Darüber kann man sich später Gedanken machen.

Nennen Sie alle Elemente / Vorschläge, die Ihnen einfallen und machen Sie sich **Gedanken** über **Zusammenhänge und Wirkungen**. Erklären Sie falls nötig in Stichworten, wie die einzelnen Elemente zu verstehen sind.

Leistungsmaß für diese Aufgabe: Umfang des Konzepts. Je mehr unterschiedliche Elemente, desto mehr Punkte erhalten Sie.

Konzept

Colophon

Diese Dissertation wurde mit LaTeX 2_ε und R in RStudio erstellt.